Exploiting Geometry for Treatment Effect Estimation via Optimal Transport

Yuguang Yan[1], Zeqin Yang[1], Weilin Chen[1], Ruichu Cai[1,2*], Zhifeng Hao[3], Michael Kwok-Po Ng[4]

[1]School of Computer Science, Guangdong University of Technology, Guangzhou, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]College of Science, Shantou University, Shantou, China
[4]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
ygyan@gdut.edu.cn, yangzeqin1999@163.com, {chenweilin.chn,cairuichu}@gmail.com,
haozhifeng@stu.edu.cn, michael-ng@hkbu.edu.hk

Abstract

Estimating treatment effects from observational data suffers from the issue of confounding bias, which is induced by the imbalanced confounder distributions between the treated and control groups. As an effective approach, re-weighting learns a group of sample weights to balance the confounder distributions. Existing methods of re-weighting highly rely on a propensity score model or moment alignment. However, for complex real-world data, it is difficult to obtain an accurate propensity score prediction. Although moment alignment is free of learning a propensity score model, accurate estimation for high-order moments is computationally difficult and still remains an open challenge, and first and second-order moments are insufficient to align the distributions and easy to be misled by outliers. In this paper, we exploit geometry to capture the intrinsic structure involved in data for balancing the confounder distributions, so that confounding bias can be reduced even with outliers. To achieve this, we construct a connection between treatment effect estimation and optimal transport, a powerful tool to capture geometric information. After that, we propose an optimal transport model to learn sample weights by extracting geometry from confounders, in which geometric information between groups and within groups is leveraged for better confounder balancing. A projected mirror descent algorithm is employed to solve the derived optimization problem. Experimental studies on both synthetic and real-world datasets demonstrate the effectiveness of our proposed method.

Introduction

Treatment effect estimation aims to predict the effect of a treatment and has been widely used in real-world applications, such as public health (Glass et al. 2013) and advertisement (Li et al. 2016). The ideal approach for estimating treatment effect is to perform Randomized Controlled Trials (RCTs), which means that samples are randomly assigned to the treated group and the control group, and the treatment effect can be evaluated by comparing the outcomes of the two groups. Nevertheless, RCTs are usually expensive or even unethical. Therefore, it is more feasible to estimate the treatment effect from observational data.

*Corresponding author.

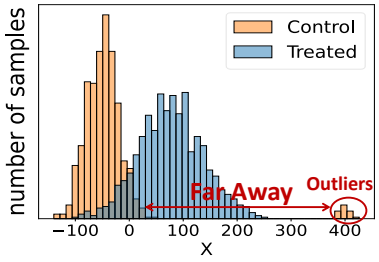


Figure 1: An example with 1-D confounders. First and second moment-based methods are misled by outliers and tend to assign large weights to them for distribution balancing.

The major issue in the observational study is the confounding bias (Yao et al. 2021), which is induced by the imbalanced confounder distributions between the treated and control groups due to the fact that the treatment is affected by the confounder instead of randomly assigned. For example, the decision on medical treatment is usually affected by age, resulting in different age distributions between different groups. Re-weighting is an effective approach for overcoming confounding bias, which creates a pseudo-population with learned sample weights so that confounder distributions between the treated and control groups are balanced.

As a classic method of re-weighting, propensity score-based method (Rosenbaum and Rubin 1983) heavily depends on the correct model specification on treatment assignment, which is difficult to obtain from complex real-world data. In order to avoid learning a propensity score model, moment-based methods (Hainmueller 2012) optimize weights directly by minimizing the difference between the confounder moments of two groups. However, estimation for high-order moments is computationally difficult and still remains an open challenge. As a result, researchers usually adopt only the first and second moments in practice, which are insufficient to balance the complex distributions of real data and are easy to be misled by outliers.

Fig. 1 shows an example where first and second moment-based methods are misled by outliers, whose weights should be reduced since their unreliable covariates and outcomes can negatively affect the treatment effect estimation. However, to achieve distribution balancing, the outliers in the

control group will be assigned large weights, since they are helpful for reducing the difference between these two dsitributions in terms of the mean and variance.

In this paper, we propose a novel re-weighting method for balancing the complex confounder distributions by exploiting geometric information, which capture the intrinsic structure of confounders to balance distributions and recognize outliers. In specific, we explore both inter and intra-group geometric information. The inter-group geometry captures distances between control and treated samples, and a sample with a smaller inter-group distance tends to be assigned a larger weight since it contributes to the distribution balancing. The intra-group geometry captures the internal similarity within one group, which leverages metric information for further distribution balancing and also helps reduce the weights of outliers, since the outliers are far away from other samples and have different internal similarity properties.

Motivated by this, we propose an optimal transport model to learn sample weights by extracting geometric information from confounders, which is also theoretically supported by our finding that the estimation error of the treatment effect can be upper bounded by the objective of optimal transport. In specific, we model the problem of confounder balancing as a semi-relaxed optimal transport model, which learns weights by minimizing the distribution discrepancy measured by the transport cost from control samples to treated samples. We leverage Wasserstein and Gromov-Wassersein discrepancies to extract inter and intra-group geometric information, respectively. Furthermore, in order to encourage more samples to be transported for improving the data efficiency, we employ a negative entropic regularization on the sample weights, which is modeled as the empirical marginal distribution in optimal transport. We develop a projected mirror descent algorithm equipped with the Kull-Leibler (KL) divergence to solve the derived optimization problem.

We summarize our principal contributions as follows:

- We propose a semi-relaxed fused Gromov-Wasserstein discrepancy model with regularized marginal distribution to learn weights for samples, which extracts both inter and intra-group geometries to reduce confounding bias.
- We develop a projected mirror descent algorithm to solve the resultant optimization problem, which learns coupled sample weights and a transport plan to balance the confounder distributions.
- We conduct extensive experiments on both synthetic and real-world data sets to demonstrate the advantages of our proposed method in terms of treatment effect estimation and robustness to outliers.

Related Works

Causal Effect Estimation

Re-weighting is an effective class of methods to overcome confounding bias. Inverse Propensity Weighting (IPW) treats the inverse of propensity scores as weights, which can be estimated via regression algorithms (Rosenbaum and Rubin 1983; McCaffrey, Ridgeway, and Morral 2004; Westreich, Lessler, and Funk 2010; Zhao 2019). Various methods then have been proposed to make IPW more robust via combining outcome regression (Robins, Rotnitzky, and Zhao 1994) or exploiting dual characteristics of propensity score (Imai and Ratkovic 2014). However, these methods still heavily depend on the correct specification of the propensity or outcome regression models.

Recently, researchers propose to learn weights directly. (Hainmueller 2012) proposes to maximize the entropy of the weights while aligning the moments between treated and control groups. (Kuang et al. 2017) learns weights by aligning moments while distinguishing confounders. In practice, researchers usually align only the first and second moments, because accurate estimation for high-order moments is computationally difficult. However, the first and second moments are insufficient to balance the complex distribution of real data and are easy to be misled by outliers. (Arbour, Dimmery, and Sondhi 2021) learns weights using a standard binary classifier. Different from these methods, our method exploit both inter and intra-group geometric information via optimal transport, resulting in outstanding performance while being robust to outliers.

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 ability in different kinds of applications (Peyré, Cuturi et al. 2019; Yan et al. 2019; Zhao and 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 metric for image retrieval (Rubner, Tomasi, and Guibas 2000). For transfer learning, data from one distribution is transported to another distribution based on the optimal transport plan for label information transfer (Courty, Flamary, and Tuia 2014; Courty et al. 2017b,a). For generative modeling, the Wasserstein distance derived by optimal transport is minimized to train deep generative models (Tolstikhin et al. 2018; Arjovsky, Chintala, and Bottou 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 some researchers trying to introduce optimal transport into causal inference. (Gunsilius and Xu 2021) employs unbalanced optimal transport for matching. (Torous, Gunsilius, and Rigollet 2021) generalizes Changes-in-Changes (CiC) to high-dimensional setting based on optimal transport. (Li et al. 2021) proposes to infer counterfactual outcome via transporting factual distribution to the counterfactual distribution. (Dunipace 2021) applies optimal transport to achieve distribution balance by finding a intermediate distribution with learned weights. Compared with them, our contributions lie in two aspects. First, we establish a connection between treatment effect estimation and optimal transport, providing theoretical support for our method that learns weights via optimal transport. Second, we further explore the property that optimal transport can leverage geometric information, and then propose to extract both inter and intra-group geometries to remove confounding bias.

Problem Statement and Notations

We consider the Rubin-Neyman potential outcome framework (Rubin 1974; Splawa-Neyman, Dabrowska, and Speed 1990) with n observational samples $\{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^n$ and binary treatment, *i.e.*, $t_i \in \{0, 1\}$. $\mathbf{x}_i \in \mathbb{R}^d$ is a feature vector, and $y_i \in \mathbb{R}$ is the observational outcome, where d is the dimension of features. Given the potential outcomes $Y_0(\cdot)$ and $Y_1(\cdot)$, the observational outcome is $y_i = t_i Y_1(\mathbf{x}_i) + (1 - t_i)Y_0(\mathbf{x}_i)$. In specific, the observational data set includes a control group $\{(\mathbf{x}_i^c, t_i^c, y_i^c)\}_{i=1}^{n_c}$ with $t_i^c = 0$, and a treated group $\{(\mathbf{x}_j^t, t_j^t, y_j^t)\}_{j=1}^{n_t}$ with $t_j^t = 1$, where $n = n_c + n_t$.

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 causal identification (Rosenbaum and Rubin 1983; Imbens and Wooldridge 2009).

We focus on estimating the Average Treatment effect on the Treated group (ATT), which is the average difference between the potential outcomes under treated and control situation on the treated group. ATT is defined as

$$\text{ATT} = \mathbb{E}[Y_1(\mathbf{x}_i)|t_i = 1] - \mathbb{E}[Y_0(\mathbf{x}_i)|t_i = 1]. \quad (1)$$

The first term $\mathbb{E}[Y_1(\mathbf{x}_i)|t_i = 1]$ can be easily estimated based on the observational data by $\frac{1}{n_t}\sum_{i=1}^{n_t} y_i^t$. However, the second term $\mathbb{E}[Y_0(\mathbf{x}_i)|t_i = 1]$ involves unobservational potential outcomes $Y_0(\mathbf{x}_i)$ on the treated group. Under the strong ignorability assumption, the second term can be estimated by removing the confounding bias, which is usually achieved by the re-weighting approach (Kuang et al. 2017). In specific, re-weighting aims to learn weights $\{w_j\}_{j=1}^{n_c}$ for the control samples $\{\mathbf{x}_j^c\}_{j=1}^{n_c}$, so that the distributions of control and treated groups are aligned and the confounding bias is reduced. With the learned weights ATT can be estimated by

$$\widehat{\text{ATT}} = \frac{1}{n_t}\sum_{j=1}^{n_t} y_j^t - \sum_{i=1}^{n_c} w_i y_i^c. \quad (2)$$

Throughout the paper, $[n]$ denotes a set including the elements $\{1, \ldots, n\}$. $\mathbf{1}_n$ 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 denoted as $\text{tr}(\mathbf{A})$. Given two matrices $\mathbf{A}$ and $\mathbf{B}$ with the same size, the inner product of them is denoted as $\langle \mathbf{A}, \mathbf{B} \rangle$. The Hadamard product between $\mathbf{A}$ and $\mathbf{B}$ is denoted as $\mathbf{A} \odot \mathbf{B}$, *i.e.*, $(\mathbf{A} \odot \mathbf{B})_{ij} = A_{ij}B_{ij}$.

Learning Model

In this section, we first describe the key concepts of optimal transport, then connect it to the ATT estimation problem via the dual form of optimal transport. After that, we propose our learning model for ATT via optimal transport.

Connection between Optimal Transport and ATT

Optimal transport seeks an transport 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 $\mu \in P(\mathcal{X}), \nu \in P(\mathcal{Y})$ and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the Kantorovich problem seeks a *transport plan* $\pi(x, y)$ via optimizing the following problem:

$$(KP) = \inf_{\pi \in \Pi(\mu,\nu)} \int c(x, y) d\pi(x, y), \quad (3)$$

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):

$$(DP) = \sup_{f(x)+g(y) \le c(x,y)} \int f(x) d\mu(x) + \int g(y) d\nu(y). \quad (4)$$

We now turn our attention to the ATT estimation problem. We first decompose its estimation error as follow, in terms of $m_t(\mathbf{x}) = \mathbb{E}[Y_t(\mathbf{x})|\mathbf{x}], t \in \{0, 1\}$, as discussed in (Ben-Michael et al. 2021):

$$\begin{aligned}
&\widehat{\text{ATT}} - \text{ATT} \\
&= (\frac{1}{n_t}\sum_{j=1}^{n_t} y_j^t - \sum_{i=1}^{n_c} w_i y_i^c) - \frac{1}{n_t}\sum_{j=1}^{n_t}(m_1(\mathbf{x}_j) - m_0(\mathbf{x}_j)) \\
&= \frac{1}{n_t}\sum_{j=1}^{n_t} m_0(\mathbf{x}_j) + \frac{1}{n_t}\sum_{j=1}^{n_t}(y_j^t - m_1(\mathbf{x}_j)) - \sum_{i=1}^{n_c} w_i y_i^c \\
&= -\sum_{i=1}^{n_c} w_i m_0(\mathbf{x}_i) + \frac{1}{n_t}\sum_{j=1}^{n_t} m_0(\mathbf{x}_j) \quad (5) \\
&\quad + \frac{1}{n_t}\sum_{j=1}^{n_t}(y_j^t - m_1(\mathbf{x}_j)) + \sum_{i=1}^{n_c} w_i(m_0(\mathbf{x}_i) - y_i^c). \quad (6)
\end{aligned}$$

The term (5) is the imbalance between treated and control groups, and the term (6) means randomness of y in these two groups, which is zero under expectation. Moreover, the term (6) $\to 0$ as $n \to \infty$ under mild regularity conditions (Kong et al. 2023). Therefore, the estimation error of ATT depends on the term (5), which allows us to link the ATT estimation problem to the Kantorovich problem, as stated in the following proposition.

Proposition 1 *Let μ, ν be the distribution of weighted control and treated groups respectively. Suppose $-m_0 \in f, m_0 \in g$, and assume there exists a cost function such that $m_0(y) - m_0(x) \le c(x, y)$. We have:*

$$\widehat{ATT} - ATT \le (DP) \le (KP). \quad (7)$$

The first inequality holds because (DP) is the worst-case of the imbalance term (5) under the assumptions in Proposition 1, which controls the estimation error of ATT. The second inequality holds because the property of the dual problem. For the conditions when the strong duality holds so that the bound is tight, please refer to (Villani 2008, 2021).

Proposition 1 shows that the estimation error of ATT is bounded by the Kantorovich problem, which theoretically supports that ATT estimation error can be minimized by learning sample weights via optimal transport. We present our learning model to achieve this in the following.

Inter-group Geometric Information

Supported by Proposition 1, we propose to learn weights via solving the Kantorovich problem, which leverages inter-group geometric information to minimize ATT estimation error. Overall, we achieve distribution balancing between control and treated groups by minimizing the transport cost with estimated weights for the control samples.

For discrete samples, we consider a semi-relaxed optimal transport model with empirical distributions. In specific, let $\mathbf{C}$ be the cost matrix, and the element C_{ij} measures the transport cost from $\mathbf{x}_i^c$ to $\mathbf{x}_j^t$ and can be defined as

$$C_{ij} = \|\mathbf{x}_i^c - \mathbf{x}_j^t\|_2^2. \tag{8}$$

The transport plan is represented by the matrix $\mathbf{T}$ with the element T_{ij} being the transport mass from the i-th control sample to the j-th treated sample. $\boldsymbol{\mu}^t = \mathbf{T}^\top \mathbf{1} = [\frac{1}{n_t}, \ldots, \frac{1}{n_t}]^\top$ is the fixed marginal distribution of the treated group, which is the weights of uniform distribution used in Eq. (2) for treated samples. $\boldsymbol{\mu}^c = \mathbf{T}\mathbf{1}$ is the estimated marginal distribution of the control group used for re-weighting. Based on these, we leverage the Wasserstein discrepancy to learn weights for control group by the following semi-relaxed optimal transport model:

$$\begin{aligned} \min_{\boldsymbol{\mu}^c} \min_{\mathbf{T}} \ & \langle \mathbf{C}, \mathbf{T} \rangle \\ \text{s.t. } & \mathbf{T}\mathbf{1}_{n_t} = \boldsymbol{\mu}^c, \mathbf{T}^\top \mathbf{1}_{n_c} = \boldsymbol{\mu}^t, T_{ij} \in [0,1]. \end{aligned} \tag{9}$$

The above optimal transport model usually induces a sparse solution, which means only a limited number of control samples are transported (Blondel, Seguy, and Rolet 2018; Vincent-Cuaz et al. 2022), suffering from low data efficiency (Kaddour et al. 2021). To tackle this issue, we apply a negative entropy regularization on the marginal distributions $\boldsymbol{\mu}^c$ to encourage more control samples to be transported, which is defined as

$$\Omega(\mathbf{T}) = \sum_{i=1}^{n_c} T_{i\cdot}(\log T_{i\cdot} - 1), \tag{10}$$

where $T_{i\cdot}$ is the sum of the i-th row of $\mathbf{T}$, *i.e.*,

$$T_{i\cdot} = \sum_{j=1}^{n_t} T_{ij}. \tag{11}$$

In addition, we constrain $\mathbf{T}$ to belong to the following domain of the definition

$$\mathcal{T} = \{\mathbf{T} \mid \mathbf{T}^\top \mathbf{1}_{n_c} = \boldsymbol{\mu}^t, T_{ij} \in [0,1]\}, \tag{12}$$

which does not consider the constraint $\mathbf{T}\mathbf{1}_{n_t} = \boldsymbol{\mu}^c$ since $\boldsymbol{\mu}^c$ are also parameters to be optimized. Based on this, our semi-relaxed optimal transport model with a regularized marginal distribution with respect to $\mathbf{T}$ is given as follows:

$$\min_{\mathbf{T}} \ \langle \mathbf{C}, \mathbf{T} \rangle + \gamma_c \Omega(\mathbf{T}) \qquad \text{s.t. } \mathbf{T} \in \mathcal{T}, \tag{13}$$

where γ_c is the trade-off parameter.

Intra-group Geometric Information

Besides inter-group geometric information, we further employ optimal transport to exploit the intra-group geometric information in control and treated groups to re-weight control samples. To this end, we construct the metric matrices $\mathbf{M}^c$ and $\mathbf{M}^t$, where either a similarity or a distance metric can be adopted. The elements $M_{ii'}^c$ (*resp.*, $M_{jj'}^t$) represents the metric between $\mathbf{x}_i^c$ and $\mathbf{x}_{i'}^c$ (*resp.*, $\mathbf{x}_j^t$ and $\mathbf{x}_{j'}^t$). Based on these, we leverage the Gromov-Wasserstein discrepancy to minimize the transport cost between a control pair $(\mathbf{x}_i^c, \mathbf{x}_{i'}^c)$ to a treatment pair $(\mathbf{x}_j^t, \mathbf{x}_{j'}^t)$. Here, the transport cost between two pairs is measured by the difference between the metrics $M_{ii'}^c$ and $M_{jj'}^t$, which is defined as $\ell_{(ii')(jj')} = \frac{1}{2}(M_{ii'}^c - M_{jj'}^t)^2$, so that metric information is also incorporated for distribution balancing. According to Proposition 1 in (Peyré, Cuturi, and Solomon 2016), the total transport cost can be rewritten as

$$\begin{aligned} &\sum_{i=1}^{n_c}\sum_{j=1}^{n_t}\sum_{i'=1}^{n_c}\sum_{j'=1}^{n_t} \ell_{(ii')(jj')} T_{ij} T_{i'j'} \\ &\quad = \sum_{i=1}^{n_c}\sum_{j=1}^{n_t}\Big(\sum_{i'=1}^{n_c}\sum_{j'=1}^{n_t} \ell_{(ii')(jj')} T_{i'j'}\Big) T_{ij} \\ &\quad = \langle (\mathbf{M}^c \odot \mathbf{M}^c)\mathbf{T}\mathbf{1}_{n_t}\mathbf{1}_{n_t}^\top - 2\mathbf{M}^c\mathbf{T}(\mathbf{M}^t)^\top, \mathbf{T} \rangle \\ &\qquad + \text{tr}(\boldsymbol{\mu}^t(\boldsymbol{\mu}^t)^\top(\mathbf{M}^t \odot \mathbf{M}^t)^\top), \end{aligned} \tag{14}$$

where the first term is related to $\mathbf{T}$, and the second term is constant. For simplicity, we define the matrix $\mathbf{G}$ as

$$\mathbf{G} = (\mathbf{M}^c \odot \mathbf{M}^c)\mathbf{T}\mathbf{1}_{n_t}\mathbf{1}_{n_t}^\top - 2\mathbf{M}^c\mathbf{T}(\mathbf{M}^t)^\top, \tag{15}$$

and achieve the semi-relaxed fused Gromov-Wasserstein model with an entropic regularization on the marginal distribution as follows:

$$\begin{aligned} \min_{\mathbf{T}} \ & \alpha\langle \mathbf{C}, \mathbf{T} \rangle + (1-\alpha)\langle \mathbf{G}, \mathbf{T} \rangle + \gamma_c \Omega(\mathbf{T}) \\ \text{s.t. } & \mathbf{T} \in \mathcal{T}, \end{aligned} \tag{16}$$

where $\alpha \in [0,1]$ is a trade-off parameter between inter and intra geometric terms. The first term in Eq. (16) leverages inter-group geometry to achieve distribution balancing by assigning small weights (small $T_{i\cdot}$) to control samples which are far from treated group, since they have large transport costs C_{ij}. The second term further leverages intra-group geometry for balancing since it prefers to transport pairs with similar metrics (*i.e.*, $M_{ii'}^c$ and $M_{jj'}^t$). In addition, the outliers can be recognized based on their different metric property, which results from their large distances to other samples.

The optimization algorithm is given in next section. After obtaining the solution $\mathbf{T}$, the estimated marginal distribution $\{T_{i\cdot}\}_{i=1}^{n_c}$ can be calculated by Eq. (11) and taken as the weights for control samples, and ATT can be estimated by

$$\widehat{\text{ATT}} = \frac{1}{n_t}\sum_{j=1}^{n_t} y_j^t - \sum_{i=1}^{n_c} T_{i\cdot} y_i^c. \tag{17}$$

Optimization

In this section, we develop a projected mirror descent (Nemirovskij and Yudin 1983; Raskutti and Mukherjee 2015)

based on the Kullback-Leibler (KL) divergence to solve Problem (16), which is non-trivial to address because of the equality constrints. For simplicity, we define the objective function in Problem (16) as

$$F(\mathbf{T}) = \alpha\langle \mathbf{C}, \mathbf{T}\rangle + (1-\alpha)\langle \mathbf{G}, \mathbf{T}\rangle + \gamma_c \Omega(\mathbf{T}). \quad (18)$$

With the symmetric $\mathbf{M}^c$ and $\mathbf{M}^t$, the (i,j)-th element of the gradient $\nabla F(\mathbf{T})$ is denoted by ∇_{ij}, which is calculated as

$$\nabla_{ij} = \alpha C_{ij} + 2(1-\alpha)G_{ij} + \gamma_c \log T_{i}.. \quad (19)$$

At each iteration, we solve the following problem

$$\mathbf{T}^{k+1} = \arg\min_{\mathbf{T}\in\mathcal{T}} \ \eta\langle \nabla F(\mathbf{T}^k), \mathbf{T}\rangle + \mathcal{D}(\mathbf{T}||\mathbf{T}^k), \quad (20)$$

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 $\mathcal{T}$ by projection. Next, we present the details of these two operations.

Proximal Gradient Descent

Let $\mathbf{\Upsilon}^k$ be the solution to Problem (20) without considering the constraint $\mathbf{T} \in \mathcal{T}$, *i.e.*,

$$\mathbf{\Upsilon}^k = \arg\min_{\mathbf{T}} \ \eta\langle \nabla F(\mathbf{T}^k), \mathbf{T}\rangle + \mathcal{D}(\mathbf{T}||\mathbf{T}^k). \quad (21)$$

We adopt the KL divergence between two distributions $\mathbf{T}$ and $\mathbf{T}^k$ as the Bregman divergence $\mathcal{D}(\mathbf{T}||\mathbf{T}^k)$, which is defined as

$$\mathcal{D}(\mathbf{T}||\mathbf{T}^k) = \sum_{i=1}^{n_c}\sum_{j=1}^{n_t} T_{ij}\log(\frac{T_{ij}}{T^k_{ij}}) - T_{ij} + T^k_{ij}. \quad (22)$$

Then the closed-form solution to Problem (21) is given as

$$\mathbf{\Upsilon}^k = \mathbf{T}^k \odot \exp(-\eta \nabla F(\mathbf{T}^k)). \quad (23)$$

Projection Operation To make sure $\mathbf{T}^{k+1}$ satisfies the constraints in Eq. (12), we update $\mathbf{T}^{k+1}$ by finding $\mathbf{T} \in \mathcal{T}$ which is most close to $\mathbf{\Upsilon}^k$ under the KL metric. This is achieved by solving the following projection problem

$$\begin{aligned} \min_{\mathbf{T}} \quad & \mathcal{D}(\mathbf{T}||\mathbf{\Upsilon}^k) := \sum_{i=1}^{n_c}\sum_{j=1}^{n_t} T_{ij}\log(\frac{T_{ij}}{\Upsilon^k_{ij}}) - T_{ij} + \Upsilon^k_{ij} \\ \text{s.t.} \quad & \mathbf{T}^\top \mathbf{1}_{n_c} = \boldsymbol{\mu}^t. \end{aligned} \quad (24)$$

In the following, we provide the closed-form solution to the problem (24), and show that the box constraints $T_{ij} \in [0,1]$ can be safely removed.

By introducing the Lagrangian multipliers $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_{n_t}]^\top$ for the equality constraint $\mathbf{T}^\top \mathbf{1}_{n_c} = \boldsymbol{\mu}^t$, we obtain the Lagrangian $\mathcal{L}(\mathbf{T}, \boldsymbol{\lambda})$ as follows

$$\begin{aligned} \mathcal{L}(\mathbf{T}, \boldsymbol{\lambda}) = \sum_{i=1}^{n_c}\sum_{j=1}^{n_t} T_{ij}\log(\frac{T_{ij}}{\Upsilon^k_{ij}}) - T_{ij} + \Upsilon^k_{ij} \\ + \boldsymbol{\lambda}^\top(\mathbf{T}^\top \mathbf{1}_{n_c} - \boldsymbol{\mu}^t). \end{aligned} \quad (25)$$

Next, we take the partial derivative of $\mathcal{L}(\mathbf{T}, \boldsymbol{\lambda})$ with respect to T_{ij} to zero

$$\frac{\partial \mathcal{L}(\mathbf{T}, \boldsymbol{\lambda})}{\partial T_{ij}} = \log(\frac{T_{ij}}{\Upsilon^k_{ij}}) + T_{ij}\frac{\Upsilon^k_{ij}}{T_{ij}}\frac{1}{\Upsilon^k_{ij}} - 1 + \lambda_j = 0,$$

Algorithm 1: Optimal Transport for Causal Inference.

Input: Data matrices $\mathbf{X}^c$, $\mathbf{X}^t$. Metric matrices $\mathbf{M}^c$, $\mathbf{M}^t$. The cost matrix $\mathbf{C}$. Trade-off parameters α, γ_c.

Initialize $\mathbf{T}^0 : T^0_{ij} = \frac{1}{n_c n_t}, \forall i, j$. Set $k = 1$.
repeat
  Calculate $\mathbf{\Upsilon}^k$ according to Eq. (23).
  Update $\mathbf{T}^k$ according to Eq. (28).
  $k := k + 1$.
until Convergence.
Estimate ATT according to Eq. (17).

and then obtain the optimality condition as

$$\log T_{ij} = \log \Upsilon^k_{ij} - \lambda_j \Rightarrow T_{ij} = \Upsilon^k_{ij}\exp(-\lambda_j). \quad (26)$$

According to the equality constraint $\mathbf{T}^\top \mathbf{1}_{n_c} = \boldsymbol{\mu}^t$, we have $\sum_{i=1}^{n_c} T_{ij} = \frac{1}{n_t}$. By combining the condition in Eq. (26), we further obtain

$$\begin{aligned} \sum_{i=1}^{n_c} T_{ij} = \sum_{i=1}^{n_c} \Upsilon^k_{ij}\exp(-\lambda_j) = \exp(-\lambda_j)\sum_{i=1}^{n_c}\Upsilon^k_{ij} = \frac{1}{n_t} \\ \Rightarrow \exp(-\lambda_j) = \frac{1}{n_t \sum_{i=1}^{n_c}\Upsilon^k_{ij}}. \end{aligned} \quad (27)$$

By plugging Eq. (27) into (26), we finally achieve the closed-form solution

$$T_{ij} = \Upsilon^k_{ij} \Big/ \Big(n_t \sum_{i=1}^{n_c} \Upsilon^k_{ij}\Big). \quad (28)$$

For an initial value $T^0_{ij} \geq 0$, given $\Upsilon^k_{ij} > 0$ which is guaranteed by the update rule in Eq. (23), it is obvious that the solution obtained by Eq. (28) satisfies the box constraint $T_{ij} \in [0,1]$. Therefore, Problem (24) does not consider this constraint explicitly.

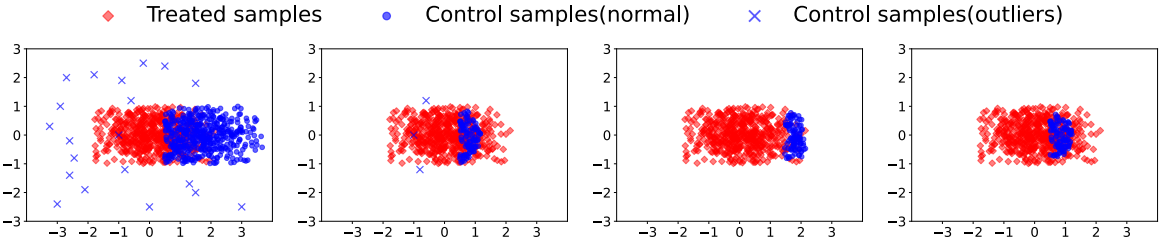
Algorithm 1 summarizes our proposed method.

Experiments

Compared Methods and Evaluation Metric

We compare OTCI with the following methods: (i) Methods based on propensity score or outcome regression: inverse propensity weighting **IPW** (Rosenbaum and Rubin 1983), doubly robust estimator **DR** (Robins, Rotnitzky, and Zhao 1994), covariate balancing propensity score **CSPS** (Imai and Ratkovic 2014), approximate residual balancing **ARB** (Athey, Imbens, and Wager 2018). (ii) Methods based on moment alignment: entropy balancing **Ebal** (Hainmueller 2012), differentiated confounder balancing **DCB** (Kuang et al. 2017). (iii) Methods based on machine learning: **BART** (Chipman, George, and McCulloch 2010) and **CFR** (Shalit, Johansson, and Sontag 2017). (iv) Methods based on optimal transport: optimal transport weights **OTW** (Dunipace 2021), causal optimal transport **causalOT** (Li et al. 2021).

We implement Ebal as Ebal(1) and Ebal(2), which correspond to aligning only the first moment and aligning the first and second moments, respectively. Similarly, we implement DCB as DCB(X) considering original features for



(a) Data (b) Result of OTCI$_{inter}$ (c) Result of OTCI$_{intra}$ (d) Result of OTCI

Figure 2: A toy example with the presence of outliers. (a) shows all the treated and control samples (with 20 outliers); (b), (c), (d) show all the treated samples and 100 control samples with the largest weights, which are learned by OTCI with only inter-group geometry (OTCI$_{inter}$), only intra-group geometry (OTCI$_{intra}$), both inter and intra-group geometries (OTCI).

	Ebal(1)	Ebal(2)	Inter	Intra	Both
#($\geq 10^{-2}$)	10	13	0	0	0
#($10^{-3}, 10^{-2}$)	9	4	17	3	0
#($10^{-4}, 10^{-3}$)	1	3	3	3	5
#($\leq 10^{-4}$)	0	0	0	14	15
MAE	.473	.517	.407	.355	**.347**

Table 1: Results on a toy example.

the first moment alignment, and DCB(A) considering augmented features for the first and second moments alignment.

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

Experiments on Toy Example

We first design a 2D toy example with outliers. Specifically, we generate 500 samples for the treated and control groups, respectively, as follows: $\mathbf{x}^t \sim \mathcal{N}([0,0]^\top, \mathbf{\Sigma}), \mathbf{x}^c \sim \mathcal{N}([1.5,0]^\top, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = [[0.8,0];[0,0.4]]$. In addition, we randomly generate 20 outliers in the control group. We do not consider outliers in the treated group since the target is to estimate ATT, in which the weights of the treated groups are fixed instead of learnable. The covariates are visualized in Fig. 2a, and the outcomes are generated as $y = \sin(\mathbf{w}_1^\top \mathbf{x}) + \cos(\mathbf{w}_2^\top (\mathbf{x} \odot \mathbf{x})) + t + \epsilon$, where $\mathbf{w}_1 = [8.0, 1.5], \mathbf{w}_2 = [1.5, 2.0]$, and $\epsilon \sim \mathcal{N}(0, 0.1)$.

We compare OTCI with its two variants, OTCI$_{inter}$ with only inter-group geometry, and OTCI$_{intra}$ with only intra-group geometry, and the representative moment-based method Ebal. Table 1 summarizes the weights of the 20 outliers learned by these methods and their estimation errors of ATT. Figs. 2b, 2c, 2d visualize all the treated samples and 100 control samples with the largest weights learned by OTCI and its variants. We have the following observations:

- Table 1 indicates that Ebal(1) and Ebal(2) fail to correctly identify the outliers, leading to large MAE. This issue arises because Ebal assigns large weights to outliers if they assist in achieving moment alignment.
- Fig. 2b and 2c show that OTCI$_{inter}$ performances well on distribution alignment by allocating larger weights to high-overlap regions. However, it lacks the capability to detect those outliers that are close to the treated group, which can be addressed by exploiting the metric information between sample pairs. By integrating both inter and intra-group geometries, OTCI not only performs well for distribution alignment but also demonstrates robustness to outliers, as shown in Fig. 2d and Table 1.

Experiments on Simulation Data

In this part, following similar protocols as in (Yao et al. 2018; Hatt and Feuerriegel 2021), we conduct experiments on simulation data with two different settings:

- For Gaussian distribution, we generate 1500 treated samples from $\mathcal{N}(\mu_t^{10\times 1}, 0.5 \times \mathbf{\Sigma}_t \mathbf{\Sigma}_t^T)$ and 1500 control samples from $\mathcal{N}(\mu_c^{10\times 1}, 0.5 \times \mathbf{\Sigma}_c \mathbf{\Sigma}_c^T)$, where $\mathbf{\Sigma}_\cdot \sim \mathcal{U}((0,\mu_\cdot)^{10\times 10})$. We fix $\mu_t = 0.5$ and vary μ_c to generate data with different levels of confounding bias.
- For Non-Gaussian distribution, we use Gaussian mixture distribution. We first generate two Gaussian distribution $\mathcal{N}_1 = \mathcal{N}(0.5^{10\times 1}, 0.5 \times \mathbf{\Sigma}_1 \mathbf{\Sigma}_1^T), \mathcal{N}_2 = \mathcal{N}(1^{10\times 1}, 0.5 \times \mathbf{\Sigma}_2 \mathbf{\Sigma}_2^T)$ where $\mathbf{\Sigma}_1 \sim \mathcal{U}((0,0.5)^{10\times 10})$ and $\mathbf{\Sigma}_2 \sim \mathcal{U}((0,1)^{10\times 10})$, and then generate 1500 treated samples as $\mathbf{x}^t \sim \alpha_t \mathcal{N}_1 + (1-\alpha_t)\mathcal{N}_2$ and 1500 control samples as $\mathbf{x}^c \sim \alpha_c \mathcal{N}_1 + (1-\alpha_c)\mathcal{N}_2$. We fix $\alpha_t = 0.5$ and vary α_c to generate data with different confounding bias.
- For the above two covariate distributions, the outcomes are both generated as $y = \sin(\mathbf{w}_1^\top \mathbf{x}) + \cos(\mathbf{w}_2^\top (\mathbf{x} \odot \mathbf{x})) + t + \epsilon$, where $\mathbf{w}_\cdot \sim \mathcal{U}((0,1)^{10\times 1})$ and $\epsilon \sim \mathcal{N}(0, 0.1)$.

Table 2 reports the results under different settings. IPW, DR, CBPS, and ARB obtain limited performance since they heavily depend on the correct specification of propensity or regression models, which are difficult to be satisfied. Ebal and DCB perform better when aligning the first and second moments. Achieving better distribution alignment usually requires higher-order moments, which are difficult to estimate with limited samples. OTW and causalOT outperform other baselines due to their usage of optimal transport. However, these two methods do not leverage intra-group geometry, which captures intrinsic structure in data. OTCI fully extracts geometric information from both inter and intra-group via optimal transport, leading to significant improvements over the compared methods in different settings.

	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	.137±.055	.288±.062	.339±.079	.329±.108	.038±.025	.049±.035	.062±.045	.099±.054
DR	.125±.056	.260±.072	.294±.102	.302±.095	.037±.024	.049±.034	.059±.042	.091±.054
CBPS	.124±.053	.276±.058	.322±.082	.317±.102	.037±.025	.049±.035	.059±.044	.094±.056
ARB	.134±.051	.282±.057	.334±.073	.341±.103	.037±.025	.049±.035	.062±.044	.094±.055
Ebal(1)	.131±.067	.247±.111	.303±.098	.320±.083	.037±.025	.049±.035	.059±.044	.091±.058
Ebal(2)	.117±.082	.196±.105	.249±.100	.281±.069	.037±.024	.045±.033	.053±.043	.085±.052
DCB(X)	.138±.095	.234±.069	.236±.083	.248±.080	.035±.032	.056±.037	.078±.046	.081±.078
DCB(A)	.124±.113	.212±.072	.216±.089	.220±.082	.035±.036	.043±.032	.071±.062	.081±.065
OTW	.141±.197	.192±.075	.272±.227	.266±.133	.033±.025	.041±.027	.048±.029	.051±.048
CausalOT	.059±.029	.136±.043	.212±.052	.241±.060	.035±.032	.043±.030	.052±.031	.056±.031
OTCI	**.041±.022**	**.042±.040**	**.067±.039**	**.072±.051**	**.019±.023**	**.021±.018**	**.024±.013**	**.034±.023**

Table 2: Results on simulation data in different settings. The mean and standard deviation of MAE for ATT are reported.

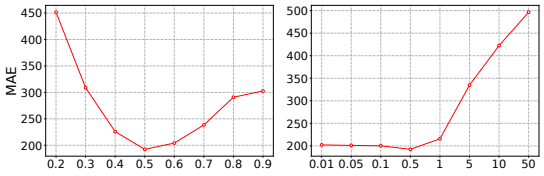
	LaLonde	Twins
IPW	853.513 ± 78.003	.009±.004
DR	555.262±243.713	.007±.004
CBPS	562.232±106.226	.007±.004
ARB	658.561±91.844	.005±.004
Ebal(1)	562.387±105.917	.009±.003
Ebal(2)	476.013±180.443	.008±.003
DCB(X)	533.172±268.699	.007±.004
DCB(A)	406.180±302.601	.006±.004
BART	532.578 ± 200.600	.008±.004
CFR	569.022 ± 250.001	.009±.004
OTW	289.689±198.828	.006±.006
CausalOT	382.277±107.453	.007±.006
OTCI	**192.364±120.783**	**.003±.002**

Table 3: Results on LaLonde and Twins datasets. The mean and standard deviation of MAE about ATT is reported.

Experiments on Real-world Data Finally, we conduct experiments on two real datasets for ATT estimation, including LaLonde and Twins. **LaLonde**[1] consists of two parts. The first part comes from a RCT (NSW). In the second part, as (Kuang et al. 2017) did, we replace the control group in NSW with another control group from the observational data (CPS). The treatment is whether the participant attend the job training program, and the outcome is the earning in 1978. The data contains 8 covariates. **Twins** is collected from the twins born in USA between 1989-1991 (Almond, Chay, and Lee 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: $t \sim Bern(\text{sigmoid}(\mathbf{w}^\top \mathbf{x} + b))$ where $\mathbf{w} \sim \mathcal{U}((-0.1, 1)^{30\times 1})$ and $b \sim \mathcal{N}(0, 0.1)$.

The results are reported in Table 3. We have similar observations as in simulation data. The performance of IPW, DR, DCB and ARB is relatively lower. Ebal and DCB obtain limited performance since the first and second moments are insufficient to achieve good distribution alignment on complex data in the real world. Benefiting from optimal transport, OTW and causalOT achieve relatively good results. OTCI achieves the best performance by exploiting both inter and intra-group geometries to eliminate confounding bias.

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



(a) parameter: α (b) parameter: $\gamma_c(1\times10^{-2})$

Figure 3: Sensitivity analysis on LaLonde dataset

Sensitivity Analysis

We take LaLonde as an example to evaluate the parameter sensitivity of OTCI. We vary the parameters α and γ_c in Eq. (16), and plot the results in Fig. 3. From Fig. 3a, MAE increases when α becomes small or large, indicating that both inter and intra-group geometric information play important roles in removing confounding bias. From Fig. 3b, MAE increases when γ_c is larger than 10^{-2}, since a large γ_c will push the learned weights close to the uniform distribution, resulting in a failure of confounding bias elimination. Overall, when parameters $0.4 \leq \alpha \leq 0.7$ and $\gamma_c \leq 1 \times 10^{-2}$, OTCI achieves promising performance, which demonstrates its stability to the trade-off parameters in a certain range.

Conclusion

In this paper, we exploit geometry to reduce confounding bias in treatment effect estimation under the re-weighting paradigm. To this end, the connection between the estimation error of treatment effect and optimal transport is discussed, and inter as well as intra-group geometric information is captured by optimal transport for confounder balancing. By doing this, the distributions of the control and treated groups are balanced, and the negative effects of outliers can be reduced. We conducted experiments on synthetic and real-world datasets to demonstrate the efficacy of our method. We present an insight regarding the connection between treatment effect estimation and optimal transport, which provides potential possibility to employ optimal transport to address the problem of causal inference.

Acknowledgments

This research was supported in part by National Key R&D Program of China (2021ZD0111501), National Natural Science Foundation of China (62206061, 61876043, 61976052, 62206064), National Science Fund for Excellent Young Scholars (62122022), Guangzhou Basic and Applied Basic Research Foundation (2023A04J1700), and the major key project of PCL (PCL2021A12). The work of Michael K. Ng was supported in part by Hong Kong Research Grant Council GRF (17201020, 17300021), CRF (C1013-21GF), and Joint NSFC-RGC (N-HKU76921).

References

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

Arbour, D.; Dimmery, D.; and Sondhi, A. 2021. Permutation weighting. In *International Conference on Machine Learning*, 331–341. PMLR.

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

Athey, S.; Imbens, G. W.; and Wager, S. 2018. 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.

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

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

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

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

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

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

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

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

Glass, T. A.; Goodman, S. N.; Hernán, M. A.; and Samet, J. M. 2013. Causal inference in public health. *Annual review of public health*, 34: 61–75.

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

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

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

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

Imbens, G. W.; and Wooldridge, J. M. 2009. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1): 5–86.

Kaddour, J.; Zhu, Y.; Liu, Q.; Kusner, M. J.; and Silva, R. 2021. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems*, 34: 24841–24854.

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

Kong, I.; Park, Y.; Jung, J.; Lee, K.; and Kim, Y. 2023. Covariate balancing using the integral probability metric for causal inference. *arXiv preprint arXiv:2305.13715*.

Kuang, K.; Cui, P.; Li, B.; Jiang, M.; and Yang, S. 2017. 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*, 265–274.

Li, Q.; Wang, Z.; Liu, S.; Li, G.; and Xu, G. 2021. Causal Optimal Transport for Treatment Effect Estimation. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, S.; Vlassis, N.; Kawale, J.; and Fu, Y. 2016. Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns. In *IJCAI*, 3768–3774.

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

McCaffrey, D. F.; Ridgeway, G.; and Morral, A. R. 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4): 403.

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

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

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

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

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

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

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

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

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

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, 3076–3085. PMLR.

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

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

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

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

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

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

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

Westreich, D.; Lessler, J.; and Funk, M. J. 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology*, 63(8): 826–833.

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

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

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

Yao, L.; Li, S.; Li, Y.; Huai, M.; Gao, J.; and Zhang, A. 2018. Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems*, 31.

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

Zhao, Q. 2019. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47(2): 965 – 993.