Reducing Balancing Error for Causal Inference via Optimal Transport

Yuguang Yan [1] **Hao Zhou** [1] **Zeqin Yang** [1] **Weilin Chen** [1] **Ruichu Cai** [1 2] **Zhifeng Hao** [3]

Abstract

Most studies on causal inference tackle the issue of confounding bias by reducing the distribution shift between the control and treated groups. However, it remains an open question to adopt an appropriate metric for distribution shift in practice. In this paper, we define a generic balancing error on reweighted samples to characterize the confounding bias, and study the connection between the balancing error and the Wasserstein discrepancy derived from the theory of optimal transport. We not only regard the Wasserstein discrepancy as the metric of distribution shift, but also explore the association between the balancing error and the underlying cost function involved in the Wasserstein discrepancy. Motivated by this, we propose to reduce the balancing error under the framework of optimal transport with learnable marginal distributions and the cost function, which is implemented by jointly learning weights and representations associated with factual outcomes. The experiments on both synthetic and real-world datasets demonstrate the effectiveness of our proposed method.

1. Introduction

Causal inference from observational data aims to estimate the effect of the treatment from data, which are collected from a real-world scenario rather than well-designed randomized control trials (RCTs) (Concato et al., 2000). For data from RCTs, the average treatment effect (ATE) can be easily estimated by comparing the outcomes of treated and control groups (Hernán & Robins, 2010). Nevertheless, it is non-trivial to estimate ATE from observational data since treated and control groups follow different covariate distributions (Shalit et al., 2017). For example, to treat heart disease, a doctor typically prescribes surgery to younger patients and medication to older ones, resulting in different age distributions for these two groups. In other words, age is the confounder here that brings confounding bias.

Most studies for ATE estimation address the confounding bias by balancing the confounder distributions of the treated and control groups, so that ATE can be estimated by comparing the difference of the average outcomes between the balanced groups (Yao et al., 2021). However, the performance of distribution balancing highly relies on the metric of distribution discrepancy, *e.g.*, moment difference (Hainmueller, 2012) or integral probability metric (IPM) (Kong et al., 2023; Wei et al., 2023), and how to adopt an appropriate metric remains a challenging problem. For example, regarding the moment-based method, the first and second moments frequently used in existing works are insufficient to model complex distributions, while the higher-order moments are difficult to estimate in practice. For the IPM-based method, different function spaces derive different distance metrics, *e.g.*, maximum mean discrepancy or Hilbert-Schmidt Independence Criterion, while how to select the optimal function space is still under-explored.

In this paper, we seek to address the confounding bias from a generic balancing error, which is not necessarily a metric of distribution discrepancy on confounders and could be built based on factual and counterfactual outcomes, propensity scores, or other factors. To this end, we first define a balancing error on reweighted samples to describe the confounding bias, and then construct a theoretical connection between the balancing error and the Wasserstein discrepancy through the theory of optimal transport, which studies how to move masses from one distribution to another with a minimal transport cost (Villani, 2008; Peyré et al., 2019). We not only regard the Wasserstein discrepancy as a metric of distribution shift, but also deeply explore the association between our defined balancing error and the underlying cost function involved in optimal transport. Based on this, we propose an optimal transport model with learnable sample weights and the cost function associated with the balancing error to reduce the confounding bias.

In particular, we model the balancing error by incorporating factual outcomes, and design an optimal transport problem

[1]School of Computer Science, Guangdong University of Technology, Guangzhou, China [2]Pazhou Laboratory (Huangpu), Guangzhou, China [3]College of Science, Shantou University, Shantou, China. Correspondence to: Ruichu Cai <cairuichu@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

that is motivated by our theoretical connection between balancing error and optimal transport, in which the marginal distributions of data are modeled as probability mass functions and estimated as the sample weights, and the cost function is learned based on a representation subspace that is guided by factual outcomes. By doing this, we integrate two major learning paradigms, *i.e.*, reweighting and representation learning, into a unified model of optimal transport to minimize the Wasserstein discrepancy with the underlying cost function guided by factual outcomes. We develop an alternate algorithm to solve the resultant optimization problem, and conduct experiments on both synthetic and real-world datasets to evaluate the performance of our proposed method.

We summarize our principal contributions as follows:

- To address the issue of confounding bias, we define a generic balancing error on reweighted samples and theoretically connect it with the Wasserstein discrepancy via the theory of optimal transport.
- Motivated by our connections, we reduce the confounding bias by minimizing the Wasserstein discrepancy with learnable marginal distributions and the cost function associated with the balancing error.
- Based on our learning model, we design a balancing error based on potential outcomes, and develop an algorithm to simultaneously learn sample weights and representations with the guidance of factual outcomes.

2. Related Works

2.1. Causal Effect Estimation

There are two classes of methods to eliminate confounding bias for causal effect estimation. The first class is the reweighting method, which aims to create a pseudo-balanced group. Rosenbaum & Rubin (1983) proposed IPW, treating the inverse of propensity score as sample weight. To make IPW more robust, Robins et al. (1994) combined it with outcome regression and Imai & Ratkovic (2014) further exploited the dual characteristic of propensity score. To avoid the model misspecification problem of IPW, Hainmueller (2012) and Kuang et al. (2017) proposed to learn weights directly by aligning the moments between treated and control groups.

The other class is the representation-based method. Shalit et al. (2017) first introduced representation learning to causal effect estimation, showing that balancing the distributions of treated and control groups in the representation space can improve performance. Recently, various algorithms have been proposed to incorporate both representation-based learning and reweighting (Johansson et al., 2018; Assaad et al., 2021; Johansson et al., 2021). Compared with them, we not only integrate reweighting and representation learning in one unified optimal transport model to minimize the Wasserstein discrepancy between two groups, but also reveal that the underlying cost function in our optimal transport model is associated with a balancing error metric measured on data.

Recently, some empirical studies show that hyperparameter tuning can balance the covariates to some extent (Athey & Imbens, 2016; Curth & Van Der Schaar, 2023; Machlanski et al., 2023; Mahajan et al., 2022). For example, Kong et al. (2023) applied kernel MMD to measure the confounding bias, and Wei et al. (2023) tuned the hyperparameters within the IPW framework. Nevertheless, a theoretical analysis regarding hyperparameter tuning is still under-explored.

2.2. Optimal Transport

Optimal transport seeks to find an optimal plan for moving mass from one distribution to another with the minimal transport cost (Monge, 1781; Kantorovitch, 1958; Villani, 2008). Recently, optimal transport has shown powerful abilities in different kinds of applications (Peyré et al., 2019; Yan et al., 2019; Zhao & Zhou, 2018). For computer vision, the Earth Mover's Distance, which is calculated based on the solution to the optimal transport problem, is used as a similarity metric for image retrieval (Rubner et al., 2000). For transfer learning, data from one distribution is transported to another distribution based on the optimal transport plan for label information transfer (Courty et al., 2014; 2017b;a). For generative modeling, the Wasserstein distance derived from optimal transport is minimized to train deep generative models to generate high-quality data (Tolstikhin et al., 2018; Arjovsky et al., 2017). For structured data, Wasserstein (Maretic et al., 2022), Gromov-Wasserstein (Xu, 2020) and Fused Gromov-Wasserstein (Titouan et al., 2019) are applied for graph data analysis.

There are also a few researches trying to introduce optimal transport into causal inference (Wang et al., 2023). Gunsilius & Xu (2021) employed unbalanced optimal transport for matching. Torous et al. (2021) generalized Changes-in-Changes (CiC) to a high-dimensional setting based on optimal transport. Li et al. (2021) proposed to infer counterfactual outcomes via transporting factual distribution to the counterfactual distribution. Dunipace (2021) applied optimal transport to achieve distribution balance by finding an intermediate distribution with learned weights. Compared with them, we construct a theoretical connection between causal effect estimation and optimal transport, and further propose to learn the underlying cost function, which is shown to be associated with a bias involved in data. As a result, the optimal transport cost is not only a metric of distribution shift, but also a metric for reducing a balancing error with the corresponding underlying cost function.

3. Notations and Problem Statement

Throughout the paper, $[n]$ denotes a set including the elements $\{1, \dots, n\}$. $\mathbf{1}_n$ denotes a vector in the space $\mathbb{R}^n$ with all the elements being 1. For a matrix $\mathbf{A}$, the (i, j)-th element of $\mathbf{A}$ is denoted as A_{ij}, and $\mathbf{A}^\top$ is the transpose of $\mathbf{A}$. The trace of a square matrix $\mathbf{A}$ is defined as $\text{tr}(\mathbf{A}) = \sum_i A_{ii}$. Given two matrices $\mathbf{A}$ and $\mathbf{B}$ with the same size, the inner product of them is defined as

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_i \sum_j A_{ij} B_{ij} = \text{tr}(\mathbf{A}^\top \mathbf{B}) = \text{tr}(\mathbf{A}\mathbf{B}^\top). \quad (1)$$

The Hadamard product between $\mathbf{A}$ and $\mathbf{B}$ is denoted as $\mathbf{A} \odot \mathbf{B}$, *i.e.*, $(\mathbf{A} \odot \mathbf{B})_{ij} = A_{ij} B_{ij}$. For the two distributions $\mathbf{A}$ and $\mathbf{B}$ with the same size, the Kullback-Leibler (KL) divergence $\text{KL}(\mathbf{A}||\mathbf{B})$ is defined as

$$\text{KL}(\mathbf{A}||\mathbf{B}) = \sum_{i=1} \sum_{j=1} A_{ij} \log(\frac{A_{ij}}{B_{ij}}) - A_{ij} + B_{ij}. \quad (2)$$

The probability simplex Σ_n is defined as

$$\Sigma_n = \{\mathbf{u} \in \mathbb{R}^n \mid \sum_{i=1}^n u_i = 1, u_i \geq 0 \; \forall \, i \in [n]\}. \quad (3)$$

In this paper, we consider the Rubin-Neyman potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990) with n observational samples $\{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^n$, where $t_i \in \{0, 1\}$ is the binary treatment with 1 for treated groups and 0 for control groups, $\mathbf{x}_i \in \mathbb{R}^d$ is feature vector with d being the dimension of features, and $y_i \in \mathbb{R}$ is the observational outcome. We denote the sample numbers of treated and control groups as n_t, n_c, and we have $n = n_t + n_c$. Given the potential outcomes $Y_0(\cdot)$ and $Y_1(\cdot)$, the observational outcome is $y_i = t_i Y_1(\mathbf{x}_i) + (1 - t_i) Y_0(\mathbf{x}_i)$. For simplicity, we define data matrices $\mathbf{X}^c \in \mathbb{R}^{n_c \times d}$, $\mathbf{X}^t \in \mathbb{R}^{n_t \times d}$, $\mathbf{X} \in \mathbb{R}^{n \times d}$ including all the control samples, treated samples, and all the samples, respectively. And $\mathbf{y}^c \in \mathbb{R}^{n_c}$, $\mathbf{y}^t \in \mathbb{R}^{n_t}$ include the observational outcomes of the control and treated groups, *i.e.*, $\{y_i^c\}_{i=1}^{n_c}$ and $\{y_i^t\}_{i=1}^{n_t}$, respectively.

We focus on the Average Treatment Effect (ATE), which is the average difference between the potential outcomes of the treated and control groups on the entire population:

$$\text{ATE} = \mathbb{E}[Y_1(\mathbf{x}_i)] - \mathbb{E}[Y_0(\mathbf{x}_i)]. \quad (4)$$

The simplest approach to calculate Eq. (4) is directly comparing the average outcome between the treated and control groups, which will be biased because the confounding bias is not taken into account.

In this paper, we assume that the standard *strong ignorability* assumption is satisfied: $t \perp (Y_1(\mathbf{x}), Y_0(\mathbf{x}))|\mathbf{x}$ and $0 < p(t = 1|\mathbf{x}) < 1$ for all $\mathbf{x}$. Strong ignorability is a sufficient condition for ATE identification (Rosenbaum & Rubin, 1983). Under this assumption, ATE can be estimated unbiasedly by reweighting. In specific, reweighting aims to learn weights $\mathbf{w} = \{w_i\}_{i=1}^n$ for each sample to reduce the confounding bias, and ATE can be estimated by

$$\widehat{\text{ATE}} = \sum_{i=1}^{n_t} w_i^t y_i^t - \sum_{i=1}^{n_c} w_i^c y_i^c. \quad (5)$$

4. Learning Model

In this section, we first define the balancing error to measure the level of confounding bias, and then bound the balancing error via the Wasserstein discrepancy from optimal transport. After that, we propose our model for ATE via optimal transport, which consists of the learnable sample weights and the cost function guided by factual outcomes.

4.1. Connection Between ATE and Optimal Transport

Optimal transport finds an optimal plan to move mass from one distribution to another with the minimal transport cost. Among the rich theory of optimal transport, we focus on the *Kantorovich Problem*. Consider two distributions $\mu \in P(\mathcal{U}), \nu \in P(\mathcal{V})$ and a cost function $D : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$, the Kantorovich problem seeks an optimal plan $\pi(u, v)$ via optimizing the following problem:

$$KP_{\mu,\nu} = \inf_{\pi \in \Pi(\mu,\nu)} \int D(u, v) d\pi(u, v), \quad (6)$$

where $\Pi(\mu, \nu)$ denotes the set of all joint probability couplings whose first and second marginals are μ and ν, respectively. Kantorovich also provided a *Dual Problem*, known as the Kantorovich duality ((Villani, 2021), Theorem 1.3):

$$DP_{\mu,\nu} = \sup_{f(u)+g(v) \leq D(u,v)} \int f(u) d\mu(u) + \int g(v) d\nu(v). \quad (7)$$

We now turn our attention to the ATE estimation problem, whose main challenge comes from the confounding bias between treated and control groups. Hainmueller (2012) proposed a reweighting scheme that assigns a weight to each sample such that the reweighted groups satisfy the prespecified balance constraint $m(\cdot)$ that are imposed on covariate distributions, *i.e.*, $\sum_{i=1}^{n_c} w_i^c m(\mathbf{x}_i^c) = \frac{1}{n} \sum_{i=1}^n m(\mathbf{x}_i)$, $\sum_{i=1}^{n_t} w_i^t m(\mathbf{x}_i^t) = \frac{1}{n} \sum_{i=1}^n m(\mathbf{x}_i)$. Inspired by this, we define the *balancing error* to measure the level of confounding bias between different groups after reweighting:

$$err_m^{\mathbf{w}} = |\sum_{i=1}^{n_t} w_i^t m(\mathbf{x}_i^t) - \sum_{i=1}^n \frac{m(\mathbf{x}_i)}{n}| + |\sum_{i=1}^{n_c} w_i^c m(\mathbf{x}_i^c) - \sum_{i=1}^n \frac{m(\mathbf{x}_i)}{n}|. \quad (8)$$

We do not restrict the form of the function $m(\cdot)$, as long as the balancing error $err_m^{\mathbf{w}}$ can characterize the degree of confounding bias in some sense. Instead, we connect this error with optimal transport through the following theorem, in which the relationship between the function $m(\cdot)$ and the cost function $D(\cdot,\cdot)$ is established:

Theorem 4.1. *Let μ^c, μ^t, μ be the empirical distributions of the weighted control group, weighted treated group, and entire population, respectively, with the corresponding sample weight vector $\boldsymbol{\mu}^c \in \Sigma_{n_c}$, $\boldsymbol{\mu}^t \in \Sigma_{n_t}$, $\boldsymbol{\mu} \in \Sigma_n$. Suppose $m(\cdot) = f(\cdot), -m(\cdot) = g(\cdot)$, and assume there exists a cost function such that $m(u) - m(v) = f(u) + g(v) \leq D(u, v)$. We have the following results*

$$\begin{aligned} err_m^{\mathbf{w}} &\leq DP_{\mu^t,\mu} + DP_{\mu^c,\mu} \\ &\leq KP_{\mu^t,\mu} + KP_{\mu^c,\mu}. \end{aligned} \tag{9}$$

Proof. For the first inequality, under the assumption of Theorem 4.1, $DP_{\mu^t,\mu}$ is the worst-case of $\sum_{i=1}^{n_t} w_i^t m(\mathbf{x}_i^t) - \frac{1}{n}\sum_{i=1}^{n} m(\mathbf{x}_i)$, and so is $DP_{\mu^c,\mu}$. The second inequality holds because of the property of the dual problem, and the non-negative property of cost function $D(\cdot,\cdot)$ that leads to $|KP_{\mu^t,\mu}| = KP_{\mu^t,\mu}$. The second inequality could be an equality under the following assumptions: let $(\mathcal{U}^t, \mu^t)$, $(\mathcal{U}^c, \mu^c)$ and $(\mathcal{U}, \mu)$ be Polish spaces, $D(\cdot,\cdot)$ be a lower semi-continuous cost function, $m(\cdot) \in L^1(\cdot)$ and $-m(\cdot) \in L^1(\cdot)$, where L^1 is the Lebesgue space of exponent 1. For the condition when the strong duality holds so that the bound is tight, please refer to (Villani, 2008; 2021). □

Remark 4.2. Theorem 4.1 inspires a way to estimate weights by minimizing KP. Specifically, we could align the reweighted empirical distribution of treated/control group $\sum_i^{n_z} w_i^z \delta_{x_i}$ to that of full samples with uniform probability measure $\sum_i^n \frac{1}{n}\delta_{x_i}$, where $z \in \{t, c\}$, δ_x is the Dirac delta. At the same time, by aligning the empirical distribution, we could avoid the extreme case that one can just map treated and control samples to a point mass and lose all the information.

Remark 4.3. In intuition, Theorem 4.1 shows that the balancing error, which is used to characterize the lever of confounding bias, is bounded by the Wasserstein discrepancy with an underlying cost function $D(u, v)$ and marginal distributions as the sample weights, which theoretically supports that the confounding bias can be minimized by learning weights and the cost function in optimal transport.

We now discuss the design of the cost function $D(\cdot,\cdot)$ considering the assumption $m(\mathbf{x}_i) - m(\mathbf{x}_j) \leq D(\mathbf{x}_i, \mathbf{x}_j)$. One simple implementation of the cost function is the squared Euclidean distance, which is the default choice in optimal transport (Courty et al., 2017b) because of its simplicity and efficacy. In intuition, this is a reasonable choice since if the squared Euclidean distance between two samples is small which means they have similar representations, the difference of $m(\cdot)$ between them is usually small, too. In this situation, the vanilla Wasserstein discrepancy is adopted as the metric of distribution shift and minimized by learning weights to reduce confounding bias.

Nevertheless, considering that the target is to minimize the total transport cost with the underlying cost function, it is possible to further reduce the balancing error by learning the cost function $D(\cdot,\cdot)$. In fact, a small $m(\mathbf{x}_i) - m(\mathbf{x}_j)$ allows a small $D(\mathbf{x}_i, \mathbf{x}_j)$ while a large $m(\mathbf{x}_i) - m(\mathbf{x}_j)$ requires a large $D(\mathbf{x}_i, \mathbf{x}_j)$, which implicitly describe an approximation of the relevance between $m(\mathbf{x}_i) - m(\mathbf{x}_j)$ and $D(\mathbf{x}_i, \mathbf{x}_j)$. Motivated by this, we implement the balancing error as the bias between the potential outcomes of the two groups, which means that the potential outcome is involved in the function $m(\cdot)$, and propose to learn a cost function that is relevant to observational outcomes.

As a result, in order to reduce the balancing error, we propose to minimize the Wasserstein discrepancy with learnable weights as well as the learnable cost function by introducing the factual outcomes.

4.2. Reducing Balancing Error via Optimal Transport

Let $D(\cdot,\cdot;\phi)$ be the cost function between two samples parameterized by ϕ, $\mathbf{D}^c$ and $\mathbf{D}^t$ are the corresponding cost matrices from $\mathbf{X}^c$ to $\mathbf{X}$ and from $\mathbf{X}^t$ to $\mathbf{X}$, respectively. The elements in the cost matrices are defined as

$$D_{ij}^c = D(\mathbf{x}_i^c, \mathbf{x}_j; \phi), \quad D_{ij}^t = D(\mathbf{x}_i^t, \mathbf{x}_j; \phi). \tag{10}$$

We omit the parameters ϕ in the absence of ambiguity.

Motivated by Theorem 4.1, to obtain lower *KP*'s in Eq. (9), we minimize the Wasserstein discrepancy between $(\mathbf{X}^c, \boldsymbol{\mu}^c)$ to $(\mathbf{X}, \boldsymbol{\mu})$ and that between $(\mathbf{X}^t, \boldsymbol{\mu}^t)$ to $(\mathbf{X}, \boldsymbol{\mu})$ by learning sample weights and the transport cost simultaneously. To this end, we propose the following optimal transport model with estimated weights for the control and treated groups and the learnable transport cost,

$$\min_{\boldsymbol{\mu}^c, \boldsymbol{\mu}^t, \phi} \mathcal{W}(\mathbf{X}^c, \mathbf{X}, \phi) + \mathcal{W}(\mathbf{X}^t, \mathbf{X}, \phi), \tag{11}$$

where $\boldsymbol{\mu}^c$ is the estimated marginal distribution for reweighting the control group, and $\boldsymbol{\mu}^t$ is the estimated marginal distribution for reweighting the treated group. The Wasserstein discrepancy between two groups is defined as

$$\mathcal{W}(\mathbf{X}^c, \mathbf{X}, \phi) = \min_{\mathbf{T}^c \in \mathcal{T}(\boldsymbol{\mu}^c, \boldsymbol{\mu})} \langle \mathbf{D}^c, \mathbf{T}^c \rangle, \tag{12}$$

$$\mathcal{W}(\mathbf{X}^t, \mathbf{X}, \phi) = \min_{\mathbf{T}^t \in \mathcal{T}(\boldsymbol{\mu}^t, \boldsymbol{\mu})} \langle \mathbf{D}^t, \mathbf{T}^t \rangle, \tag{13}$$

where the domain of the transport matrix is defined as:

$$\mathcal{T}^c(\boldsymbol{\mu}^c, \boldsymbol{\mu}) = \{\mathbf{T}^c \in \mathbb{R}^{n_c \times n} \mid \mathbf{T}^c \mathbf{1}_n = \boldsymbol{\mu}^c, (\mathbf{T}^c)^\top \mathbf{1}_{n_c} = \boldsymbol{\mu}, T^c_{ij} \in [0,1] \,\forall\, i \in [n_c], j \in [n]\}, \quad (14)$$

$$\mathcal{T}^t(\boldsymbol{\mu}^t, \boldsymbol{\mu}) = \{\mathbf{T}^t \in \mathbb{R}^{n_t \times n} \mid \mathbf{T}^t \mathbf{1}_n = \boldsymbol{\mu}^t, (\mathbf{T}^t)^\top \mathbf{1}_{n_t} = \boldsymbol{\mu}, T^t_{ij} \in [0,1] \,\forall\, i \in [n_t], j \in [n]\}. \quad (15)$$

4.3. Factual Outcome Guided Cost Learning

For the cost function in our model, by implementing $m(\mathbf{x})$ as the potential outcome, we propose to learn the transport cost function $D(\cdot,\cdot;\phi)$ by considering its relevance with the factual outcomes in the following model

$$\min_{\boldsymbol{\mu}^c, \boldsymbol{\mu}^t, \phi} \mathcal{W}(\boldsymbol{\mu}^c, \boldsymbol{\mu}, \phi) + \mathcal{W}(\boldsymbol{\mu}^t, \boldsymbol{\mu}, \phi) + \mathcal{F}(\mathbf{y}^c, \mathbf{y}^t, \phi), \quad (16)$$

where the factual outcome guidance term $\mathcal{F}(\mathbf{y}^c, \mathbf{y}^t, \phi)$ exploits the outcomes $\mathbf{y}^c$ and $\mathbf{y}^t$ to learn a better transport cost function $D(\cdot,\cdot;\phi)$. Considering that samples of two groups lie in the same metric space with a common cost function $D(\cdot,\cdot;\phi)$, we reuse the $D(\cdot,\cdot;\phi)$ to construct the intra-group cost matrices $\mathbf{D}^{c,c}$ and $\mathbf{D}^{t,t}$ as

$$D^{c,c}_{ij} = D(\mathbf{x}^c_i, \mathbf{x}^c_j; \phi), \quad D^{t,t}_{ij} = D(\mathbf{x}^t_i, \mathbf{x}^t_j; \phi), \quad (17)$$

and also construct the affinity matrices $\mathbf{K}^c$ and $\mathbf{K}^t$ based on $\mathbf{y}^c$ and $\mathbf{y}^t$ as

$$K^c_{ij} = \exp(-\frac{(y^c_i - y^c_j)^2}{\sigma}), \quad K^t_{ij} = \exp(-\frac{(y^t_i - y^t_j)^2}{\sigma}). \quad (18)$$

Based on these, we implement $\mathcal{F}(\mathbf{y}^c, \mathbf{y}^t, \phi)$ as follows to leverage factual outcomes to guide the learning of the cost function

$$\mathcal{F}(\mathbf{y}^c, \mathbf{y}^t, \phi) = \lambda_c \langle \mathbf{D}^{c,c}, \mathbf{K}^c \rangle + \lambda_t \langle \mathbf{D}^{t,t}, \mathbf{K}^t \rangle, \quad (19)$$

where λ_c and λ_t are the trade-off parameters. In intuition, for the control group, if the factual outcome of $\mathbf{x}^c_i$ and $\mathbf{x}^c_j$ are close, *i.e.*, y^c_i and y^c_j are close, we will get a large K^c_{ij}, which will induce a small $D^{c,c}_{ij}$, indicating that the transport cost between $\mathbf{x}^c_i$ and $\mathbf{x}^c_j$ is small. So as the treated group. In this sense, the intra-group costs $\mathbf{D}^{c,c}$ and $\mathbf{D}^{t,t}$ are guided by the factual outcomes $\mathbf{y}^c$ and $\mathbf{y}^t$, and reflect the relevance between the outcomes of the cost function.

4.3.1. Implementation of Cost Function

Now we provide an implementation to learn the cost function $D(\cdot,\cdot,\phi)$. The common choice for transport cost is the squared Euclidean distance because of its simplicity and effectiveness (Courty et al., 2017b; Yan et al., 2019). Inspired by this, we extend the squared Euclidean distance in a learned subspace, which is achieved by defining $D(\cdot,\cdot,\phi)$ based on a projection matrix $\mathbf{P} \in \mathbb{R}^{d \times d'}$, *i.e.*,

$$D^c_{ij} = \|\mathbf{P}^\top \mathbf{x}^c_i - \mathbf{P}^\top \mathbf{x}_j\|_2^2, D^t_{ij} = \|\mathbf{P}^\top \mathbf{x}^t_i - \mathbf{P}^\top \mathbf{x}_j\|_2^2, \quad (20)$$

where d' is the dimension of the subspace. Similarly, the costs within the control and treated groups are defined as

$$D^{c,c}_{ij} = \|\mathbf{P}^\top \mathbf{x}^c_i - \mathbf{P}^\top \mathbf{x}^c_j\|_2^2, \; D^{t,t}_{ij} = \|\mathbf{P}^\top \mathbf{x}^t_i - \mathbf{P}^\top \mathbf{x}^t_j\|_2^2. \quad (21)$$

As a result, we optimize the projection matrix $\mathbf{P}$ rather than the cost function $D(\cdot,\cdot,\phi)$. By considering the guidance of factual outcomes in Eq. (19), we learn a subspace associated with the potential outcomes. We further constrain $\mathbf{P}$ to live in the Stiefel manifold:

$$\mathcal{M} = \{\mathbf{P} \in \mathbb{R}^{d \times d'} \mid \mathbf{P}^\top \mathbf{P} = \mathbf{I}\}, \quad (22)$$

which defines orthogonal subspaces.

By re-introducing $\mathbf{P}$, $\mathbf{T}^c$ and $\mathbf{T}^t$ into $D(\cdot,\cdot,\phi)$ and $\mathcal{W}(\cdot,\cdot,\phi)$, we implement Problem (16) to achieve the following relaxed optimal transport model with estimated transport cost function and marginal distributions:

$$\begin{aligned} \min_{\boldsymbol{\mu}^c, \boldsymbol{\mu}^t, \mathbf{P}, \mathbf{T}^c, \mathbf{T}^t} \; & \langle \mathbf{D}^c, \mathbf{T}^c \rangle + \langle \mathbf{D}^t, \mathbf{T}^t \rangle \\ & + \lambda_c \langle \mathbf{D}^{c,c}, \mathbf{K}^c \rangle + \lambda_t \langle \mathbf{D}^{t,t}, \mathbf{K}^t \rangle \\ \text{s.t.} \; & \mathbf{P} \in \mathcal{M}, \mathbf{T}^c \in \mathcal{T}(\boldsymbol{\mu}^c, \boldsymbol{\mu}), \mathbf{T}^t \in \mathcal{T}(\boldsymbol{\mu}^t, \boldsymbol{\mu}). \end{aligned} \quad (23)$$

This learning model integrates reweighting and representation learning into the unified target of minimizing the Wasserstein discrepancy, in which marginal distributions are estimated as sample weights and the learnable cost function is associated with the balancing error.

4.4. Learning with Entropic Regularization

The above optimal transport model could induce a sparse solution, which means only a limited number of samples are transported, suffering from low data efficiency (Blondel et al., 2018; Vincent-Cuaz et al., 2022). Motivated by the (Cuturi, 2013), we apply a negative entropy regularization on the marginal distributions $\boldsymbol{\mu}^c$ and $\boldsymbol{\mu}^t$ to encourage more samples to be transported, and also avoid solving linear programming problems with heavy computation. The entropic regularizations are defined as

$$\Omega(\mathbf{T}^c) = \sum_{i=1}^{n_c} T^c_{i\cdot}(\log T^c_{i\cdot} - 1), \quad (24)$$

$$\Omega(\mathbf{T}^t) = \sum_{i=1}^{n_t} T^t_{i\cdot}(\log T^t_{i\cdot} - 1), \quad (25)$$

where $T^c_{i\cdot}$ is the sum of the i-th row of $\mathbf{T}^c$, where $T^t_{i\cdot}$ is the sum of the i-th column of $\mathbf{T}^t$, *i.e.*,

$$T^c_{i\cdot} = \sum_{j=1}^{n_t} T^c_{ij}, \qquad T^t_{i\cdot} = \sum_{j=1}^{n_c} T^t_{ij}. \quad (26)$$

Note that the entropic regularization here is different from the one in the Sinkhorn algorithm as mentioned in (Cuturi, 2013), which is on the joint distribution (*i.e.*, the optimal transport matrix $\mathbf{T}$) to induce a smooth $\mathbf{T}$.

In addition, we constrain $\mathbf{T}^c$ and $\mathbf{T}^t$ to belong to the following domains of the definition

$$\mathcal{T}^c = \{\mathbf{T}^c \mid (\mathbf{T}^c)^\top \mathbf{1}_{n_c} = \boldsymbol{\mu}, T^c_{ij} \in [0,1]\}, \quad (27)$$
$$\mathcal{T}^t = \{\mathbf{T}^t \mid (\mathbf{T}^t)^\top \mathbf{1}_{n_t} = \boldsymbol{\mu}, T^t_{ij} \in [0,1]\}, \quad (28)$$

which does not consider the constraints $\mathbf{T1} = \boldsymbol{\mu}^c$ and $\mathbf{T}^\top\mathbf{1} = \boldsymbol{\mu}^t$ explicitly, since $\boldsymbol{\mu}^c$ and $\boldsymbol{\mu}^t$ are also parameters to be optimized.

Finally, we obtain the following learning model

$$\begin{aligned}\min_{\mathbf{P},\mathbf{T}^c,\mathbf{T}^t} \ & \langle \mathbf{D}^c, \mathbf{T}^c\rangle + \langle \mathbf{D}^t, \mathbf{T}^t\rangle + \gamma_c\Omega(\mathbf{T}^c) + \gamma_t\Omega(\mathbf{T}^t) \\ & + \lambda_c\langle \mathbf{D}^{c,c}, \mathbf{K}^c\rangle + \lambda_t\langle \mathbf{D}^{t,t}, \mathbf{K}^t\rangle \\ \text{s.t. } & \mathbf{P} \in \mathcal{M}, \mathbf{T}^c \in \mathcal{T}^c, \mathbf{T}^t \in \mathcal{T}^t, \end{aligned} \quad (29)$$

where γ, λ_c and λ_t are the trade-off parameters.

The optimization algorithm is presented in the next section. After obtaining the solutions $\mathbf{T}^c$ and $\mathbf{T}^t$, the estimated marginal distributions $\{T^c_{i\cdot}\}_{i=1}^{n_c}$ and $\{T^t_{i\cdot}\}_{i=1}^{n_t}$ can be calculated by Eq. (26) and taken as the weights for control and treated samples, and ATE can be estimated by

$$\widehat{\text{ATE}} = \sum_{i=1}^{n_t} T^t_{i\cdot} y^t_i - \sum_{i=1}^{n_c} T^c_{i\cdot} y^c_i. \quad (30)$$

4.5. Analysis of Estimator

We now discuss the consistency and sample efficiency of our proposed ATE estimator.

In addition, we need the following assumptions to guarantee that the cost function exists and is continuously differentiable, and the measures are either subgaussian or defined on subsets of the real numbers (Dunipace, 2021).

Assumption 4.4. $\exists x_0 \in \mathcal{X} : \int_{\mathcal{X}} D(x_0, x) d\mu < \infty$.

Assumption 4.5. $D(\cdot,\cdot) \in \mathcal{C}^\infty$ and is L-Lipschitz, and either μ^z and μ are σ^2 - subgaussian with $D = ||\cdot||_2^2$ or $\mathcal{X} \in \mathbb{R}^d$.

Theorem 4.6. *Under the strong ignorability assumption, the estimated weights $\hat{w}^z$ are balancing weights, then our estimated $\widehat{ATE}$ is consistent, i.e., $\widehat{ATE} \to ATE$.*

Theorem 4.7. *Suppose Assumptions 4.4 and 4.5 hold, the weights $\hat{w}^z$ estimated by our proposed method converge at an $n^{-1/d}$-rate, which implies the sample efficiency of estimators based on optimal transport.*

The proofs of the above theorems are given in Appendix C.

5. Optimization

Problem (29) involves three groups of variables to optimize, *i.e.*, the projection matrix $\mathbf{P}$ for representation learning, the optimal transport matrices $\mathbf{T}^c$ and $\mathbf{T}^t$ for reweighting. We alternately update these groups of variables as follows.

5.1. Update P

Subproblem with respect to $\mathbf{P}$ is

$$\begin{aligned}\min_{\mathbf{P}} \ & \langle \mathbf{D}^c, \mathbf{T}^c\rangle + \langle \mathbf{D}^t, \mathbf{T}^t\rangle + \lambda_c\langle \mathbf{D}^c, \mathbf{K}^c\rangle + \lambda_t\langle \mathbf{D}^t, \mathbf{K}^t\rangle \\ \text{s.t. } & \mathbf{P} \in \mathcal{M}, \end{aligned} \quad (31)$$

in which the cost matrices $\mathbf{D}^c$, $\mathbf{D}^t$, $\mathbf{D}^{c,c}$ and $\mathbf{D}^{t,t}$ depend on the matrix $\mathbf{P}$. We rewrite the terms in the objective function, and provide a closed-form solution to this problem in the following proposition, whose proof is given in the appendix.

Proposition 5.1. *Problem (31) is equivalent to the following problem*

$$\begin{aligned}\min_{\mathbf{P}} \ & tr\Big(\mathbf{P}^\top(\boldsymbol{\Theta}^c + \boldsymbol{\Theta}^t + \lambda_c\boldsymbol{\Theta}^{c,c} + \lambda_t\boldsymbol{\Theta}^{t,t})\mathbf{P}\Big) \\ \textit{s.t. } & \mathbf{P}^\top\mathbf{P} = \mathbf{I}, \end{aligned} \quad (32)$$

where the matrices are defined as

$$\begin{aligned}\boldsymbol{\Theta}^c &= (\mathbf{X}^c)^\top diag(\mathbf{T}^c\mathbf{1})\mathbf{X}^c + (\mathbf{X})^\top diag((\mathbf{T}^c)^\top\mathbf{1})\mathbf{X} \\ &\quad - 2(\mathbf{X})^\top(\mathbf{T}^c)^\top\mathbf{X}^c, \quad (33)\\ \boldsymbol{\Theta}^t &= (\mathbf{X}^t)^\top diag(\mathbf{T}^t\mathbf{1})\mathbf{X}^t + (\mathbf{X})^\top diag((\mathbf{T}^t)^\top\mathbf{1})\mathbf{X} \\ &\quad - 2(\mathbf{X})^\top(\mathbf{T}^t)^\top\mathbf{X}^t, \quad (34)\\ \boldsymbol{\Theta}^{c,c} &= 2(\mathbf{X}^c)^\top(diag(\mathbf{K}^c\mathbf{1}) - \mathbf{K}^c)\mathbf{X}^c, \quad (35)\\ \boldsymbol{\Theta}^{t,t} &= 2(\mathbf{X}^t)^\top(diag(\mathbf{K}^t\mathbf{1}) - \mathbf{K}^t)\mathbf{X}^t. \quad (36)\end{aligned}$$

The closed-form solution to this problem is the first d' eigenvectors of the matrix $\boldsymbol{\Theta}^c + \boldsymbol{\Theta}^t + \lambda_c\boldsymbol{\Theta}^{c,c} + \lambda_t\boldsymbol{\Theta}^{t,t}$ with the smallest eigenvalues.

The proof is given in Appendix A.

5.2. Update $\mathbf{T}^c$ and $\mathbf{T}^t$

The subproblems with respect to $\mathbf{T}^c$ and $\mathbf{T}^t$ are given as follows

$$\min_{\mathbf{T}^c} \ \langle \mathbf{D}^c, \mathbf{T}^c\rangle + \gamma_c\Omega(\mathbf{T}^c) \quad \text{s.t. } \mathbf{T}^c \in \mathcal{T}^c, \quad (37)$$
$$\min_{\mathbf{T}^t} \ \langle \mathbf{D}^t, \mathbf{T}^t\rangle + \gamma_t\Omega(\mathbf{T}^t) \quad \text{s.t. } \mathbf{T}^t \in \mathcal{T}^t. \quad (38)$$

The two problems have similar forms and can be solved by similar methods. We develop a projected mirror descent algorithm (Nemirovskij & Yudin, 1983; Raskutti & Mukherjee, 2015) based on the KL divergence defined in Eq. (2) to solve the Problems (37) and (38) with respect to $\mathbf{T}^c$ and

$\mathbf{T}^t$, which are non-trivial to address because of the equality constraints. For simplicity, we define the objective functions of the two problems as

$$F(\mathbf{T}^\pi) = \langle \mathbf{D}^\pi, \mathbf{T}^\pi \rangle + \gamma_\pi \Omega(\mathbf{T}^\pi), \tag{39}$$

where $\pi \in \{c, t\}$ indicates the group. At the k-th iteration, we update $\mathbf{T}^\pi$ by solving the following problem

$$(\mathbf{T}^\pi)^{k+1} = \arg\min_{\mathbf{T}^\pi} \ \eta \langle \nabla F((\mathbf{T}^\pi)^k), \mathbf{T}^\pi \rangle + \mathcal{D}(\mathbf{T}^\pi || (\mathbf{T}^\pi)^k), \quad \text{s.t. } \mathbf{T}^\pi \in \mathcal{T}^\pi, \tag{40}$$

which firstly performs proximal gradient descent with the Bregman divergence (Banerjee et al., 2005) and the stepsize η, and then obtains a feasible solution in the set $\mathcal{T}^\pi$ by projection. Next, we show that each of these two operations has a closed-form solution.

5.2.1. Proximal Gradient Descent

Let $(\mathbf{\Upsilon}^\pi)^k$ be the solution to Problem (40), without considering the constraints $\mathbf{T}^\pi \in \mathcal{T}^\pi$, *i.e.*,

$$(\mathbf{\Upsilon}^\pi)^k = \arg\min_{\mathbf{T}^\pi} \eta \langle \nabla F((\mathbf{T}^\pi)^k), \mathbf{T}^\pi \rangle + \mathcal{D}(\mathbf{T}^\pi || (\mathbf{T}^\pi)^k). \tag{41}$$

By adopting the KL divergence defined in Eq. (2) as the Bregman divergence $\mathcal{D}(\mathbf{T}^\pi || (\mathbf{T}^\pi)^k)$, the closed-form solution to the above problem is given as

$$(\mathbf{\Upsilon}^\pi)^k = (\mathbf{T}^\pi)^k \odot \exp(-\eta \nabla F((\mathbf{T}^\pi)^k)), \tag{42}$$

where the gradient is calculated as

$$\nabla_{ij} F((\mathbf{T}^\pi)^k) = D^\pi_{ij} + \gamma_\pi \log T^\pi_{i\cdot}. \tag{43}$$

5.2.2. Projection Operation

To make sure the transport plans $(\mathbf{T}^\pi)^{k+1}$ and $(\mathbf{T}^\pi)^{k+1}$ satisfy the constraints in Eqs. (27) and (28), We update $(\mathbf{T}^\pi)^{k+1}$ by finding $\mathbf{T}^\pi \in \mathcal{T}^\pi$ which is most close to $(\mathbf{\Upsilon}^\pi)^k$ under the KL metric, which is achieved by solving the following projection problem

$$\min_{\mathbf{T}^\pi} \ \mathcal{D}(\mathbf{T}^\pi || (\mathbf{\Upsilon}^\pi)^k) := \sum_{i=1}^{n_\pi} \sum_{j=1}^{n} T^\pi_{ij} \log(\frac{T^\pi_{ij}}{(\Upsilon^\pi)^k_{ij}}) - T^\pi_{ij} + (\Upsilon^\pi)^k_{ij}$$
$$\text{s.t.} \ \ (\mathbf{T}^\pi)^\top \mathbf{1}_{n_\pi} = \boldsymbol{\mu}. \tag{44}$$

The closed-form solutions to the problems are given as

$$T^\pi_{ij} = \frac{(\Upsilon^\pi)^k_{ij}}{n \sum_{i=1}^{n_\pi} (\Upsilon^\pi)^k_{ij}}, \tag{45}$$

Algorithm 1 summarizes the whole procedure of our proposed method, named **O**ptimal transport for causal **I**nference by **C**ost **L**earning[1].

[1]Our code is available at https://github.com/ygyan/OICL.

Algorithm 1 Optimal transport for causal Inference by Cost Learning (OICL).

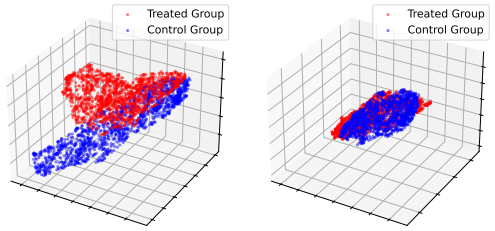
Input: The data matrices $\mathbf{X}^c$ and $\mathbf{X}^t$, and the corresponding outcomes $\mathbf{Y}^c$ and $\mathbf{Y}^t$.
Initialize $\mathbf{T}^c$ and $\mathbf{T}^t$: $T^c_{ij} = \frac{1}{n_c n}$, $T^t_{ij} = \frac{1}{n_t n}$.
repeat
  Update $\mathbf{P}$ according to Proposition (5.1).
  repeat
    Calculate $\mathbf{\Upsilon}^c$ according to Eq. (42).
    Update $\mathbf{T}^c$ according to Eq. (45).
  until Convergence.
  repeat
    Calculate $\mathbf{\Upsilon}^t$ according to Eq. (42).
    Update $\mathbf{T}^t$ according to Eq. (45).
  until Convergence.
until Convergence.
Obtain sample weights based on Eq. (26).
Estimate ATE according to Eq. (30).

6. Experiment

6.1. Experiment Setup

We compare the performance of OICL with the following methods: **IPW** (Rosenbaum & Rubin, 1983) estimates ATE via reweighting with the inverse of propensity scores. **DR** (Robins et al., 1994) estimates ATE with a combination of IPW and outcome regression model. **CBPS** (Imai & Ratkovic, 2014) exploits dual characteristics of the propensity score, which models treatment assignment while optimizing the covariate balance. **ARB** (Athey et al., 2018) combines weighting adjustment via directly balancing on confounders and regression adjustment on outcomes. **EBAL** (Hainmueller, 2012) estimates ATE by moment alignment with a maximum entropy scheme. We achieve EBAL(1) ensuring the first moment is balanced, and EBAL(2) ensuring the first and second moments are balanced. **CFR** (Shalit et al., 2017) learns a representation to balance the distributions of treated and control groups via Integral Probability Metric (IPM). Specifically, we use the Wasserstein distance as the IPM term for baseline. **OTW** (Dunipace, 2021) learns weights by minimizing the Sinkhorn divergence between treated and control groups, and adopts the LBFGS algorithm to solve. **CBIPM** (Kong et al., 2023) uses the kernel MMD method to achieve the smallest IPM value across treated and control groups. Both parametric CBIPM(P-CBIPM) and nonparametric CBIPM(N-CBIPM) algorithms are adopted. **ℓ_1-TCL** (Wei et al., 2023) trains a rough estimator first, and uses ℓ_1 regularization to correct the bias. **DKLITE** (Zhang et al., 2020) uses deep kernel regression algorithm and posterior regularization framework to estimate treatment effects.

For evaluating the performance of the conducted methods,



(a) Original covariates (b) Learned representations

Figure 1. Illustration results of the original covariates (left) and learned representations (right) based on UMAP. A red point means a treated sample while a blue point means a control sample.

we adopt the mean absolute errors (MAE) $|\widehat{ATE} - ATE|$ as the metric. We carry out the experiments 10 times and report the mean and standard deviation.

6.2. Results on Simulation Data

Following similar protocols in (Yao et al., 2018; Hatt & Feuerriegel, 2021), we conduct our simulation experiment in two different settings:

- For the setting of Gaussian distribution, we generate 1500 treated samples from $\mathcal{N}(\mu_t^{10\times1}, 0.5 \times \Sigma_t\Sigma_t^T)$ and 1500 control samples from $\mathcal{N}(\mu_c^{10\times1}, 0.5 \times \Sigma_c\Sigma_c^T)$, where $\Sigma_\cdot \sim \mathcal{U}((0, \mu_\cdot)^{10\times10})$. We fix μ_t to 0.5 and vary μ_c to simulate different confounding biases.
- For the setting of Non-Gaussian distribution, we generate data from Gaussian mixture distribution. We first generate two Gaussian distributions: $\mathcal{N}_1 = \mathcal{N}(0.5^{10\times1}, 0.5 \times \Sigma_1\Sigma_1^T)$, $\mathcal{N}_2 = \mathcal{N}(1^{10\times1}, 0.5 \times \Sigma_2\Sigma_2^T)$, where $\Sigma_1 \sim \mathcal{U}((0, 0.5)^{10\times10})$, $\Sigma_2 \sim \mathcal{U}((0, 1)^{10\times10})$. Then, we generate 1500 treated and control samples from $\mathbf{x}^t \sim \alpha_t\mathcal{N}_1 + (1-\alpha_t)\mathcal{N}_2$, $\mathbf{x}^c \sim \alpha_c\mathcal{N}_1 + (1-\alpha_c)\mathcal{N}_2$. We fix α_t to 0.5 and vary the value of α_c to simulate different selection bias.
- For the distributions above, the outcomes are both generated as $y = \sin(\mathbf{w}_1^\top\mathbf{x}) + \cos(\mathbf{w}_2^\top(\mathbf{x} \odot \mathbf{x})) + \mathrm{t} + \epsilon$, where $\mathbf{w}_\cdot \sim \mathcal{U}((0, 1)^{10\times1})$, $\epsilon \sim \mathcal{N}(0, 0.1)$.

The results are reported in Table 1. IPW, DR, CBPS, and ARB have limited performance, because these methods depend heavily on the correct specification of the propensity score or the conditional outcome regression models, which is usually difficult to obtain in practice. EBAL(2) performs better than EBAL(1), since it additionally uses the second moment as the distribution discrepancy metric. Although CFR takes some advantages of the neural network, it still performs not well in some complex settings without a large number of samples. OTW demonstrates competitive performance compared with other baselines, reflecting the superiority of the optimal transport technology. Based on optimal transport, OICL not only regards the optimal transport cost as the metric of distribution shift, but also deeply explores the underlying cost function by leveraging the guidance of factual outcomes, which brings significant improvements.

Additionally, to verify the role of the projection matrix $\mathbf{P}$ in cost learning which maps the original covariates into a balanced subspace, we further visualize the result using simulation data with $\mu_c = 1.2$. Specifically, we use Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) to find low-dimensional embeddings of the original covariates and learned representations induced from $\mathbf{P}$ in 3-dimensional spaces. The results are shown in Figure 1, in which different colors indicate binary treatments t. We observe that the learned representations are more overlapped compared with the original covariates, suggesting the effectiveness of the projection matrix $\mathbf{P}$ for distribution balancing, which can significantly reduce confounding bias.

6.3. Results on Real-world Data

We also conduct experiments on two real-world datasets, including LaLonde and Twins. **LaLonde**[2] consists of two parts. The first part comes from the RCT (NSW), and we replace the control group in NSW with another control group from the observational data (PSID3) in the second part. The treatment is whether the participant attends the job training program, and the outcome is the earning in the year 1978. The data contains 8 covariates. **Twins** is collected from the twins born in USA between 1989-1991 (Almond et al., 2005). Each twin pair has 30 covariates. For each twin pair, we observe both the cases $t = 0$ (lighter) and $t = 1$ (heavier). The outcome is the one-year mortality. To simulate the confounding bias, we choose one of the twins as follows: $t \sim Bern(\mathrm{sigmoid}(\mathbf{w}^\top\mathbf{x} + b))$ where $\mathbf{w} \sim \mathcal{U}((-0.1, 1)^{30\times1})$ and $b \sim \mathcal{N}(0, 0.1)$. Infant Health and Development Program (**IHDP**) aims to study the treatment effect of specialist home visits on infants' future cognitive test scores. Each pair comprises 25 covariates that measure aspects of children and their mothers. We consider setting "A" in the NPCI package (Dorie, 2016).

The results are reported in Table 2. The performance of IPW and ARB is relatively lower, while CBPS achieves better performance by exploiting the dual characteristic of propensity score. DR performs well in Twins and in LaLonde compared with other baselines, possibly because the specified linear parametric methods are suitable. EBAL achieves modest performance in Twins but performs badly in Lalonde, possibly because the low-order moments are insufficient to balance the complex distribution in LaLonde. OICL performs best by exploiting optimal transport with

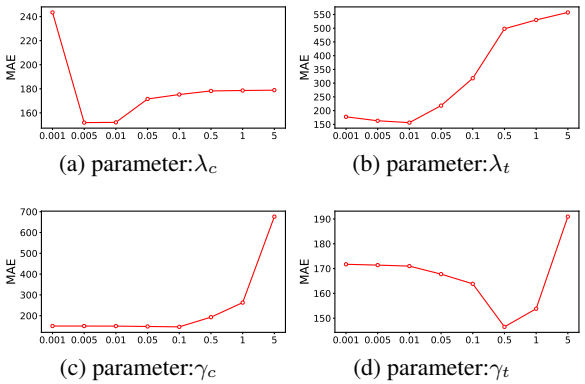
[2]https://users.nber.org/ rdehejia/data/.nswdata2.html

Table 1. Result on simulation data. We report mean and standard deviation of MAE and highlight the best result in bold.

	Gaussian				Non-Gaussian			
	$\mu_c = 0.6$	$\mu_c = 0.8$	$\mu_c = 1.0$	$\mu_c = 1.2$	$\alpha_c = 0.4$	$\alpha_c = 0.3$	$\alpha_c = 0.2$	$\alpha_c = 0.1$
IPW	0.1242 ± 0.0492	0.2317 ± 0.0425	0.2407 ± 0.0621	0.2035 ± 0.1050	0.0387 ± 0.0240	0.0493 ± 0.0325	0.0583 ± 0.0422	0.0866 ± 0.0499
DR	0.1122 ± 0.0455	0.1874 ± 0.0457	0.1669 ± 0.0722	0.0858 ± 0.0692	0.0387 ± 0.0241	0.0492 ± 0.0324	0.0572 ± 0.0414	0.0823 ± 0.0484
CBPS	0.1289 ± 0.0508	0.2223 ± 0.0596	0.2209 ± 0.0633	0.1832 ± 0.0825	0.0404 ± 0.0249	0.0517 ± 0.0332	0.0620 ± 0.0464	0.0930 ± 0.0568
ARB	0.1219 ± 0.0503	0.2146 ± 0.0536	0.1905 ± 0.0700	0.0924 ± 0.0934	0.0388 ± 0.0254	0.0502 ± 0.0340	0.0593 ± 0.0439	0.0872 ± 0.0503
EBAL(1)	0.1234 ± 0.0742	0.2320 ± 0.1045	0.2168 ± 0.1111	0.2036 ± 0.1521	0.0386 ± 0.0240	0.0493 ± 0.0323	0.0574 ± 0.0413	0.0846 ± 0.0467
EBAL(2)	0.0882 ± 0.0558	0.1557 ± 0.0749	0.1949 ± 0.0928	0.1989 ± 0.1134	0.0367 ± 0.0248	0.0397 ± 0.0285	0.0452 ± 0.0309	0.0603 ± 0.0410
CFR	0.0522 ± 0.0424	0.1726 ± 0.0742	0.2374 ± 0.0699	0.2683 ± 0.1006	0.0799 ± 0.0615	0.0854 ± 0.0942	0.0784 ± 0.0713	0.0625 ± 0.0406
OTW	0.0811 ± 0.0824	0.0986 ± 0.0902	0.1249 ± 0.1199	0.1387 ± 0.0976	0.1092 ± 0.0878	0.0891 ± 0.0813	0.0837 ± 0.0824	0.0762 ± 0.0764
P-CBIPM	0.1248 ± 0.0511	0.2187 ± 0.0524	0.2056 ± 0.0577	0.1646 ± 0.0763	0.0410 ± 0.0270	0.0581 ± 0.0379	0.0736 ± 0.0470	0.1087 ± 0.0533
N-CBIPM	0.0360 ± 0.0350	0.1204 ± 0.0510	0.1217 ± 0.0563	**0.1100 ± 0.0702**	0.0301 ± 0.0288	0.0458 ± 0.0320	0.0495 ± 0.0367	0.0568 ± 0.0412
ℓ_1-TCL	0.1269 ± 0.0489	0.2258 ± 0.0456	0.2148 ± 0.0624	0.1448 ± 0.0941	0.0420 ± 0.0255	0.0543 ± 0.0336	0.0640 ± 0.0435	0.0902 ± 0.0529
DKLITE	0.0408 ± 0.0278	0.0918 ± 0.0536	0.1761 ± 0.0716	0.1267 ± 0.1295	0.0306 ± 0.0160	0.0301 ± 0.0227	0.0336 ± 0.0198	0.0511 ± 0.0324
OICL	**0.0272 ± 0.0212**	**0.0720 ± 0.0358**	**0.1125 ± 0.0570**	0.1178 ± 0.0932	**0.0138 ± 0.0177**	**0.0168 ± 0.0168**	**0.0207 ± 0.0187**	**0.0292 ± 0.0185**

Table 2. Result on real-world data. We report the mean and standard deviation of MAE and highlight the best result in bold.

	Twins(1e-2)	Lalonde	IHDP
IPW	0.3456 ± 0.2156	414.0558 ± 231.6505	0.2156 ± 0.0997
DR	0.1740 ± 0.1209	211.1198 ± 143.7868	0.1262 ± 0.1343
CBPS	0.2263 ± 0.1187	248.9927 ± 272.2510	0.0931 ± 0.0686
ARB	0.1725 ± 0.1238	300.1334 ± 190.0572	0.1160 ± 0.0789
EBAL(1)	0.2249 ± 0.1491	593.1440 ± 205.0838	0.1260 ± 0.1338
EBAL(2)	0.2051 ± 0.1895	549.1707 ± 380.8954	0.2413 ± 0.3902
CFR	0.4539 ± 0.2115	439.6378 ± 376.2243	0.3016 ± 0.3049
OTW	0.6688 ± 0.4421	730.1346 ± 197.5947	0.1429 ± 0.1176
P-CBIPM	0.2017 ± 0.1768	148.2635 ± 89.1910	0.0904 ± 0.0610
N-CBIPM	0.1738 ± 0.1625	179.5005 ± 179.4790	0.0864 ± 0.0622
ℓ_1-TCL	0.3239 ± 0.3120	393.2358 ± 245.1366	0.1208 ± 0.1277
DKLITE	0.2074 ± 0.1327	335.9644 ± 259.2227	0.0874 ± 0.0566
OICL	**0.1674 ± 0.1733**	**146.3449 ± 141.7637**	**0.0698 ± 0.0617**



(a) parameter:λ_c (b) parameter:λ_t

(c) parameter:γ_c (d) parameter:γ_t

Figure 2. Results of varying values of parameters on LaLonde.

learnable sample weights and the cost function.

6.4. Sensitivity Analysis

We take LaLonde as an example to evaluate the effects of the parameters of our model. We vary the parameters λ_c, λ_t, γ_c, γ_t in Eqs. (29), from 10^{-3} to 5 and plot the result in Figure 1. From Figures 2a and 2b, we observe that MAE increases as λ_c and λ_t become too large or too small, indicating that moderate strength of the factual outcome guidance term can successfully guide the cost function learning and achieve good performance. From Figures 2c and 2d, we have a similar observation since a large strength of entropic regularization will push the learned weights close to the uniform distribution, while a small strength of the regularization makes the transport too sparse.

7. Conclusion

In this paper, we address confounding bias in causal inference by investigating the connection between the balancing error and optimal transport. We show that the balancing error can be reduced by minimizing the Wasserstein discrepancies with learnable marginal distributions and the underlying cost function, which is associated with our defined balancing error. In specific, we incorporate the potential outcomes and propose a learning problem that unifies reweighting and representation learning. Our learning model provides additional possibilities to exploit optimal transport for causal inference with different designs of the balancing error, such as one considering propensity scores.

Impact Statement

This research advances the fields of causal effect estimation and optimal transport by constructing a theoretical connection between them, which motivates us to develop a causal effect estimation method under the framework of optimal transport. Our method could be applied to a wide range of applications, such as decision-making in healthcare, public policy, business, *etc*.

Acknowledgements

This research was supported in part by National Key R&D Program of China (2021ZD0111501), Natural Science Foundation of China (62206061, 62206064), Guangdong Basic and Applied Basic Research Foundation (2024A1515011901), Guangzhou Basic and Applied Basic Research Foundation (2023A04J1700), National Science Fund for Excellent Young Scholars (62122022), and CCF-DiDi GAIA Collaborative Research Funds (CCF-DiDi GAIA 202311).

References

Almond, D., Chay, K. Y., and Lee, D. S. The costs of low birth weight. *The Quarterly Journal of Economics*, 120 (3):1031–1083, 2005.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Assaad, S., Zeng, S., Tao, C., Datta, S., Mehta, N., Henao, R., Li, F., and Carin, L. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pp. 1972–1980. PMLR, 2021.

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Athey, S., Imbens, G. W., and Wager, S. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4): 597–623, 2018.

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. Clustering with bregman divergences. *Journal of machine learning research*, 6(10), 2005.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Blondel, M., Seguy, V., and Rolet, A. Smooth and sparse optimal transport. In *International conference on artificial intelligence and statistics*, pp. 880–889. PMLR, 2018.

Concato, J., Shah, N., and Horwitz, R. I. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England journal of medicine*, 342(25):1887–1892, 2000.

Courty, N., Flamary, R., and Tuia, D. Domain adaptation with regularized optimal transport. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 274–289, 2014.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation. In *Annual Conference on Neural Information Processing Systems*, pp. 3733–3742, 2017a.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (9):1853–1865, 2017b.

Curth, A. and Van Der Schaar, M. In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. In *International Conference on Machine Learning*, pp. 6623–6642. PMLR, 2023.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Annual Conference on Neural Information Processing Systems*, pp. 2292–2300, 2013.

Dorie, V. Npci: Non-parametrics for causal inference. *URL: https://github. com/vdorie/npci*, 11:23, 2016.

Dudley, R. M. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1): 40–50, 1969.

Dunipace, E. Optimal transport weights for causal inference. *arXiv preprint arXiv:2109.01991*, 2021.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1574–1583. PMLR, 2019.

Gunsilius, F. and Xu, Y. Matching for causal effects via multimarginal unbalanced optimal transport. *arXiv preprint arXiv:2112.04398*, 2021.

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.

Hatt, T. and Feuerriegel, S. Estimating average treatment effects via orthogonal regularization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 680–689, 2021.

Hernán, M. A. and Robins, J. M. Causal inference, 2010.

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

Johansson, F. D., Kallus, N., Shalit, U., and Sontag, D. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. Generalization bounds and representation learning for estimation of potential outcomes and causal effects, 2021.

Kantorovitch, L. On the translocation of masses. *Management Science*, 5(1):1–4, 1958.

Kong, I., Park, Y., Jung, J., Lee, K., and Kim, Y. Covariate balancing using the integral probability metric for causal inference. In *International Conference on Machine Learning*, pp. 17430–17461. PMLR, 2023.

Kuang, K., Cui, P., Li, B., Jiang, M., and Yang, S. Estimating treatment effect in the wild via differentiated confounder balancing. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 265–274, 2017.

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.

Li, Q., Wang, Z., Liu, S., Li, G., and Xu, G. Causal optimal transport for treatment effect estimation. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Machlanski, D., Samothrakis, S., and Clarke, P. Hyperparameter tuning and model evaluation in causal effect estimation. *arXiv preprint arXiv:2303.01412*, 2023.

Mahajan, D., Mitliagkas, I., Neal, B., and Syrgkanis, V. Empirical analysis of model selection for heterogeneous causal effect estimation. *arXiv preprint arXiv:2211.01939*, 2022.

Maretic, H. P., El Gheche, M., Chierchia, G., and Frossard, P. Fgot: Graph distances based on filters and optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7710–7718, 2022.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Monge, G. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.

Peyré, G., Cuturi, M., and Solomon, J. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pp. 2664–2672, 2016.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Raskutti, G. and Mukherjee, S. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

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.

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.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Rubner, Y., Tomasi, C., and Guibas, L. J. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

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.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pp. 465–472, 1990.

Titouan, V., Courty, N., Tavenard, R., and Flamary, R. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pp. 6275–6284. PMLR, 2019.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schölkopf, B. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Torous, W., Gunsilius, F., and Rigollet, P. An optimal transport approach to causal inference. *arXiv preprint arXiv:2108.05858*, 2021.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Villani, C. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., and Courty, N. Semi-relaxed gromov wasserstein divergence with applications on graphs. In *International Conference on Learning Representations*, 2022.

Wang, H., Chen, Z., Fan, J., Li, H., Liu, T., Liu, W., Dai, Q., Wang, Y., Dong, Z., and Tang, R. Optimal transport for treatment effect estimation. *arXiv preprint arXiv:2310.18286*, 2023.

Wei, S., Kong, X., Huestis-Mitchell, S. A., Xie, Y., Zhu, S., Xavier, A. S., and Qiu, F. Unfairness detection within power systems through transfer counterfactual learning. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023.

Xu, H. Gromov-wasserstein factorization models for graph clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6478–6485, 2020.

Yan, Y., Tan, M., Xu, Y., Cao, J., Ng, M., Min, H., and Wu, Q. Oversampling for imbalanced data via optimal transport. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 5605–5612, 2019.

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.

Zhang, Y., Bellot, A., and Schaar, M. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pp. 1005–1014. PMLR, 2020.

Zhao, P. and Zhou, Z.-H. Label distribution learning by optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

A. Proof of Proposition 5.1

Subproblem with respect to $\mathbf{P}$ is

$$\min_{\mathbf{P}} \ \langle \mathbf{D}^c, \mathbf{T}^c \rangle + \langle \mathbf{D}^t, \mathbf{T}^t \rangle + \lambda_c \langle \mathbf{D}^c, \mathbf{K}^c \rangle + \lambda_t \langle \mathbf{D}^t, \mathbf{K}^t \rangle \quad \text{s.t. } \mathbf{P} \in \mathcal{M}, \tag{46}$$

in which the cost matrices $\mathbf{D}^c$, $\mathbf{D}^t$, $\mathbf{D}^{c,c}$ and $\mathbf{D}^{t,t}$ depend on the matrix $\mathbf{P}$. We rewrite the terms in the objective function, and provide the closed-form solution to this problem.

Proposition 5.1. *Problem (31) is equivalent to the following problem*

$$\min_{\mathbf{P}} \ tr\Big(\mathbf{P}^\top(\mathbf{\Theta}^c + \mathbf{\Theta}^t + \lambda_c \mathbf{\Theta}^{c,c} + \lambda_t \mathbf{\Theta}^{t,t})\mathbf{P}\Big) \quad s.t. \ \mathbf{P}^\top\mathbf{P} = \mathbf{I}, \tag{47}$$

where the matrices are defined as

$$\mathbf{\Theta}^c = (\mathbf{X}^c)^\top diag(\mathbf{T}^c\mathbf{1})\mathbf{X}^c + (\mathbf{X})^\top diag((\mathbf{T}^c)^\top\mathbf{1})\mathbf{X} - 2(\mathbf{X})^\top(\mathbf{T}^c)^\top\mathbf{X}^c, \tag{48}$$

$$\mathbf{\Theta}^t = (\mathbf{X}^t)^\top diag(\mathbf{T}^t\mathbf{1})\mathbf{X}^t + (\mathbf{X})^\top diag((\mathbf{T}^t)^\top\mathbf{1})\mathbf{X} - 2(\mathbf{X})^\top(\mathbf{T}^t)^\top\mathbf{X}^t, \tag{49}$$

$$\mathbf{\Theta}^{c,c} = 2(\mathbf{X}^c)^\top(diag(\mathbf{K}^c\mathbf{1}) - \mathbf{K}^c)\mathbf{X}^c, \tag{50}$$

$$\mathbf{\Theta}^{t,t} = 2(\mathbf{X}^t)^\top(diag(\mathbf{K}^t\mathbf{1}) - \mathbf{K}^t)\mathbf{X}^t. \tag{51}$$

The closed-form solution to this problem is the first d' eigenvectors of the matrix $\mathbf{\Theta}^c + \mathbf{\Theta}^t + \lambda_c\mathbf{\Theta}^{c,c} + \lambda_t\mathbf{\Theta}^{t,t}$ with the smallest eigenvalues.

Proof. First of all, we rewrite the transport cost between $\mathbf{x}_i^c$ and $\mathbf{x}_j$ as

$$D_{ij}^c = \|\mathbf{P}^\top\mathbf{x}_i^c - \mathbf{P}^\top\mathbf{x}_j\|_2^2 = \|\mathbf{P}^\top\mathbf{x}_i^c\|_2^2 + \|\mathbf{P}^\top\mathbf{x}_j\|_2^2 - 2\langle\mathbf{P}^\top\mathbf{x}_i^c, \mathbf{P}^\top\mathbf{x}_j\rangle. \tag{52}$$

Therefore, we have

$$\begin{aligned}
&\langle \mathbf{D}^c, \mathbf{T}^c \rangle \\
=&\sum_{i=1}^{n_c}\sum_{j=1}^{n} D_{ij}^c T_{ij}^c, \\
=&\sum_{i=1}^{n_c}\Big(\|\mathbf{P}^\top\mathbf{x}_i^c\|_2^2\Big)\sum_{j=1}^{n} T_{ij}^c + \sum_{j=1}^{n}\Big(\|\mathbf{P}^\top\mathbf{x}_j\|_2^2\Big)\sum_{i=1}^{n_c} T_{ij}^c \\
&\quad - 2\sum_{i=1}^{n_c}\sum_{j=1}^{n}\Big(\langle\mathbf{P}^\top\mathbf{x}_i^c, \mathbf{P}^\top\mathbf{x}_j\rangle\Big)T_{ij}^c \\
=&\langle(\mathbf{X}^c\mathbf{P})(\mathbf{X}^c\mathbf{P})^\top, \mathrm{diag}(\mathbf{T}^c\mathbf{1})\rangle + \langle(\mathbf{X}\mathbf{P})(\mathbf{X}\mathbf{P})^\top, \\
&\quad \mathrm{diag}((\mathbf{T}^c)^\top\mathbf{1})\rangle - 2\langle(\mathbf{X}^c\mathbf{P})(\mathbf{X}\mathbf{P})^\top, \mathbf{T}^c\rangle \\
=&\mathrm{tr}\big(\mathbf{P}^\top(\mathbf{X}^c)^\top\mathrm{diag}(\mathbf{T}^c\mathbf{1})\mathbf{X}^c\mathbf{P}\big) \\
&\quad + \mathrm{tr}\big(\mathbf{P}^\top\mathbf{X}^\top\mathrm{diag}((\mathbf{T}^c)^\top\mathbf{1})\mathbf{X}\mathbf{P}\big) \\
&\quad - 2\mathrm{tr}\big(\mathbf{P}^\top\mathbf{X}^\top(\mathbf{T}^c)^\top\mathbf{X}^c\mathbf{P}\big).
\end{aligned} \tag{53}$$

Similarly, we have

$$\begin{aligned}
\langle \mathbf{D}^t, \mathbf{T}^t \rangle =& \mathrm{tr}\big(\mathbf{P}^\top(\mathbf{X}^t)^\top\mathrm{diag}(\mathbf{T}^t\mathbf{1})\mathbf{X}^t\mathbf{P}\big) \\
&+ \mathrm{tr}\big(\mathbf{P}^\top\mathbf{X}^\top\mathrm{diag}((\mathbf{T}^t)^\top\mathbf{1})\mathbf{X}\mathbf{P}\big) \\
&- 2\mathrm{tr}\big(\mathbf{P}^\top\mathbf{X}^\top(\mathbf{T}^t)^\top\mathbf{X}^t\mathbf{P}\big).
\end{aligned} \tag{54}$$

For the symmetric and non-negative matrices $\mathbf{K}^c$ and $\mathbf{K}^t$ in Problem (31), based on the property of the Laplacian matrix, we have

$$\begin{aligned}
\langle \mathbf{D}^{c,c}, \mathbf{K}^c \rangle &= \sum_{i=1}^{n_c}\sum_{j=1}^{n_c}\|\mathbf{P}^\top\mathbf{x}_i^c - \mathbf{P}^\top\mathbf{x}_j^c\|_2^2 K_{ij}^c \\
&= 2\mathrm{tr}\Big(\mathbf{P}^\top(\mathbf{X}^c)^\top(\mathrm{diag}(\mathbf{K}^c\mathbf{1}) - \mathbf{K}^c)\mathbf{X}^c\mathbf{P}\Big),
\end{aligned} \tag{55}$$

$$\begin{aligned}
\langle \mathbf{D}^{t,t}, \mathbf{K}^t \rangle &= \sum_{i=1}^{n_t}\sum_{j=1}^{n_t}\|\mathbf{P}^\top\mathbf{x}_i^t - \mathbf{P}^\top\mathbf{x}_j^t\|_2^2 K_{ij}^t \\
&= 2\mathrm{tr}\big(\mathbf{P}^\top(\mathbf{X}^t)^\top(\mathrm{diag}(\mathbf{K}^t\mathbf{1}) - \mathbf{K}^t)\mathbf{X}^t\mathbf{P}\big).
\end{aligned} \tag{56}$$

Combining Eqs. (53), (54), (55), (56), the proposition can be proved immediately. □

B. Computational Complexity Analysis

Let n_c, n_t, n be the numbers of control samples, treated samples, all the samples, respectively. And d, d' are the numbers of features before and after projection, respectively. For each iteration of the outer loop in Algorithm 1, in Line 3, the complexity of updating $\mathbf{P} \in \mathbb{R}^{d\times d'}$ is $O(n^2d + d^3)$, and the complexity of calculating $\mathbf{D}^c$ and $\mathbf{D}^t$ is $O(ndd' + n^2d')$. For the update of $\mathbf{T}^c$ and $\mathbf{T}^t$ in Lines 4 - 11, the complexity is $O(t_c n n_c + t_t n n_t)$, where t_c and t_t are the numbers of iterations to repeatedly perform Eqs. (42) and (45) for control and treated groups, respectively. Since t_c and t_t are much smaller compared with n_c and n_t, the complexity of obtaining $\mathbf{T}^c$ and $\mathbf{T}^t$ is $O(n^2)$. Therefore, the complexity of each iteration of the outer loop in Lines 2 to 12 is $O(n^2d + ndd' + d^3)$. After iteration, in Lines 13 and 14, the complexities of Eqs. (26) and (29) are $O(n^2)$ and $O(n)$, respectively.

Moreover, we empirically evaluate the running time of all methods on synthetic data. Hardware used in this experiment are: CPU: Intel i5-12600K, GPU: NVIDIA GeForce RTX 4090. The result shows in Table 3.

C. Consistency and Sample Efficiency

Our proposed ATE estimator is consistent, and the weights learned by optimal transport (KP) will converge at a $n^{-1/d}$-rate (d is the feature dimensions), which implies the sample efficiency of estimators based on KP.

Table 3. Running time on synthetic data (sec.)

IPW	DR	CBPS	ARB	EBAL(1)	EBAL(2)	CFR	OTW	P-CBIPM	N-CBIPM	ℓ_1-TCL	DKLITE	OICL
0.4	2.8	70.3	23.8	14.0	15.4	116.8	727.8	9.2	10.5	303.5	13.2	19.2

To prove that, we first give the notations required to prove. Let $z = \{t, c\}$ to denote the treatment or control group indicator, μ_n^z and μ_n to denote the empirical distribution for group z and the full population, μ^z and μ to denote the corresponding population distributions.

After that, we define the importance sampling weights for group z as $w_i^{z,*} = \frac{\tilde{w}_i^{z,*}}{\sum_i \tilde{w}_i^{z,*}}$, $\tilde{w}_i^{z,*} = \frac{d\mu(x)}{d\mu^z(x)}$. Also, we have the following Lemma C.1 about the convergence of $w^{z,*}$ as shown in (Dunipace, 2021).

Lemma C.1. *Under the strong ignorability assumption, the importance sampling weights $w^{z,*}$ converges to μ, i.e., $w^{z,*} \to \mu$.*

Based on Lemma C.1, we can derive that the weights learned from optimal transport $\hat{w}^z$ also converges to μ. Specifically, based on definition of the estimated weights $\hat{w}^z = \arg\min_{w^z} KP(w^z, \mu_n)$, we have $KP(\hat{w}^z, \mu_n) \leqslant KP(w^{z,*}, \mu_n)$. When $n \to \infty$, we have $w^{z,*} \to \mu, KP(w^{z,*}, \mu_n) \to 0 \Rightarrow KP(\hat{w}^z, \mu_n) \to 0, \hat{w}^z \to \mu$.

Theorem C.2. *Under the strong ignorability assumption, the estimated weights $\hat{w}^z$ are balancing weights, then our estimated $\widehat{ATE}$ is consistent, i.e., $\widehat{ATE} \to ATE$.*

Proof. We firstly show that the estimated weights $\hat{w}^z$ are balancing weights. According to the Theorem 4.1 in our paper, we use balancing error err_m^w to capture the covariate balancing degree, and we have: $err_m^w \leqslant KP(\hat{w}^t, \mu_n) + KP(\hat{w}^c, \mu_n) \to 0$, which shows when n is large enough, the weights $\hat{w}^z$ learned by KP is balancing weights that balance the weighted covariates distributions of between groups, i.e., $P^t(x)\hat{w}^t(x) = P^c(x)\hat{w}^c(x) = P(x)$.

Therefore, according to the Theorem 1 in (Li et al., 2018), the estimated ATE based on balancing weights is consistent, i.e., $\widehat{ATE} \to ATE$. □

Theorem C.3. *Suppose Assumptions 4.4 and 4.5 hold, the weights $\hat{w}^z$ estimated by our proposed method converge at an $n^{-1/d}$-rate, which implies the sample efficiency of estimators based on optimal transport.*

Proof. Based on definition of $\hat{w}^z$ and property of KP, we have $KP(\mu_n, \mu_n) \leqslant KP(\hat{w}^z, \mu_n) \leqslant KP(w^{z,*}, \mu_n)$. If we add $-KP(\mu, \mu)$ to each term, we obtain $KP(\mu_n, \mu_n) - KP(\mu, \mu) \leqslant KP(w^z, \mu_n) - KP(\mu, \mu) \leqslant KP(w^{z,*}, \mu_n) - KP(\mu, \mu)$. Under the assumption 4.4 and 4.5, as stated in (Dudley, 1969; Genevay et al., 2019), we have $\mathbb{E}[KP(\mu_n, \mu_n) - KP(\mu, \mu)] = \mathcal{O}\left(\frac{1}{n^{1/d}}\right)$ and $\mathbb{E}\left[KP\left(w^{z,*}, \mu_n\right) - KP(\mu, \mu)\right] = \mathcal{O}\left(\frac{1}{n^{1/d}}\right)$. Thus, we have $\mathbb{E}\left[KP\left(\hat{w}^z, \mu_n\right) - KP(\mu, \mu)\right] = \mathcal{O}\left(\frac{1}{n^{1/d}}\right)$. □

D. The balance constraint $m(\cdot)$

$m(\cdot)$ is a function that captures the information of data samples. We do not restrict the form of $m(\cdot)$ as long as the balancing error $err_m^{\mathbf{w}}$ defined on $m(\cdot)$ can characterize the degree of confounding bias in some sense.

Besides factual outcome, we can also introduce propensity scores into the function $m(\cdot)$ to reduce the balancing error. Based on this, the affinity matrices $\mathbf{K}^c$ and $\mathbf{K}^t$ can be constructed based on the propensity scores to learn the cost function. We empirically evaluate the performance based on propensity scores and report the results in Table 4. We observe that our algorithm with propensity scores also achieves promising performance.

E. Convergence Analysis

The subproblems with respect to $\mathbf{T}^c$ and $\mathbf{T}^t$ in Eqs. (37) and (38) are convex, while the subproblem with respect to $\mathbf{P}$ in Eq. (31) is non-convex since the constraint $\mathbf{P}^\top\mathbf{P} = \mathbf{I}$ is not convex.

The objective value of Problem (29) is convergent. For the subproblem with respect to $\mathbf{T}^c$ and $\mathbf{T}^t$, we can obtain a function value convergence from (Benamou et al., 2015; Peyré et al., 2016). In addition, the subproblem with respect to $\mathbf{P}$ has a closed-form solution, we obtain that the objective value is non-increasing during iteration. Combining the condition that the objective value is lower-bounded, we can obtain the convergence of the objective value by using the monotone convergence theorem.

We also empirically evaluate the convergence result of our algorithm. At the k-th iteration, we report $\Delta(\mathbf{T}^c)^{k+1} = \sum_{i,j} \mid (T_{ij}^c)^{k+1} - (T_{ij}^c)^k \mid$ as example to show the convergence, as well as the objective function in Eq. (29). The result is shown in Figure 3.

F. Extension to Estimate ATT or ITE

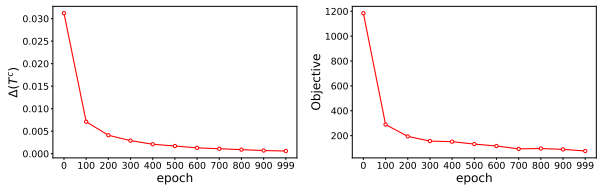
It is possible to extend our algorithm to estimate ATT (average treatment effect on the treated group) or ITE (individual treatment effect). For ATT, we can fix the weights of treated samples with $\frac{1}{n_t}$, and reduce the balancing error by the

Table 4. The performance based on propensity scores.

μ_c=0.6	μ_c=0.8	μ_c=1.0	μ_c=1.2	α_c=0.4	α_c=0.3	α_c=0.2	α_c=0.1	Twins(1e-2)	Lalonde	IHDP
0.0273 ± 0.0211	0.0720 ± 0.0360	0.1125 ± 0.0570	0.1178 ± 0.0932	0.0138 ± 0.0177	0.0168 ± 0.0168	0.0208 ± 0.0185	0.0292 ± 0.0186	0.1674 ± 0.1733	167.8991 ± 97.3491	0.1266 ± 0.0931

Table 5. Confidence Interval Results.

μ_c=0.6	μ_c=0.8	μ_c=1.0	μ_c=1.2	α_c=0.4	α_c=0.4	α_c=0.4	α_c=0.4	Twins(1e-2)	Lalonde	IHDP
(0.0162, 0.0423)	(0.0500, 0.0926)	(0.0819, 0.1449)	(0.0765,0.1885)	(0.0064, 0.0277)	(0.0102, 0.0297)	(0.0125, 0.0330)	(0.0195,0.0407)	(0.0863, 0.2887)	(68.2025, 231.5016)	(0.0433, 0.1178)



(a) $\Delta(\mathbf{T}^c)$ (b) Objective

Figure 3. Empirically Convergence Results

following model to learn weights for control samples:

$$\min_{\boldsymbol{\mu}^c,\phi} \mathcal{W}(\mathbf{X}^c, \mathbf{X}^t, \phi). \tag{57}$$

After reweighting the control samples based on $\boldsymbol{\mu}^c$, we can estimate ATT by

$$\widehat{\text{ATT}} = \sum_{i=1}^{n_t} \frac{1}{n_t} y_i^t - \sum_{i=1}^{n_c} \mu_i^c y_i^c. \tag{58}$$

For ITE, we can leverage the weights learned by our model to train a reweighting regression model to predict the counterfactual outcome as the approach in (Assaad et al., 2021).

G. Confidence Interval

For the confidence interval, we use bootstrap to calculate 95% confidence interval of each dataset with 500 times resamples to form the bootstrap distribution. The results are shown in Table 5.