Causal Optimal Transport for Treatment Effects Estimation

Journal:	*IEEE Transactions on Neural Networks and Learning Systems*
Manuscript ID	TNNLS-2020-P-14789.R1
Manuscript Type:	Regular Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Li, Qian; University of Technology Sydney, Faulty of Engineering and Information Technology Wang, Zhichao; University of New South Wales Liu, Shaowu; University of Technology Sydney Li, Gang; Deakin Univesty Xu, Guandong; University of Technology Sydney, Advanced Analytics Institute
Keywords:	causal inference, treatment effect estimation, optimal transport, counterfactual inference

Causal Optimal Transport for Treatment Effect Estimation

Qian Li, Zhichao Wang, Shaowu Liu, Gang Li and Guandong Xu[1]

***Abstract*—Treatment effect estimation helps answer questions such as *whether a specific treatment affects the outcome of interest*. One fundamental issue in this research is to alleviate the treatment assignment bias among those treated units and controlled units. Classical causal inference methods resort to the propensity score estimation, which unfortunately tends to be misspecified when only limited overlapping exists between the treated and the controlled units. Moreover, existing supervised methods mainly consider the treatment assignment information underlying the factual space, and thus their performance of counterfactual inference may be degraded due to overfitting of the factual results. To alleviate those issues, we build on the optimal transport theory and propose a novel *Causal Optimal Transport* model (CausalOT) to estimate individual treatment effect. With the proposed propensity measure, CausalOT can infer the counterfactual outcome by solving a novel regularized optimal transport problem, which allows the utilization of global information on observational covariates to alleviate the issue of limited overlapping. In addition, a novel counterfactual loss is designed for CausalOT to align the factual outcome distribution with the counterfactual outcome distribution. Most importantly, we prove the theoretical generalization bound for the counterfactual error of CausalOT. Empirical studies on benchmark datasets confirm that the proposed CausalOT outperforms state-of-the-art causal inference methods.**

***Index Terms*—**

I. Introduction

In the past decades, estimating the causal effect of a treatment (or intervention) from observational study greatly contributes to applications ranging from public health [1], economics [2], [3] to education [4]. In those areas, causal inference usually investigates the treatment effect when the intervention is applied. For example, a typical question concerned in public health is *whether an alternative medication treatment for a certain illness will lead to better results*. Treatment effect could be measured at either the group-level or individual-level, which is known as *individual treatment effect* (ITE) or *average treatment effect* (ATE), respectively. For better decision making, treatment effect estimation is necessary to answer those questions mentioned above. In this paper, we focus on the individual treatment effect (ITE) estimation.

Qian Li, Shaowu Liu and Guandong Xu (Corresponding author, e-mail: Guandong.Xu@uts.edu.au.) are with the Advanced Analytics Institute, Faulty of Engineering and Information Technology, University of Technology Sydney, Australia

Zhichao Wang is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia

Gang Li is with the Center for Cyber Security Research and Innovation, Deakin University, Australia

According to the binary treatment assignments, units in the observational study include the treated units and the controlled units. ITE is defined as the expected difference between the treated outcome and controlled outcome. Inferring ITE is different from standard supervised learning, because only the factual outcome for a specific treatment assignment (say, treatment `A`) is observable, while the counterfactual outcome corresponding to alternative treatment `B` is unknown. A simple comparison of units with different treatments may be biased due to the fact that the treatments are not randomly assigned to the units. Figure 1 shows one example on the treatment effect estimation, which investigates whether positive Yelp ratings (treatment) motivate customers to go to restaurants (outcome) [1]. For brevity, we consider the binary settings of Yelp ratings (i.e., positive or negative). Obviously, we can only get one factual outcome for a restaurant along with the observed customer reviews. Estimating the causal effect of review requires to predict what would have happened if customers flipped their reviews comments or ratings. Many customers' ratings (i.e., treatment assignment) are not random but biased, which can be affected by external factors such as the context or restaurant type. For example, the average rating of Chinese restaurants is usually higher than that of fast food restaurants. The treatment assignment bias results in considerable distribution discrepancy between two groups with different treatments, thus easily leads to an inaccurate counterfactual inference. This bias further renders ITE estimation a challenging task.

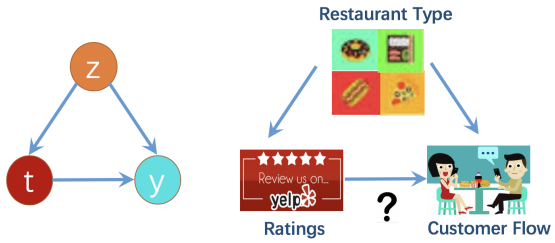


Fig. 1: An example of treatment effect estimation: an arrow represents a causal relation, t is a treatment, y is the outcome, and z is the confounder that is the common cause of treatment and outcome.

To overcome above challenges, many existing research works resort to a set of representative covariates (e.g., *age* and *health status*) of treatment assignment estimation, and

[1]*Treatment* and *outcome* are terms in the theory of causal inference, which denote a decision made or action taken and its result, respectively.

this category of methods is known as the *propensity score* methods [5], [6]. Specifically, the propensity score as the probability of receiving a treatment is estimated by adjusting the covariates. Popular methods of this category are the propensity score matching (PscoreMatch) [5], and adjustment based on propensity score [7], [8]. Although those methods have gained grounds in many applications, their performance is sensitive to inaccurate covariates selection when estimating the unknown treatment assignment [9], [10]. Another category of methods are the supervised models, which attempt to increase the accuracy of causal effect estimation, and to learn treatment effect via modelling the correlation between the covariates, treatment and factual outcomes. such as adjusted regression models [11], tree-based methods [12], [13]. While those supervised methods minimise the factual errors, but they can easily over-fit to the treated units, and thus may not generalise well to the entire population [14]. Representation-based methods as the third category recently are suggested to reduce the treatment assignment bias by learning a high-level representation for which the covariate distributions are balanced across the overall units [15], [16]. However, the training of a balanced representation requires sufficient overlapping between the treated and the controlled units, otherwise the sparsity in the units would decrease the accuracy and confidence of predicting counterfactuals [17]. Moreover, representation-based methods have more parameters, which may result in model-blindness when compared with the propensity score matching [18], [19].

To alleviate these issues, we propose a *causal optimal transport* (CausalOT) model to estimate the individual treatment effect (ITE). Leveraging the inherent interpretability of the propensity score, CausalOT is insensitive to the overlapping size between the treated and the controlled units. The proposed CausalOT method made three contributions:

- First, CausalOT builds on the optimal transport theory, and it exploits the global information of the factual space and the counterfactual space to alleviate the sparsity issue caused by the limited size of overlapping. By proposing the propensity measure, CausalOT reformulates the counterfactual inference as the task of transporting from the factual space to the counterfactual space, and it can achieve satisfactory performance even with limited overlapping between the treated and the controlled units.
- Second, we design a novel counterfactual loss for the transported samples to offset the outcome prediction bias. Moreover, a new proximal point algorithm based on Bregman divergence is proposed to improve the computation efficiency of CausalOT.
- Finally, we prove that the counterfactual error of CausalOT is bounded by the Wasserstein distance for our novel propensity measure. Extensive numerical results further confirm the effectiveness of the proposed CausalOT method.

II. Related work

Although the effectiveness of treatment in observational studies has traditionally been measured by the *average treatment effect* (ATE), modern studies have shifted the research efforts toward the *individual treatment effect* (ITE) [20]. In general, ITE from observational data has proven to be a challenge for two reasons: firstly, we can only observe one factual outcome once the treatment is chosen for the individual unit; secondly, the treatment assignment is typically biased. In the past decade, a wide variety of methods has been proposed for ITE, and they can be grouped into three categories: the propensity score methods, the supervised models and the representation based methods.

Methods in the first category are based on matching, which provides a way to estimate the counterfactual while reducing the confounding bias brought by the confounders. According to the (binary) treatment assignments, a set of individuals can be divided into a treatment group and a control group. For each treated individual, matching methods select its counterpart in the control group based on certain criteria, and treat the selected individual as a counterfactual. The outcome of counterpart is viewed as the counterfactual outcome that is used for computing ITE [6]. Various distance metrics have been adopted to compare the closeness between individuals and to select counterparts. For example, the propensity score matching [5] selects the counterpart in the controlled (or treated) units from the treated (or controlled) units with similar propensity scores (e.g., one-to-one or one-to-many). However, theoretical analysis suggests that the existing matching estimators have poor performance when the distributions of control and treatment groups are unbalanced [21]. Rather than in original covariate space, a *balanced and nonlinear representation* (BNR) [21] is learned from observational data, and a novel matching estimator named BNR-NNM is performed on BNR to provide a robust estimation of causal effect. Another example is *feature selection representation matching* (FSRM) [22] method that maps the original covariate space into a selective nonlinear, and balanced representation space, and then finds the counterpart individual based on the learned representations.

Methods in the second category consider the treatment and covariates as features, and they infer the potential outcomes by exploiting the correlations with the features. Various regression models such as linear regression can be used to build either an outcome model with the treatment as the input feature, or multiple separated outcome models, one for each treatment [23]. More sophisticated regression models include *Bayesian additive regression tree* (BART) [24] and *Causal random forest* (CausalForest) [13]. Among them, BART [24] can be intuitively considered as a Bayesian regularized tree boosting procedure, because the algorithm repeatedly refits the tree residuals; CausalForest [13] views tree and forests as an adaptive neighbourhood metric, and it estimates the treatment effect at the leaf node. As those supervised methods focus on minimising the factual errors only and can easily over-fit to the treated group [14], they may not generalise well to the entire population.

Representation based methods in the third category learn a balanced representation for which the covariate distributions are balanced across the treated and the controlled units, and then they predict the counterfactual outcomes using balanced feature representations. Early examples of this category in-

clude *balancing neural networks* (BalanceNN) [15] and *counterfactual Regression Networks* (CFRNET) [16]. Particularly, the balanced feature representation can be learned by minimising the discrepancy between the treated and the controlled units. BalanceNN [15] learns a balanced representation that adjusts the mismatch between the entire sample distribution and treated/controlled distributions in order to account for the confounding bias. CFRNET [16] provides an intuitive generalization-error bound. The expected ITE representation error is bounded by the generalization-error and the distribution distance. However, the local similarity information is largely unexplored, which prevents the generalization error from decreasing in estimating the counterfactual outcomes. SITE [25] is a deep representation learning based method that preserves local similarity and simultaneously balances data distributions for predicting counterfactual outcomes. To capture the uncertainty in the counterfactual distributions, *Generative Adversarial Network* for Individualized Treatment Effect (GANITE) [26] builds a complex GAN framework including the counterfactual outcome generator and ITE generator.

III. Preliminary

We consider an observational dataset $\{\mathbf{X}, \boldsymbol{t}, \mathbf{Y}\}$, with covariates matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ of n observed units of d-dimensional covariates $\boldsymbol{x}_i$, the binary treatment vector $\boldsymbol{t} \in \{0,1\}$ and the outcome vector $\mathbf{Y} \in \mathbb{R}^{n\times 1}$. According to Rubin-Neyman causal model [27], two potential outcomes $y_0(\boldsymbol{x})$, $y_1(\boldsymbol{x})$ exist for $\boldsymbol{x}$ with treatments $\{0,1\}$, respectively.

A. Treatment Effect Estimation

Individual Treatment Estimation. Based on the potential outcomes $y_0(\cdot)$ and $y_1(\cdot)$, we can define the individual treatment effect (ITE) as the difference between two potential outcomes.

$$\tau_{\text{ITE}}(\boldsymbol{x}_i) = y_1(\boldsymbol{x}_i) - y_0(\boldsymbol{x}_i) \tag{1}$$

When only one potential outcome is observed as the assigned treatment t, it is called the *factual outcome* y. In addition, we refer the unobserved potential outcome as the *counterfactual outcome* $\hat{y}$. Given the treatment t_i, the relationship between y and two potential outcomes are

$$y_i = t_i y_1(\boldsymbol{x}) + (1-t_i) y_0(\boldsymbol{x}) \tag{2}$$

With the knowledge above, ITE can be alternatively estimated by comparing the factual outcome and the corresponding counterfactual.

$$\tau_{\text{ITE}}(\boldsymbol{x}_i) = \begin{cases} y_i - \hat{y}, & t_i = 1 \\ \hat{y} - y_i, & t_i = 0 \end{cases} \tag{3}$$

where the counterfactual outcome $\hat{y}$ is unobserved in practice.

Estimating ITE can be transformed to counterfactual inference, which lies precisely in the treatment assignment mechanism. A machine learning model trained to minimise the factual error may over-fit the treated units, but not generalise well to the entire population. This is mainly because the assignment of cases to treatments is typically biased, because cases for which a treatment is known as effective are more likely to receive the same treatment. The distribution of samples may therefore differ significantly between the treated units and the overall units.

Propensity Score. The propensity score technique considers the mechanism of treatment assignment [5]. The propensity score $p(t|\boldsymbol{x})$ is the conditional probability for a unit being assigned to a particular treatment given a set of observed covariates. One widely-adopted parametric model of propensity score $p(t|\boldsymbol{x})$ is the logistic regression:

$$p(t|\boldsymbol{x}) = \frac{1}{1+\exp\left(-\boldsymbol{w}^\top \boldsymbol{x} - \omega_0\right)} \tag{4}$$

where $\boldsymbol{w}$ and ω_0 are estimated by minimizing the negative log-likelihood [10].

Assumptions. Counterfactual inference from observational data always requires further assumptions about the data-generating process [28], [29]. Following the general practise in causal inference literature [14], [3], [24], the following two assumptions are related to *unconfoundedness* (or *ignorability*) that ensures the identifiability of the treatment effect [29]:

Assumption 1. *For all values of $\boldsymbol{x}$, it is possible to observe all treatments with non-zero probability.*

$$p(t|\boldsymbol{x}) > 0, \quad \forall \boldsymbol{x} \text{ and } t \tag{5}$$

Given some values of $\boldsymbol{x}$, the treatment assignment is not deterministic; otherwise, at least for one treatment, the outcomes could never be observed, which is infeasible to estimate the treatment effect.

Assumption 2. *The assignment to treatment t is independent of the outcome y given the covariates $\boldsymbol{x}$, i.e.,*

$$y_0, y_1 \perp t|\boldsymbol{x} \tag{6}$$

With this *unconfoundedness* assumption, the values of the potential outcomes (y_0 and y_1) are independent of the observed treatment, given the set of confounding variables. Namely, for the units with the same covariates $\boldsymbol{x}$, their treatment assignment can be viewed as random. We then have $p(y_1|t,\boldsymbol{x}) = p(y|t,\boldsymbol{x})$ to infer the unknown counterfactual outcomes from the observed datasets, which further leads to causal identification.

B. Optimal Transport

Optimal transport [30], [31], also known as *Earth Mover's Distance* in engineering-related fields, was first introduced by French mathematician Gaspard Monge [32]. Originally, optimal transport aims to transport a given mass of dirt to a given hole with a minimal effort solution. Due to its appealing ability in improving the accuracy of numerous pattern recognition-related problem, optimal transport has recently received significant attention from the machine learning communities [33], [34]. Applications of the optimal transport include various transport-based learning methods [34], domain adaptation [33], [35], Bayesian inference [36], and sampling [37], [38].

Note that Acciaio et al. [39] use a same name as our method, i.e., causal optimal transport. In fact, our method is totally different from their work with regarding to both the aim and technical methodology. Particularly, Acciaio et al. [39] define an optimal transport over causal couplings to addresses the stochastic analysis problem of filtrations enlargement. By contrast, our paper defines transportation plan in the propensity measure space to improve the causal effect estimation.

IV. Causal Optimal Transport

This work aims to infer the counterfactual outcomes for individual treatment effect (ITE) estimation. To this end, we have n observed samples $\{(\boldsymbol{x}_i, t_i, y_i)\}_{i=1}^n$ in the factual space Ω_F. Similarly, the counterfactual space Ω_C includes the samples $\{(\boldsymbol{x}_i, 1-t_i)\}_{i=1}^n$, by assuming that each sample i receives the opposite treatment $1-t_i$.

The Propensity Measure. We infer the counterfactual outcome by leveraging the information $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^n$ from the factual space. The distribution of this information can not be fully captured by the propensity score $p(t|x)$, because $p(t|x)$ is merely the conditional probability of receiving treatment given x. This motivates us to define two propensity measures, on the factual space and on the counterfactual space, respectively.

$$\begin{aligned} \mu(\boldsymbol{x}, t) &= p(\boldsymbol{x})p(t|\boldsymbol{x}) \\ \nu(\boldsymbol{x}, 1-t) &= p(\boldsymbol{x})p(1-t|\boldsymbol{x}) \end{aligned} \tag{7}$$

where $p(t|x)$ estimated by eq. (4) refers to the propensity score of the factual samples. In this paper, we consider the binary treatment settings, hence a unit with treatment t in the factual space will have treatment $1-t$ in the counterfactual space. Accordingly, the propensity score for the counterfactual space can be represented by $p(1-t|x)$. Apparently, the factual space and the counterfactual space share the same covariates but with different propensity measure values $\mu(\boldsymbol{x}, t)$ and $\nu(\boldsymbol{x}, t)$, respectively.

Based on the propensity measures in eq. (7), we define the joint distribution of factual space Ω_F and counterfactual space Ω_C. For brevity, let $\boldsymbol{u} = (\boldsymbol{x}, t)$ denote the factual feature consisting of covariate and treatment in the factual space, and let $\boldsymbol{v} = (\boldsymbol{x}, 1-t)$ denote the counterfactual feature consisting of covariate and treatment in the counterfactual space. The joint distribution on the factual space is then:

$$p_\mu(\boldsymbol{u}, y) = p(y|\boldsymbol{u})\mu(\boldsymbol{u}) \tag{8}$$

Note that the factual outcome y is observed but the counterfactual outcome is unobserved in practice. As the goal of estimating ITE in eq. (3) requires the inference of counterfactual outcome, we use the counterfactual predictor $h : \Omega_C \to \mathbb{R}$ as the proxy of counterfactual outcome, and then we define $p_\nu(\boldsymbol{v}, h)$ as the joint distribution on the counterfactual space Ω_C.

$$p_\nu(\boldsymbol{v}, h) = p(h|\boldsymbol{v})\nu(\boldsymbol{v}) \tag{9}$$

Building on the optimal transport theory, the following part will generate an unbiased h by searching an optimal mapping from the factual joint distribution $p_\mu(\boldsymbol{u}, y)$ to the counterfactual joint distribution $p_\nu(\boldsymbol{v}, h)$.

A. Propensity Measure Transport

Because skewed data distributions naturally arise in causal inference, where the treated or the controlled units occur with reduced frequency, most propensity score based methods overlook the information of minority treated or minority controlled units, and hence suffer from the biased treatment effect estimation. Inspired by the optimal transport theory [31], we propose a novel causal optimal transport (CausalOT) method as illustrated in Figure 2. CausalOT exploits the global information of the observational covariates and the treatment to learn an unbiased treatment effect.

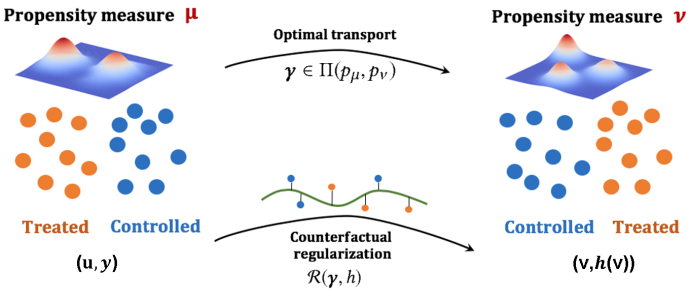


Fig. 2: Causal optimal transport method.

According to the propensity measures on the factual space and on the counterfactual space, we have two joint distributions $p_\mu(\boldsymbol{u}, y)$ on Ω_F and $p_\nu(\boldsymbol{v}, h)$ on Ω_C, as defined in eqs. (8) and (9). When the factual joint distribution is similar to the counterfactual joint distribution, the data is close to randomized experimental study that involves less treatment selection bias. According to the optimal transport theory [31], we assume that the discrepancy between two spaces Ω_F and Ω_C is due to an unknown mapping $\boldsymbol{\gamma} : \Omega_F \to \Omega_C$. In this work, we propose to find a counterfactual predictor h by minimizing the transport loss corresponding to the unknown $\boldsymbol{\gamma}$, so the optimal $\boldsymbol{\gamma}$ can be identified as a mapping that transports $p_\mu(\boldsymbol{u}, y)$ to $p_\nu(\boldsymbol{v}, h)$ with the minimum transport cost. By defining $\Pi(p_\mu, p_\nu)$ as the space of probability distributions over $\mathbb{R}^2$ with marginals p_μ and p_ν, the optimal $\boldsymbol{\gamma} \in \Pi(p_\mu, p_\nu)$ minimizes the following quantity:

$$\begin{aligned} &\min_{\boldsymbol{\gamma}, h} \int_{\Omega_F \times \Omega_C} \mathcal{L}\left(\boldsymbol{u}, \boldsymbol{y}; \boldsymbol{v}, h(\boldsymbol{v})\right) d\,\boldsymbol{\gamma}\left(\boldsymbol{u}, \boldsymbol{y}; \boldsymbol{v}, h(\boldsymbol{v})\right) \\ &\text{s.t. } \boldsymbol{\gamma} \in \Pi(p_\mu, p_\nu) \end{aligned} \tag{10}$$

where the joint cost function is

$$\mathcal{L}\left(\boldsymbol{u}, \boldsymbol{y}; \boldsymbol{v}, h(\boldsymbol{v})\right) = \lambda \mathcal{C}(\boldsymbol{u}, \boldsymbol{v}) + \mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v})) \tag{11}$$

Note that $\mathcal{C}$ measures the discrepancy between the factual features (i.e., covariates and treatment) $\boldsymbol{u}$ and the counterfactual $\boldsymbol{v}$. In Section Section IV-B, we design a novel $\mathcal{R}(\boldsymbol{\gamma}, h)$ as a counterfactual loss that measures the discrepancy between the factual outcome $\boldsymbol{y}$ and the predicted counterfactual outcome $h(\boldsymbol{v})$. $\mathcal{R}(\boldsymbol{\gamma}, h)$ is continuous and differentiable with respects to $h(\cdot)$. Although the problem (10) does not involve any regularization on the mapping $\boldsymbol{\gamma}$, it is essentially for the sake of simplicity [35]. By minimizing the cost $\mathcal{L}$, the optimal γ

maps the factual samples to the counterfactual samples with similar features and outcomes.

So far the optimization task as in eq. (10) has no analytic solution. Rather than working with distribution functions p_μ and p_ν, we will relax it to the general case of transport between discrete measures. The discrete uniform distribution (i.e., the probability of each sample is equal) was usually adopted for the discrete settings [31]. In our case, uniform distribution can not fully exploit the joint distribution information underlying in the factual and the counterfactual spaces. For the joint distributions p_μ and p_ν as in eqs. (8) and (9), two empirical measures (i.e., $\boldsymbol{p}_\mu$ and $\boldsymbol{p}_\nu$) are proposed as the discrete approximations to them:

$$\begin{aligned}\boldsymbol{p}_\mu &= \left[\frac{p_\mu(\boldsymbol{u}_1, y_1)}{\sum_i^n p_\mu(\boldsymbol{u}_i, y_i)}, \cdots, \frac{p_\mu(\boldsymbol{u}_n, y_n)}{\sum_i^n p_\mu(\boldsymbol{u}_i, y_i)}\right]^\top \\ \boldsymbol{p}_\nu &= \left[\frac{\nu(\boldsymbol{v}_1)}{\sum_i^n \nu(\boldsymbol{v}_i)}, \cdots, \frac{\nu(\boldsymbol{v}_n)}{\sum_i^n \nu(\boldsymbol{v}_i)}\right]^\top\end{aligned} \tag{12}$$

where $p_\mu(\boldsymbol{u}_i, y_i)$ is the propensity measure on the observed factual sample $(\boldsymbol{u}_i, y_i)$, and it can be easily computed according to eq. (8). Similarly, computing $\boldsymbol{p}_\nu$ requires $p_\nu(\boldsymbol{v_i}, h(\boldsymbol{v_i}))$ in eq. (9). As $h(\cdot)$ is a proxy of unknown counterfactual outcome which also requires to be estimated, the probability $p(h(\boldsymbol{v})|\boldsymbol{v})$ is assumed to be uniform for simplicity. Consider the fact that no prior knowledge is available for the probability of counterfactual outcomes, we assume a uniform distribution in which all counterfactual outcomes are equally likely given $\boldsymbol{v}$. The probabilistic coupling between $\boldsymbol{p}_\mu$ and $\boldsymbol{p}_\nu$ is formulated as

$$\Pi(\boldsymbol{p}_\mu, \boldsymbol{p}_\nu) = \{\boldsymbol{\gamma} \in \mathbb{R}^{n\times n} | \boldsymbol{\gamma}\mathbf{1}_n = \boldsymbol{p}_\mu,\ \boldsymbol{\gamma}^\top \mathbf{1}_n = \boldsymbol{p}_\nu\} \tag{13}$$

where $\mathbf{1}_n$ represents the n-dimensional vector of ones. $\Pi(\boldsymbol{p}_\mu, \boldsymbol{p}_\nu)$ refers to a set of all admissible couplings between $\boldsymbol{p}_\mu$ and $\boldsymbol{p}_\nu$. $\boldsymbol{\gamma}_{i,j}$ represents the amount of mass shifted from the bin $\boldsymbol{p}_{\mu_i}$ to $\boldsymbol{p}_{\nu_j}$. In our case, the matrix $\boldsymbol{\gamma}$ describes a probabilistic matching of the samples in the factual space and the counterfactual space. Consequently, the transport map $\boldsymbol{\gamma}$ turns to be a coupling matrix where $\boldsymbol{\gamma}_{i,j}$ describes the amount of mass flowing from bin i to bin j.

Under discrete measures, the optimal transport problem as in eq. (10) can be generalized for the counterfactual inference as

$$\min_{\boldsymbol{\gamma}, h} \lambda\langle\boldsymbol{\gamma}, \mathbf{C}\rangle + \mathcal{R}(\boldsymbol{\gamma}, h) \quad \text{s.t. } \boldsymbol{\gamma} \in \Pi(\boldsymbol{p}_\mu, \boldsymbol{p}_\nu) \tag{14}$$

where λ is a hyperparameter to balance the alignment of features (i.e., $\boldsymbol{u}$ and $\boldsymbol{v}$) and outcomes (i.e., y and $h(\cdot)$). Since counterfactual outcome prediction $h(\cdot)$ is the main task for treatment effect estimation, λ should be less than 1 from intuition. Specifically, $\langle\cdot,\cdot\rangle$ denotes the Frobenius dot-product in the feature space. The matrix $\mathbf{C} = [\mathbf{C}_{i,j}] \in \mathbb{R}^{n\times n}$ denotes the cost matrix, in which each element $\mathbf{C}_{i,j}$ represents the cost of moving a probability mass $\boldsymbol{u}$ to $\boldsymbol{v}$. We define $\mathbf{C}_{i,j}$ as the squared Euclidean distance between i-th and j-th sample, i.e.,

$$\mathbf{C}_{ij} = \|\boldsymbol{u}_i - \boldsymbol{v}_j\|_2^2 \tag{15}$$

In the following section, the newly designed $\mathcal{R}(\boldsymbol{\gamma}, h)$ as a counterfactual loss in eq.(14) will be discussed in detail.

B. Counterfactual Loss

Recall that $h(\cdot)$ predicts a counterfactual outcome given an input of $\boldsymbol{v}$ in terms of covariates and treatment. The counterfactual loss term $\mathcal{R}(\boldsymbol{\gamma}, h)$ is proposed to reduce the shift bias when transporting the propensity measures. As the mapping $\boldsymbol{\gamma}$ transports the propensity measure μ to ν, the transported outcome $\boldsymbol{\gamma}(y)$ should be also aligned with h to guarantee $p(y|x,t) = p(h|x,1-t)$ in eq. (8). To achieve this, we use the following loss

$$\mathcal{R}(\boldsymbol{\gamma}, h) = \frac{1}{2}\sum_{j=1}^n (\hat{y}_j - h(\boldsymbol{v}_j))^2 \tag{16}$$

Based on the properties of the Euclidean quadratic loss [31], we have $\hat{y}_j$ to be a weighted mean of factual outcomes $\{y_j\}_{j=1}^n$ as follows.

$$\hat{y}_j = \frac{\sum_j \boldsymbol{\gamma}_{i,j} y_j}{p_\mu(y_j)} \tag{17}$$

where $p_\mu(y_j)$ represents the j-th element in the distribution vector $\boldsymbol{p_\nu}$. Because the factual samples with larger propensity scores indicate that these units are more likely to be treated, the estimated $\hat{y}$ weighted by the inverse probability of treatment results in the factual unit with less propensity score to contribute more. Namely, to reduce the treatment assignment bias, eq. (17) simulates a population in which baseline covariates are independent of the treatment assignment for the training of an unbiased predictor $h(\cdot)$.

Given the unconfoundedness as in Assumption 2 and Assumption 1, for simplicity, we use the linear model for the counterfactual predictor: h is the linear hypothesis conditioned on $\boldsymbol{v}$, i.e., $h(\boldsymbol{v}) = \boldsymbol{\beta}^\top \boldsymbol{v} + \xi$, and $\boldsymbol{\beta}^\top \in \mathbb{R}^{1\times d}$ is the weight vector. Accordingly, $\mathcal{R}(\boldsymbol{\gamma}, h)$ can be written as

$$\begin{aligned}\mathcal{R}(\boldsymbol{\gamma}, h) &= \frac{1}{2}\sum_{j=1}^n \left(\hat{y}_j^2 + h(\boldsymbol{v}_j)^2 - 2\hat{y}_j h(\boldsymbol{v}_j)\right) \\ &= \frac{1}{2}\sum_{j=1}^n \left(\hat{y}_j^2 + (\boldsymbol{\beta}^\top \boldsymbol{v}_j + \xi)^2 - 2\hat{y}_j(\boldsymbol{\beta}^\top \boldsymbol{v}_j + \xi)\right) \\ &= \frac{1}{2}\mathrm{tr}(\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^\top) - \mathrm{tr}(\hat{\boldsymbol{y}}\boldsymbol{\beta}^\top \mathbf{V}^\top) - \mathrm{tr}(\xi\hat{\boldsymbol{y}}\mathbf{1}_n) + \xi^2 \\ &+ \frac{1}{2}\mathrm{tr}(\mathbf{V}\boldsymbol{\beta}\boldsymbol{\beta}^\top \mathbf{V}^\top) + \xi\mathrm{tr}((\mathbf{1}_n^\top \otimes \boldsymbol{\beta})\mathbf{V}^\top)\end{aligned} \tag{18}$$

where $\hat{\boldsymbol{y}} = [\hat{y}_1, \cdots, \hat{y}_n]^\top \in \mathbb{R}^{n\times 1}$, $\mathbf{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n]^\top \in \mathbb{R}^{n\times d}$, and $\otimes$ is the Kronecker product. Based on eq. (17), the counterfactual outcome $\hat{\boldsymbol{y}}$ can be computed based on the factual outcome vector $\boldsymbol{y} = [y_1, \cdots, y_n]^\top$.

$$\hat{\boldsymbol{y}} = \mathrm{diag}(\boldsymbol{p_\mu})^{-1}\boldsymbol{\gamma}\boldsymbol{y} = \mathbf{D}\boldsymbol{\gamma}\boldsymbol{y} \tag{19}$$

Let $\mathbf{D} = \mathrm{diag}(\boldsymbol{p}_\mu)^{-1} \in \mathbb{R}^{n\times n}$ be a diagonal matrix with the vector $\boldsymbol{p}_\mu$ on the diagonal. Using eq. (19), the counterfactual

loss can be further written as

$$\begin{aligned}\mathcal{R}(\boldsymbol{\gamma}, h) &= \frac{1}{2}\operatorname{tr}\left(\mathbf{D}\boldsymbol{\gamma}\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{\gamma}^\top\mathbf{D}^\top\right) + \xi\operatorname{tr}((\mathbf{1}_n^\top\otimes\boldsymbol{\beta})\mathbf{V}^\top) + \xi^2 \\ &\quad + \frac{1}{2}\operatorname{tr}(\mathbf{V}\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{V}^\top) - \operatorname{tr}(\mathbf{D}\boldsymbol{\gamma}\boldsymbol{y}\boldsymbol{\beta}^\top\mathbf{V}^\top) - \xi\operatorname{tr}(\mathbf{D}\boldsymbol{\gamma}\boldsymbol{y}\mathbf{1}_n) \\ &= \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\gamma}\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{\gamma}^\top\mathbf{D}^\top\mathbf{D}\right) + \xi\operatorname{tr}((\mathbf{1}_n^\top\otimes\boldsymbol{\beta})\mathbf{V}^\top) + \xi^2 \\ &\quad - \operatorname{tr}(\boldsymbol{\gamma}\boldsymbol{y}\boldsymbol{\beta}^\top\mathbf{V}^\top\mathbf{D} + \boldsymbol{\gamma}\xi\boldsymbol{y}\mathbf{1}_n\mathbf{D}) + \frac{1}{2}\operatorname{tr}(\mathbf{V}\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{V}^\top)\end{aligned} \tag{20}$$

Substituting eq. (20) into the objective as in eq. (14), we organize the reformulated objective function $\mathcal{L}$ as follows:

$$\begin{aligned}\mathcal{L} &= \operatorname{tr}(\boldsymbol{\gamma}\lambda\mathbf{C}^\top) + \mathcal{R}(\boldsymbol{\gamma}, h) \\ &= \operatorname{tr}(\boldsymbol{\gamma}\lambda\mathbf{C}^\top) + \frac{1}{2}\operatorname{tr}(\mathbf{V}\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{V}^\top) + \xi\operatorname{tr}((\mathbf{1}_n^\top\otimes\boldsymbol{\beta})\mathbf{V}^\top) + \xi^2 \\ &\quad - \operatorname{tr}(\boldsymbol{\gamma}\boldsymbol{y}\boldsymbol{\beta}^\top\mathbf{V}^\top\mathbf{D} + \boldsymbol{\gamma}\xi\boldsymbol{y}\mathbf{1}_n\mathbf{D}) + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\gamma}\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{\gamma}^\top\mathbf{D}^\top\mathbf{D}\right) \\ &= \operatorname{tr}(\boldsymbol{\gamma}\Theta) + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\gamma}\Phi\boldsymbol{\gamma}^\top\Psi\right) + \xi\operatorname{tr}((\mathbf{1}_n^\top\otimes\boldsymbol{\beta})\mathbf{V}^\top) \\ &\quad + \frac{1}{2}\operatorname{tr}(\mathbf{V}\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{V}^\top) + \xi^2\end{aligned} \tag{21}$$

where

$$\begin{aligned}\Theta &= \lambda\mathbf{C}^\top - (\xi\boldsymbol{y}\mathbf{1}_n + \boldsymbol{y}\boldsymbol{\beta}^\top\mathbf{V}^\top)\mathbf{D} \\ \Phi &= \boldsymbol{y}\boldsymbol{y}^\top \\ \Psi &= \mathbf{D}^\top\mathbf{D}\end{aligned} \tag{22}$$

V. Proximal Point Algorithm

The objective function as in eq. (21) is smoothly separable according to the counterfactual predictor $h(\cdot)$ (in terms of $\boldsymbol{\beta}, \xi$) and $\boldsymbol{\gamma}$. As the objective function is with a novel counterfactual loss, $\boldsymbol{\gamma}$ can not be directly solved by traditional solver, and we propose to solve it based on a generalized proximal point algorithm.

Updating $\boldsymbol{\gamma}$. When h is fixed, eq. (21) can be considered as an optimal transport problem regularized by $\mathcal{R}$ to fit the counterfactual outcomes. However, traditional optimal transport algorithm is not appropriate for this regularized problem. Inspired by [40], [41], we propose to add the Bregman divergence to the subproblem of γ in eq. (21). The proximal point iteration for optimizing $\boldsymbol{\gamma}$ can then be solved by

$$\boldsymbol{\gamma}^{(l+1)} = \arg\min_{\boldsymbol{\gamma}} \langle\nabla\mathcal{L}(\boldsymbol{\gamma}^{(l)}), \boldsymbol{\gamma}\rangle + \alpha d_B\left(\boldsymbol{\gamma}, \boldsymbol{\gamma}^{(l)}\right) \tag{23}$$

The Bregman divergence d_B is defined as the proximal operator associated with entropy function $g(z) = \sum_{i,j}^n z_{i,j}\log(z_{i,j})$, and we have

$$d_B(\boldsymbol{\gamma}, \boldsymbol{\gamma}^{(l)}) = \sum_{i,j}^n \boldsymbol{\gamma}_{i,j}\left(\log\boldsymbol{\gamma}_{i,j} - \log\boldsymbol{\gamma}_{i,j}^{(l)}\right) - \sum_{i,j}^n \boldsymbol{\gamma}_{i,j} + \sum_{i=1}^n \boldsymbol{\gamma}_{i,j}^{(l)} \tag{24}$$

Substituting Bregman divergence (24) into proximal point iteration (23), with simplex constraints, we have

$$\begin{aligned}\boldsymbol{\gamma}^{(l+1)} = \arg\min_{\boldsymbol{\gamma}} &\left\langle\nabla\mathcal{L}(\boldsymbol{\gamma}^{(l)}), \boldsymbol{\gamma}\right\rangle + \alpha\sum_{i,j}^n \boldsymbol{\gamma}_{i,j}\left(\log\boldsymbol{\gamma}_{i,j} - \log\boldsymbol{\gamma}_{i,j}^{(l)}\right) \\ &- \alpha\sum_{i,j}^n \boldsymbol{\gamma}_{i,j} + \alpha\sum_{i=1}^n \boldsymbol{\gamma}_{i,j}^{(l)}\end{aligned} \tag{25}$$

We define $H(\boldsymbol{\gamma}) = \sum_{i,j}^n \boldsymbol{\gamma}_{i,j}(\log\boldsymbol{\gamma}_{i,j} - 1)$ and $\boldsymbol{\gamma}^{(l+1)}$ is reformulated (26) as

$$\begin{aligned}\boldsymbol{\gamma}^{(l+1)} &= \arg\min_{\boldsymbol{\gamma}} \left\langle\nabla\mathcal{L}(\boldsymbol{\gamma}^{(l)}) - \alpha\log\boldsymbol{\gamma}^{(l)}, \boldsymbol{\gamma}\right\rangle + \alpha H(\boldsymbol{\gamma}) + \alpha\sum_{i=1}^n \boldsymbol{\gamma}_{i,j}^{(l)} \\ &= \arg\min_{\boldsymbol{\gamma}} \left\langle\nabla\mathcal{L}(\boldsymbol{\gamma}^{(l)}) - \alpha\log\boldsymbol{\gamma}^{(l)}, \boldsymbol{\gamma}\right\rangle + \alpha H(\boldsymbol{\gamma})\end{aligned} \tag{26}$$

where $\boldsymbol{\gamma}_{i,j}^{(l)}$ is a fixed value that is irrelevant to the optimization variable $\boldsymbol{\gamma}$. According to [42], $H(\boldsymbol{\gamma})$ is an entropy that allows eq. (26) to have a closed-form solution and speed up the optimization. Based on eq. (29), the closed-form solution of (26) is provided as follows:

$$\boldsymbol{\gamma}^{(l+1)} = \operatorname{diag}(\boldsymbol{a})\mathbf{K}^{(l)}\operatorname{diag}(\boldsymbol{b}) \tag{27}$$

where $\operatorname{diag}(\boldsymbol{a})$ is the diagonal matrix with the vector $\boldsymbol{a}$ on the diagonal. The updates of $(\boldsymbol{a}, \boldsymbol{b})$ in Sinkhorn's algorithm [43] are defined as

$$\begin{aligned}\boldsymbol{a} = \frac{\boldsymbol{p}_{\boldsymbol{\mu}}}{\mathbf{K}^{(l)}\boldsymbol{b}} \quad \text{and} \quad \boldsymbol{b} &= \frac{\boldsymbol{p}_{\boldsymbol{\nu}}}{{\mathbf{K}^{(l)}}^\top\boldsymbol{a}} \\ \mathbf{K}^{(l)} &= \boldsymbol{\gamma}^{(l)} \odot \exp\left(\frac{-\nabla\mathcal{L}(\boldsymbol{\gamma}^{(l)})}{\alpha}\right)\end{aligned} \tag{28}$$

As the objective function $\mathcal{L}$ in eq. (21) is differentiable and quadratic w.r.t. $\boldsymbol{\gamma}$, we compute its derivative as

$$\nabla\mathcal{L}(\boldsymbol{\gamma}) = \frac{1}{2}\Psi^\top\boldsymbol{\gamma}\Phi^\top + \Theta^\top \tag{29}$$

Updating $\boldsymbol{\beta}$ and ξ. When $\boldsymbol{\gamma}$ is fixed, $\boldsymbol{\beta}$ and ξ can be obtained by the gradient descent algorithm:

$$\begin{aligned}\nabla\mathcal{L}(\xi) &= \operatorname{tr}((\mathbf{1}_n^\top\otimes\boldsymbol{\beta})\mathbf{V}^\top) - \boldsymbol{\gamma}^\top\boldsymbol{y}\mathbf{1}_n\mathbf{D} + 2\xi \\ \nabla\mathcal{L}(\boldsymbol{\beta}) &= 2\mathbf{V}^\top\mathbf{V}\boldsymbol{\beta} + \xi(\mathbf{1}_n\otimes\mathbf{1}_d)\mathbf{V}^\top + \boldsymbol{y}\mathbf{1}_d\mathbf{V}^\top\mathbf{D}\end{aligned} \tag{30}$$

We update $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and ξ iteratively until the objective function eq. (21) converges. All steps are summarized in Algorithm 1.

VI. Theoretical Results

In this section, we provide theoretical justification on the optimal transport for the counterfactual outcome inference. Specifically, we derive an upper bound on the counterfactual generalization error with respect to the propensity measures. We assume that the factual space and the counterfactual space are with the same ground-truth outcome function f. This is reasonable because from the unconfoundedness in Assumption 2, units with similar covariates (e.g., *healthy status* or *age*) tend to have similar outcomes, no matter being treated or untreated (i.e., controlled). If the hypothesis $h \in \mathcal{H}$ learned by

Algorithm 1 Causal Optimal Transport (CausalOT)

Input: Factual units $(\boldsymbol{u}_1, \boldsymbol{y}_1) \cdots (\boldsymbol{u}_n, \boldsymbol{y}_n)$ and counterfactual units $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n$.

Initialize the weight $\boldsymbol{\beta} = \mathbf{1}$ and residual $\xi = \mathbf{0}$
Estimate the propensity score $p(t|\boldsymbol{x})$ for the observed covariates by eq. (4).
Compute the propensity score measures $\mu(\boldsymbol{u})$ and $\nu(\boldsymbol{v})$ for $\boldsymbol{u}$ and $\boldsymbol{v}$ by eq. (7).
Estimate two joint distributions $p_\mu(\boldsymbol{u}_i, y_i)$ and $p_\nu(\boldsymbol{v}_i)$ by eqs. (8) and (9).
Normalize the joint distribution $\boldsymbol{p}_\mu$ and $\boldsymbol{p}_\nu$ by eq. (12).
Set the cost matrix $[\mathcal{C}]_{ij} = \|\boldsymbol{u}_i - \boldsymbol{v}_j\|_2^2$, $\boldsymbol{a} = \boldsymbol{p}_\mu$, $\boldsymbol{b} = \boldsymbol{p}_\nu$, $\mathbf{D} = \mathrm{diag}(\boldsymbol{p}_\mu)^{-1}$,
repeat
  Compute the parameters Θ, Φ, Ψ by eq. (22).
  Update the gradient

$$\nabla\mathcal{L}(\boldsymbol{\gamma}) = \frac{1}{2}\Psi^\top\boldsymbol{\gamma}\Phi^\top + \Theta^\top$$

  Update $\boldsymbol{a} = \frac{\boldsymbol{p}_\mu}{\mathbf{K}^{(l)}\boldsymbol{b}}$, $\boldsymbol{b} = \frac{\boldsymbol{p}_\nu}{\mathbf{K}^{(l)\top}\boldsymbol{a}}$
  Update $\mathbf{K}^{(l)} = \boldsymbol{\gamma}^{(l)} \& \odot \exp\left(\frac{-\nabla\mathcal{L}(\boldsymbol{\gamma}^{(l)})}{\epsilon}\right)$
  Update $\boldsymbol{\gamma}^{(l+1)} = \mathrm{diag}(\boldsymbol{a})\mathbf{K}^{(l)}\mathrm{diag}(\boldsymbol{b})$
  Update gradient

$$\nabla\mathcal{L}(\xi) = \mathrm{tr}((\mathbf{1}_n^\top \otimes \boldsymbol{\beta})\mathbf{V}^\top) - \boldsymbol{\gamma}^\top\boldsymbol{y}\mathbf{1}_n\mathbf{D} + 2\xi$$

  Update

$$\mathcal{L}(\boldsymbol{\beta}) = 2\mathbf{V}^\top\mathbf{V}\boldsymbol{\beta} + \xi(\mathbf{1}_n \otimes \mathbf{1}_d)\mathbf{V}^\top + \boldsymbol{y}\mathbf{1}_d\mathbf{V}^\top\mathbf{D}$$

  Update $\xi^{(l+1)} = \xi^{(l)} - \nabla\mathcal{L}(\xi^{(l)})$
  Update $\boldsymbol{\beta}^{(l+1)} = \boldsymbol{\beta}^{(l)} - \nabla\mathcal{L}(\boldsymbol{\beta}^{(l)})$
until convergence
Compute the counterfactual outcome $h(\boldsymbol{v}) = \boldsymbol{\beta}^\top\boldsymbol{v} + \xi$
Compute ITE using $h(\boldsymbol{v})$ and $\boldsymbol{y}$ by eq. (3)

Output: ITE

CausalOT disagrees with f, the expected counterfactual error is

$$\epsilon_{\boldsymbol{\nu}}(h, f) = \mathbb{E}_{\boldsymbol{v}\sim\boldsymbol{\nu}}[l(h(\boldsymbol{v}), f(\boldsymbol{v}))] \tag{31}$$

where the loss function $l(h(\boldsymbol{v}), f(\boldsymbol{v})) = (h(\boldsymbol{v}) - f(\boldsymbol{v}))^2$ is denoted as $l(\boldsymbol{v})$ for short in the following text. Since covariates (e.g., salary or age) and treatments in observational study are finite values, it is reasonable to assume that $\boldsymbol{v}$ is sampled from a compact (bounded and closed) set O.

Given two samples $\boldsymbol{v}_1, \boldsymbol{v}_2$ from the compact set O, we have $\boldsymbol{\theta} \in O$ that satisfies eq. (32) because of the mean value theorem.

$$\|l(\boldsymbol{v}_1) - l(\boldsymbol{v}_2)\|_2 = \|l'(\boldsymbol{\theta})(\boldsymbol{v}_1 - \boldsymbol{v}_2)\|_2 \leq \sup_{\theta\in O}\|l'(\boldsymbol{\theta})\|_2\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 \tag{32}$$

The squared loss function l is continuously differentiable, and compact set O is bounded and closed. Therefore, $\|l'(\boldsymbol{\theta})\|_2$ has upper bound, and we have the local Lipschitz condition

$$\|l(\boldsymbol{v}_1) - l(\boldsymbol{v}_2)\|_2 \leq \kappa\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 \tag{33}$$

with Lipschitz constant κ. Based on eq. (33), we apply Wasserstein distance to analyze the generalization error between the propensity measures μ and ν.

Lemma 1. *For every $h \in \mathcal{H}$, the following error holds*

$$\epsilon_\nu(h, f) \leq 4\epsilon_\mu(h, f) + 2\kappa W_1(\mu, \nu) + \eta \tag{34}$$

where $\eta = 4\epsilon_\mu(h^*, f) + 2\epsilon_\nu(h^*, f)$ is the minimal combined error achieved by the optimal hypothesis h^*.

Proof. The Kantorovich-Rubinstein theorem shows that the dual representation of the 1-Wasserstein distance can be written as a form of

$$W_1(\mu, \nu) = \sup_{\|l\|\leq 1}\mathbb{E}_{\boldsymbol{u}\sim\mu}[l(\boldsymbol{u})] - \mathbb{E}_{\boldsymbol{v}\sim\nu}[l(\boldsymbol{v})] \tag{35}$$

where the Lipschitz semi-norm $\|l\|$ is defined as $\sup|l(\boldsymbol{u}) - l(\boldsymbol{v})|/\rho(\boldsymbol{u}, \boldsymbol{v})$ and ρ is a distance function. Given the definition of η, we know $h^* = \operatorname{argmin}_h 4\epsilon_\mu(h, f) + 2\epsilon_\nu(h, f)$. Followed by the polarization identity, the error $\epsilon_\nu(h, f) = \mathbb{E}_{\boldsymbol{v}\in\nu}[(h(\boldsymbol{v}) - f(\boldsymbol{v}))^2]$ can be written as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{v}\in\nu}[(h - h^* + h^* - f)^2] \\
&\quad\leq 2\mathbb{E}_{\boldsymbol{v}\in\nu}[(h^* - f)^2] + 2\mathbb{E}_{\boldsymbol{v}\in\nu}[(h^* - h)^2] \\
&\quad= 2\epsilon_\nu(h^*, f) + 2\epsilon_\nu(h^*, h) \\
&\quad= 2\epsilon_\nu(h^*, f) + 2\epsilon_\mu(h, h^*) + 2(\epsilon_\nu(h, h^*) - \epsilon_\mu(h, h^*)) \\
&\quad= 2\epsilon_\nu(h^*, f) + 2\epsilon_\mu(h, h^*) + 2(\mathbb{E}_\nu[l(h, h^*)] - \mathbb{E}_\mu[l(h, h^*)]) \\
&\quad\overset{(33)}{\leq} 2\epsilon_\nu(h^*, f) + 2\epsilon_\mu(h, h^*) + 2\sup_{\|l\|\leq\kappa}\mathbb{E}_\nu[l(h, h^*)] - \mathbb{E}_\mu[l(h, h^*)] \\
&\quad\leq 2\epsilon_\nu(h^*, f) + 4\epsilon_\mu(h, f) + 4\epsilon_\mu(h^*, f) + 2\kappa W_1(\mu, \nu) \\
&\quad= 4\epsilon_\mu(h, f) + 2\kappa W_1(\mu, \nu) + \eta
\end{aligned}$$

□

With Lemma 1, we have proved the generalization error of applying Wasserstein distance between the true probability measures μ and ν. In order to compute the generalization bounds for finite samples rather than the true population measures, we use two empirical measures $\hat{\mu}$ and $\hat{\nu}$ as discrete approximations of μ and ν. Specifically, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \delta_i$ is defined on independent samples $\{\boldsymbol{u}_i\}_{i=1}^n$ drawn from μ. Similarly, we define $\hat{\nu} = \frac{1}{n}\sum_{j=1}^n \delta_j$ on $\{\boldsymbol{v}_i\}_{i=1}^n$ for ν. To further prove the generalization error bound of h using the empirical measures, we give an important theorem as follows.

Theorem 1. *[44] Given a probability measure μ and its associated empirical measure $\hat{\mu}$, then for any $\epsilon > 0$, there exists $n \geq n_0 \max\left(\epsilon^{-(d_1+2)}, 1\right)$, we have*

$$\mathbb{P}\left[W_1(\mu, \hat{\mu}) > \epsilon\right] \leq \exp\left(-\frac{\zeta}{2}n\epsilon^2\right) \tag{36}$$

where n_0, d_1 and ζ are constants.

Theorem 1 shows the convergence of $\hat{\mu}$ to its true measure μ w.r.t. the Wasserstein metric. We can now use Theorem 1 in combination with Lemma 1 to prove the following theorem.

Theorem 2. *For any $\varepsilon > 0$ with probability at least $1-\varepsilon$, the following bound holds for all hypothesis $h \in \mathcal{H}$:*

$$\epsilon_\nu(h,f) \leq 4\epsilon_\mu(h,f) + 2W_1(\hat{\mu},\hat{\nu}) + 4\sqrt{\frac{2\log(1/\varepsilon)}{n\zeta}} + \eta \quad (37)$$

where $\eta = 4\epsilon_\mu(h^*,f) + 2\epsilon_\nu(h^*,f)$ is the combined error of the optimal hypothesis h^*.

Proof. Follow Lemma 1, we have

$$\begin{aligned}
\epsilon_\nu(h,f) &\leq 2\epsilon_\nu(h^*,f) + 2\epsilon_\nu(h^*,h) \\
&= 2\epsilon_\nu(h^*,f) + 2\epsilon_\mu(h^*,h) + (2\epsilon_\nu(h^*,h) - 2\epsilon_\mu(h,h^*)) \\
&\leq 2\epsilon_\nu(h^*,f) + 2\epsilon_\mu(h^*,h) + 2W_1(\mu,\nu) \\
&\leq 2\epsilon_\nu(h^*,f) + 4\epsilon_\mu(h^*,f) + 4\epsilon_\mu(h,f) + 2W_1(\mu,\nu) \\
&= 4\epsilon_\mu(h,f) + 2W_1(\mu,\nu) + \eta \\
&\leq 4\epsilon_\mu(h,f) + 2W_1(\mu,\hat{\mu}) + 2W_1(\hat{\mu},\nu) + \eta \\
&= 4\epsilon_\mu(h,f) + 2W_1(\mu,\hat{\mu}) + 2W_1(\hat{\mu},\hat{\nu}) + 2W_1(\hat{\nu},\nu) + \eta \\
&= 4\epsilon_\mu(h,f) + 4\sqrt{\frac{2\log(1/\varepsilon)}{n\zeta}} + 2W_1(\hat{\mu},\hat{\nu}) + \eta
\end{aligned}$$

□

Lemma 1 and Theorem 2 ensure that the error of counterfactual predictor h is bounded by the Wasserstein distance for empirical measures. This means that reformulating the counterfactual inference as optimizing the task as in eq. (14) would have good generalization for the treatment effect estimation. With the error bound of counterfactual outcome, the treatment effect estimation in eq. (3) is thus theoretically guaranteed to be accurate.

VII. Experiments

Since the ground truth treatment effects are rarely available in real-world data, evaluating the performance of causal inference methods is a challenging task. In this section, we adopt four benchmark datasets in causal inference, i.e., `IHDP`, `News`, `Twins` and `Jobs`, among which three datasets have known two-sides (the factual and the counterfactual) outcomes.

A. Baselines

We compare the proposed CausalOT method with methods from different categories: *linear regression based methods*: Ordinary Least Squares (OLS-1,OLS-2) [45]; *classical causal methods*: Doubly Robust Linear Regression (DoubleRobust) [7], Propensity Score Matching (PscoreMatch) [5]; *tree and forest based methods*: Bayesian Additive Regression Trees (BART) [24], Causal Random Forest (CF) [13]; and *representation based methods*: Balancing Neural Network (BalanceNN) [15]. All those compared methods can predict the unknown counterfactual outcomes and then apply the results for individual treatment estimation (ITE).

- OLS-1 [45] takes the treatment as an input feature and predicts the outcome by least square regression. OLS-2 [45] uses two separate least squares regressions to fit the treated and controlled units respectively.
- PscoreMatch [5] matches the controlled units which received no treatment with those treated units which received the treatment, based on the absolute difference between their propensity scores. We apply 5-nearest neighbour matching with replacement, and impose a nearness criterion, i.e., $caliper = 0.05$.
- DoubleRobust [7] is a combination of regression model and propensity score estimation model to estimate the treatment effect robustly.
- BART [24] directly applies a prior function on the covariate and treatment to estimate the potential outcomes, i.e., Bayesian form of the boosted regression trees. The number of regression trees is set as 200.
- CF [13] is an extension of random forest algorithm. We implement CausalForest with 100 causal trees, each of which estimates the treatment effect on the leaves.
- BalanceNN [15] [2] is the balanced representation that maximizes the similarity between the treated and the controlled units for counterfactual outcome prediction. As suggested in [15], we use 2-ReLU representation-only layers, 2-ReLU layers for the added treatment variable, and a single linear output layer.

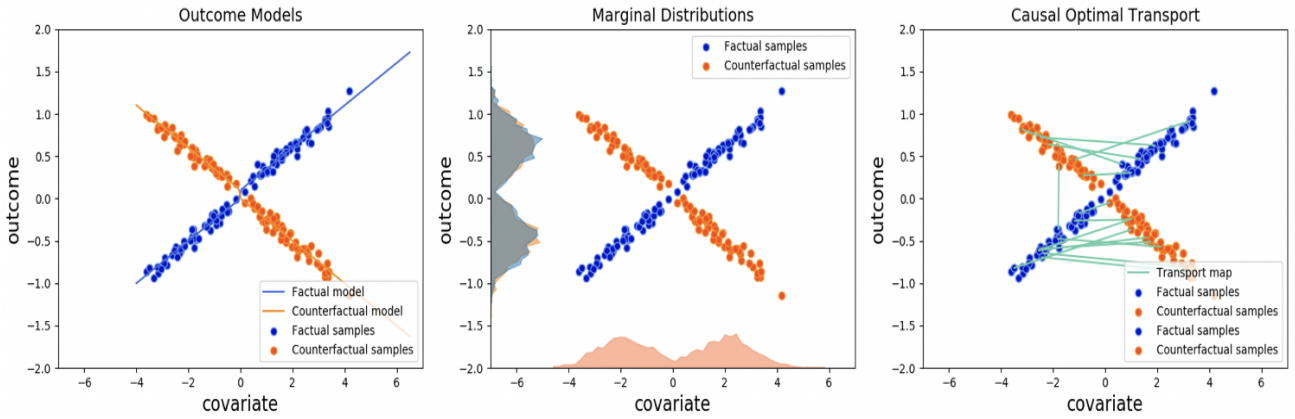
B. Benchmark Data

We use four benchmark datasets for comparison, i.e., `IHDP`, `News`, `Jobs` and `Twins` as summarized in Table I. Among them, `IHDP` is a standard semi-synthetic dataset in the *Infant Health and Development Program* (`IHDP`) [24], which is an observational program designed to study the effect of the specialist visits and parent support on future cognitive and health status of infants. The dataset contains information on 747 infants with 139 treated and 608 control units, and each with 25 real covariates (features). The outcomes are their simulated IQ scores at age 3.

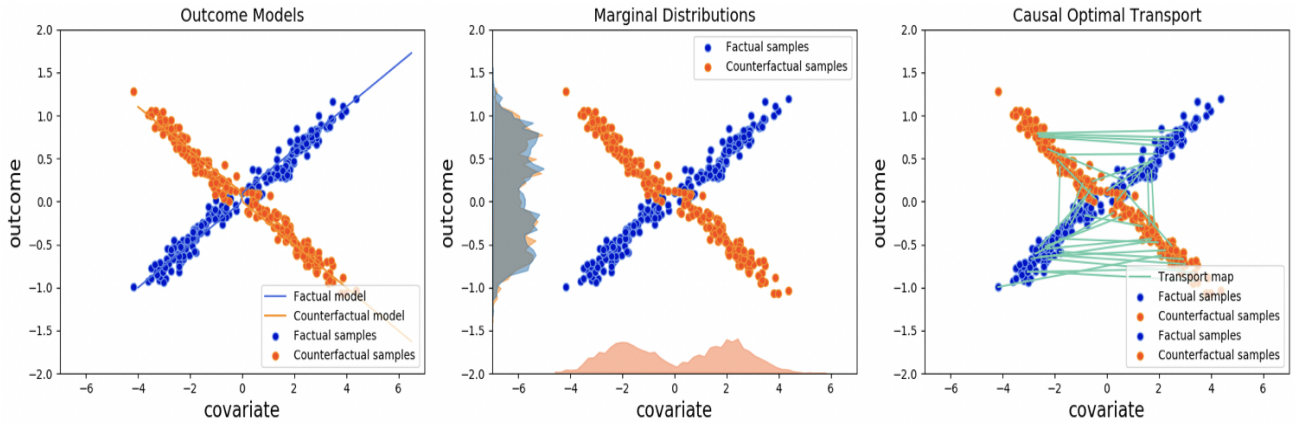
`News` dataset simulates the consumers' opinions on news items affected by different exposures of viewing devices [15]. Each record is one news item represented by word counts $\boldsymbol{x}_i \in \mathbb{R}^{d\times 1}$, where $d = 3477$ is the total number of words. The factual outcome y_i is the reader's opinion on $\boldsymbol{x}_i$ under the treatment t_i. The treatment $t = 0$ or $t = 1$ indicates that the unit views the news via desktop or mobile, respectively. The bias in treatment assignment is simulated as a function of the similarity between the topic distribution of the news items and the two centroids [15]. `Twins` dataset is collected from the twins born in the USA between 1989 to 1991 [46]. Each twin pair has 40 pre-treatment covariates related to the biological parents, the pregnancy and the birth information. We use 5409 twins records that weigh less than 2kg and without missing covariates. For each twin pair we observe both the case $t = 0$ (lighter twin) and $t = 1$ (heavier twin). The outcome is the one-year mortality. To simulate the selection bias, we choose one of the twins as the observation by following the procedures in [26].

`Jobs` dataset is based on a randomized study of a job training program [47], where the treatment is job training and

[2] https://github.com/clinicalml/cfrnet



(a) 100 samples



(b) 200 samples

Fig. 3: Toy examples: CausalOT for counterfactual inference on the synthetic dataset.

the outcome is income after training. We use $2,915$ records with 7-dimensional covariate, among which $2,490$ are control units ($t = 0$). Different from `IHDP`, `News` and `Twins`, the ground truth of ITE in `Jobs` is unknown due to the fact that only one potential outcome was observed (i.e., factual outcome).

We run each algorithm 100 times (except for the `IHDP` dataset, on which we run each algorithm 1,000 times which is the same setting as in [16]) with 70/20/10 train/validation/test splits. To find the optimal setting for the importance of propensity measure transport λ, a grid search in the range of $[10^5, 10^2]$ is performed, where the best performance is achieved with $\lambda = 10^{-2}$. For the hyper-parameter optimization of the baselines, we follow the hyper-parameter optimization code published in the GitHub with their main codes.

TABLE I: Summary of datasets. n is the number of samples, d is the dimension of covariates.

Data	Condition		Property			
	Fact	Counterfact	$t = 1$	$t = 0$	n	d
`IHDP`	✓	✓	139	608	747	25
`News`	✓	✓	2168	2832	5000	3477
`Twins`	✓	✓	1408	3996	5409	40
`Jobs`	✓		297	2915	3212	7

C. Evaluation Metrics

We compare those methods in terms of *Precision in Estimation of Heterogeneous Effect* (PEHE) [24], which evaluates the accuracy of estimated individual treatment effect (ITE), for cases in which only the covariates are observed but without the factual outcomes.

$$\epsilon_{\text{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\tau_{\text{ITE}}(i) - \hat{\tau}_{\text{ITE}}(i)\right)^2 \tag{38}$$

where $\hat{\tau}_{\text{ITE}}(i)$ is the estimated ITE by eq. (3), and $\tau_{\text{ITE}}(i)$ is the ground-truth. A lower PEHE value indicates the more accurate estimation of both the factual and the counterfactual responses.

As a second metric, we consider the absolute error in the estimated *average treatment effect* (ATE) [24].

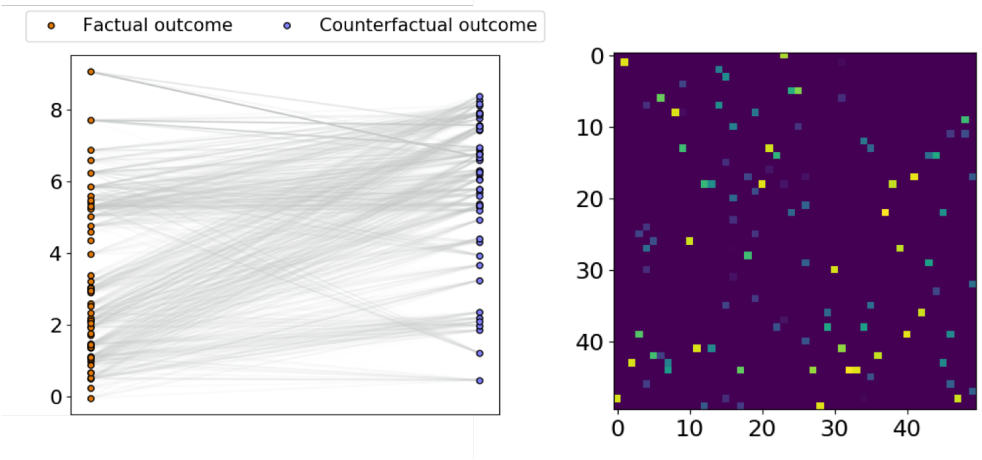
$$\epsilon_{\text{ATE}} = |\tau_{\text{ATE}} - \hat{\tau}_{\text{ATE}}| \tag{39}$$

where $\hat{\tau}_{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_{\text{ITE}}(i)$ is computed by averaging ITE, and $\tau_{\text{ATE}} =$ is the ground truth.

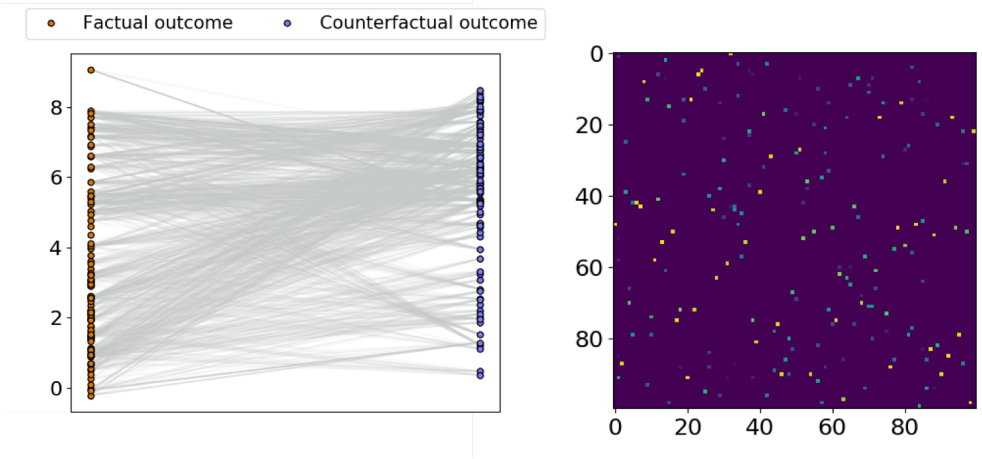
The third metric is the absolute error in the estimated *average treatment effect on the treated* (ATT).

$$\epsilon_{\text{ATT}} = |\tau_{\text{ATT}} - \hat{\tau}_{\text{ATT}}| \tag{40}$$

ATT reflects the treatment effect on the outcome among those who received the treatment $t_i = 1$, i.e., $\tau_{\text{ATT}} =$



(a) Examples of CausalOT on `IHDP` dataset: 50 samples.



(b) Examples of CausalOT on `IHDP` dataset: 100 samples.

Fig. 4: Left column: The transport mapping from factual outcomes and estimated counterfactual outcomes by CausalOT. The grey line between two points indicates the coupling between them. Right column: coupling matrix $\boldsymbol{\gamma}$ over the factual space and the counterfactual space.

$\frac{1}{n_*}\sum_{i=1}^{n_*}\hat{\tau}_{\text{ITE}}(i)$ and n_* is the number of treated units. We average over 5 splits of train/validation/test with ratios $60/30/10$, and then evaluate the criterion on the testing sample in 100 different experiments on both datasets.

As no ground truth is available for ITE in `Jobs`, the policy risk [16] is used for the dataset `Jobs`,

$$\begin{aligned}\mathcal{R}_{pol}(\pi_f) = 1 - \mathbb{E}\left[\hat{y}_t|\pi_f = 1\right]p(\pi_f = 1) \\ - \mathbb{E}\left[\hat{y}_c|\pi_f = 0\right]p(\pi_f = 0)\end{aligned} \tag{41}$$

where $\pi_f = 1$ if $\hat{y}_t - \hat{y}_c > 0$, and $\pi_f = 0$, otherwise.

D. Results

A toy example of CausalOT. CausalOT formulates the counterfactual inference as a regularized optimal transport problem that maps the factual space to the counterfactual space. Figure 3 visualizes the mapping process, which is produced by CausalOT on two synthetic datasets with n samples. We simulate n samples, and each sample has one-dimensional covariate x, treatment t and outcome y. For the convenience of visualization, Figure 3 does not include t. Given $\eta_i \sim \mathcal{N}(0,1)$, x_i is simulated by a normal distribution $\eta_i - 2$ for $0 < i < n/2$, and $\eta_i + 2$ otherwise. For $n = 100$, the treatment t_i determined by covariate x_i is binary, i.e., $t = 0$ for $x_i < 0$ otherwise $t = 1$. For $n = 200$, $t_i = 0$ if $\eta_i > 0$, otherwise $t_i = 1$. The factual outcome $y = \sigma * \mathcal{N}(0,1) + x/4 + 0.1 * t$ and the counterfactual outcome $y = \sigma * \mathcal{N}(0,1) - x/4 + 0.1 * (1-t)$. Figure 3 gives two synthetic regression models for two potential outcome models (i.e.,factual outcome and counterfactual outcome) each of which with 200 samples. The corresponding marginal distributions of covariates and outcomes for the sampling data is depicted in the middle figure. With CausalOT method, we can compute the transport maps between the factual samples and counterfactual samples for inferring unknown counterfactual outcomes. The green line in the right figure indicates the coupling $\boldsymbol{\gamma}$ between the factual outcome and the counterfactual outcome. Recall that $\boldsymbol{\gamma}_{i,j}$ indicates the amount of mass (the propensity measure) flowing from the factual sample i to the counterfactual sample j, we depict 30 green lines corresponding to top 30 values in $\boldsymbol{\gamma}$.

Illustration of CausalOT on `IHDP`. We illustrate the behavior of the proposed CausalOT method on two subsets

of IHDP with 50 and 100 samples, respectively. CausalOT aims to estimate the treatment effect from the observations, which requires to infer the unknown counterfactual outcomes. Before presenting the results of treatment effect estimation, we first illustrate our CausalOT on counterfactual inference. As shown in Figure 4, the left figure includes two colour points corresponding to the factual outcomes and the counterfactual outcomes. The grey line in the left figures indicates a coupling matrix γ between the factual outcomes and the counterfactual outcome. The lighter point is a larger value indicating that the corresponding element of coupling matrix is higher.

Treatment effect estimation. We further evaluate those methods on IHDP and News datasets by ϵ_{PEHE}, ϵ_{ATE} and ϵ_{ATT}. For dataset Twins and Jobs, we report the ϵ_{PEHE} and $\mathcal{R}_{pol}(\pi_f)$ both in-sample and out-of-sample as in Table IV. Apparently, our CausalOT method outperforms the baselines, in which the counterfactual loss accounts for the improvement in the accuracy of treatment effect estimation. BalanceNN obtains better ϵ_{PEHE} than CausalForest and BART on both datasets, as it considers the balanced property across the treated and the controlled units. On the dataset IHDP, CausalForest and BART outperform slightly BalanceNN on ϵ_{ATE} and ϵ_{ATT}, and this might due to the fact that BalanceNN has more parameters to be optimized but IHDP is a relatively small dataset. Both BalanceNN and DoubleRobust perform slightly worse than CausalForest. PscoreMatch fails to find good match for treated units as the dimension of covariates in News dataset is high, which results in its deteriorated performance when compared with that on IHDP dataset.

TABLE II: Comparison results on the IHDP dataset.

Method	IHDP		
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	ϵ_{ATT}
OLS-1	5.41±0.3	0.72±0.2	1.80±0.4
OLS-2	4.32±0.2	0.19±0.1	0.93±0.2
PscoreMatch	3.90±1.3	0.82±0.6	2.32±1.6
DoubleRobust	5.12±0.3	0.27±0.1	1.21±0.2
BART	2.65±0.4	0.24±0.3	0.49±0.6
CausalForest	4.14±0.2	0.22±0.8	0.85±1.0
BalanceNN	3.01±0.3	0.39±0.0	0.67±0.0
CausalOT	2.22±0.2	0.15±0.0	0.34±0.1

TABLE III: Comparison results on News dataset.

Method	News		
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	ϵ_{ATT}
OLS-1	4.59±0.1	0.96±0.3	1.53±0.2
OLS-2	4.20±0.2	0.87±0.0	0.74±0.1
PscoreMatch	4.62±1.0	1.3±0.7	0.89±0.7
DoubleRobust	4.38±0.1	0.92±0.6	1.12±0.3
BART	3.77±0.1	0.78±0.5	0.67±0.7
CausalForest	4.12±0.2	0.84±0.1	1.24±1.1
BalanceNN	3.65±0.2	0.65±0.0	0.72±0.3
CausalOT	3.40±0.1	0.57±0.0	0.55±0.1

To investigate how the performance of causal inference is affected by the size of overlapping between the treated and the controlled units, we vary the assignment bias for News dataset. A higher assignment bias indicates that less overlap between the treated and the controlled units. The treatment assignment of an News item x to a device $t \in \{0, 1\}$ is biased towards the preferred device for that item. As in [15], we assign the observed treatment by $t \sim \text{Bern}(\text{softmax}(\kappa y_j^F))$ with a coefficient $\kappa \geq 0$ determining the strength of the bias. We set κ in the range $(0, 10)$, where $\kappa = 0$ represents no assignment bias. We repeat the generative process 20 times for every κ.

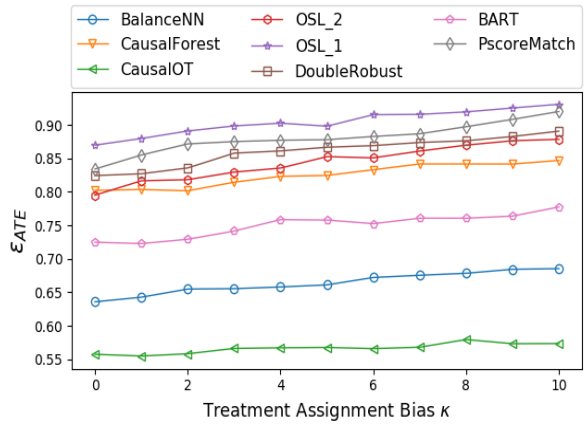
TABLE IV: Comparison results on Twins and Jobs dataset.

Method	Twins ϵ_{PEHE}		Jobs $\mathcal{R}_{pol}(\pi_f)$	
	In-sample	Out-sample	In-sample	Out-sample
OLS-1	0.32±0.01	0.33±0.00	0.24±0.00	0.25±0.00
OLS-2	0.31±0.02	0.34±0.02	0.22±0.00	0.23±0.01
PscoreMatch	0.33±0.03	0.34±0.02	0.30±0.01	0.31±0.02
DoubleRobust	0.35±0.01	0.36±0.02	0.29±0.02	0.32±0.02
BART	0.32±0.01	0.33±0.01	0.24±0.01	0.26±0.03
CausalForest	0.37±0.01	0.40±0.01	0.25±0.02	0.29±0.00
BalanceNN	0.30±0.01	0.31±0.00	0.21±0.01	0.24±0.01
CausalOT	0.29±0.00	0.30±0.01	0.20±0.01	0.21±0.03

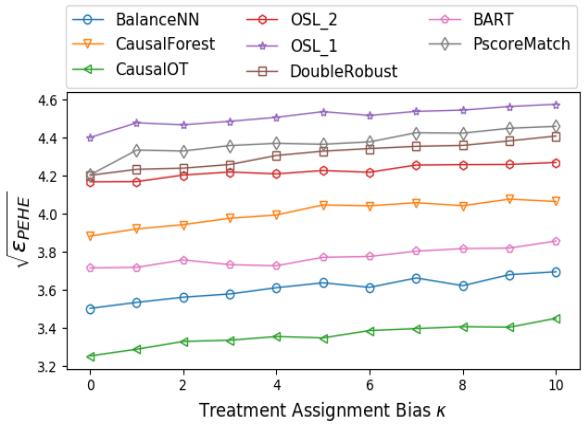
Apparently, CausalOT achieves the best results under different biases κ, as shown in Figures 5a and 5b. Namely, CausalOT is more robust to high assignment bias than existing state-of-the-art methods. This is because our method transports the propensity measures that statistically capture the global informative covariates among the factual and the counterfactual space to alleviate the issue of limited overlap. BalanceNN performs well and stably as the increasing of the treatment assignment bias, because it can learn a balanced representation for treated and controlled units. PscoreMatch, DoubleRobust and OLS-2 perform similarly on the balanced observed covariates for $\kappa = 0$. The performance of most baselines degrade dramatically as the treatment bias increases to the maximum value, which indicates that they are sensitive to the relative sample sizes of the treated and the controlled units.

Impact of trade-off parameter λ. As defined in objective function (11), the parameter $\lambda > 0$ balances the contributions of transport loss $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$ and counterfactual loss $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$. Figure 6 plots the performance of CausalOT under various λ on datasets IHDP and News. When $\lambda \to +\infty$, the objective function $\mathcal{L}$ is mainly dominated by $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$ and thus is less affected by $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$. Figure 6 indicates that relying mostly $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$ leads to a deteriorate accuracy of treatment effect estimation. Increasing the contribution of $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ (e.g., decreasing $\lambda = 10^0$ to $\lambda = 10^{-5}$) allows a performance gain, i.e., the errors of PEHE and τ_{ATE} are reduced. Namely, counterfactual loss $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ contributes more than transport loss $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$ to improving the performance on treatment effect estimation. Using λ being in the range of $(10^{-3}, 10^{-1})$ leads to the best performance on datasets IHDP and News.

Discussion of propensity score model. We defines a new propensity score measure that relies on the choice of propensity score model. The misspecification of propensity score indeed affects the performance of CausalOT. Following the pioneer work [5] in causal inference, we model the propensity score using logistic regression due to its simplicity



(a) $\sqrt{\epsilon_{PEHE}}$ on News under different κ.



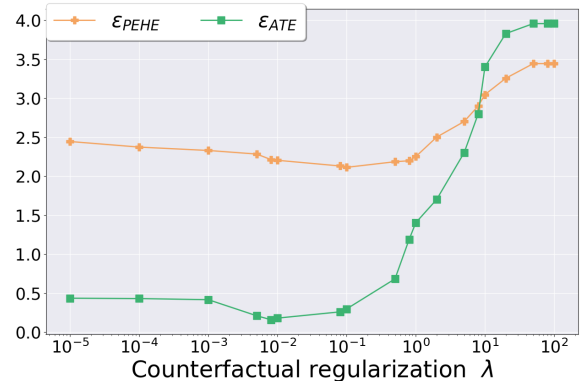
(b) ϵ_{ATE} on News under different κ.

and robustness. The parameters of the logistic regression can be tuned to achieve high accuracy. For instance, the trained logistic regression achieves accuracy of 0.901 for estimating the propensity scores on IHDP dataset, which results in $\sqrt{\epsilon_{\text{PEHE}}} = 2.22 \pm 0.2$ for treatment effect estimation. Inspired by [48], we train an advanced machine learning method (i.e., super-learner) to estimate the propensity score. We run the trained super-learner 100 times and achieve average accuracy of 0.931 on IHDP. Compared with logistic regression, the treatment effect estimation under the super-learner achieves very limited improvements with $\sqrt{\epsilon_{\text{PEHE}}} = 2.20 \pm 0.3$. Following this, we resort to the common choice of logistic regression for the estimation of propensity score.

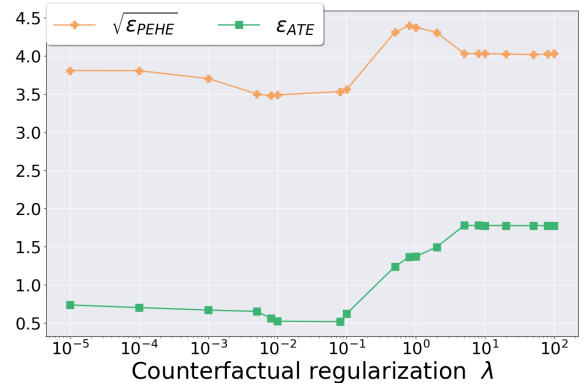
VIII. Conclusion and Future Work

In this paper, we propose an effective causal inference method CausalOT based on the optimal transport theory. First, we propose a novel propensity measure to capture the covariates and treatment assignment in factual and counterfactual space. With the propensity measure, CausalOT can reformulate the counterfactual outcome inference as a novel regularized optimal transport problem. Such reformulation is capable of utilizing the global information of observational covariates to alleviate the issue of limited overlapping. Second, to further guarantee the accuracy of counterfactual inference, we design a novel counterfactual loss for CausalOT to align the transported factual outcome during the transport of covariates. For the computation efficiency of CausalOT, we propose a proximal point algorithm based on Bregman divergence. Third, we prove that the counterfactual error of our CausalOT is bounded by the Wasserstein distance, which guarantees that CausalOT has a good generalization on the treatment effect estimation. Extensive empirical results confirm the superior performance of the proposed CausalOT method when compared with state-of-the-art methods.

CausalOT method uses raw covariates for the observational study, and our future research will incorporate available domain expertise or context knowledge to achieve higher interpretability.



(a) IHDP dataset



(b) News dataset

Fig. 6: $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE} on under different λ.

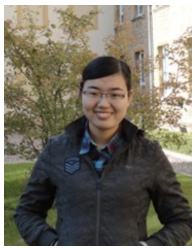
References

[1] M. M. Glymour and D. Spiegelman, "Evaluating public health interventions: 5. causal inference in public health research—do sex, race, and biological factors cause health outcomes?" *American journal of public health*, vol. 107, no. 1, pp. 81–85, 2017.

[2] N. Baum-Snow and F. Ferreira, "Causal inference in urban and regional economics," in *Handbook of regional and urban economics*. Elsevier, 2015, vol. 5, pp. 3–68.

[3] H. R. Varian, "Causal inference in economics and marketing," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7310–7315, 2016.

[4] J.-E. Gustafsson, "Causal inference in educational effectiveness research: A comparison of three methods to investigate effects of homework on student achievement," *School Effectiveness and School Improvement*, vol. 24, no. 3, pp. 275–295, 2013.
[5] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.
[6] A. Diamond and J. S. Sekhon, "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies," *Review of Economics and Statistics*, vol. 95, no. 3, pp. 932–945, 2013.
[7] M. Dudík, J. Langford, and L. Li, "Doubly robust policy evaluation and learning," *arXiv preprint arXiv:1103.4601*, 2011.
[8] M. J. Van Der Laan and D. Rubin, "Targeted maximum likelihood learning," *The international journal of biostatistics*, vol. 2, no. 1, 2006.
[9] S. Glazerman, D. M. Levy, and D. Myers, "Nonexperimental versus experimental estimates of earnings impacts," *The Annals of the American Academy of Political and Social Science*, vol. 589, no. 1, pp. 63–93, 2003.
[10] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate behavioral research*, vol. 46, no. 3, pp. 399–424, 2011.
[11] P. Z. Schochet, "Is regression adjustment supported by the neyman model for causal inference?" *Journal of Statistical Planning and Inference*, vol. 140, no. 1, pp. 246–259, 2010.
[12] H. A. Chipman, E. I. George, R. E. McCulloch *et al.*, "Bart: Bayesian additive regression trees," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 266–298, 2010.
[13] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.
[14] P. Schwab, L. Linhardt, and W. Karlen, "Perfect match: A simple method for learning representations for counterfactual inference with neural networks," *arXiv preprint arXiv:1810.00656*, 2018.
[15] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*, 2016, pp. 3020–3029.
[16] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3076–3085.
[17] N. Hassanpour and R. Greiner, "Counterfactual regression with importance sampling weights." in *IJCAI*, 2019, pp. 5880–5887.
[18] C. Kim and O. Bastani, "Learning interpretable models with causal guarantees," *arXiv preprint arXiv:1901.08576*, 2019.
[19] C. Rudin and D. Carlson, "The secrets of machine learning: Ten things you wish you had known earlier to be more effective at data analysis." *arXiv preprint arXiv:1906.01998*, 2019. [Online]. Available: https://academic.microsoft.com/paper/2948829945
[20] A. Lamont, M. D. Lyons, T. Jaki, E. Stuart, D. J. Feaster, K. Tharmaratnam, D. Oberski, H. Ishwaran, D. K. Wilson, and M. L. Van Horn, "Identification of predicted individual treatment effects in randomized clinical trials," *Statistical methods in medical research*, vol. 27, no. 1, pp. 142–157, 2018.
[21] S. Li and Y. Fu, "Matching on balanced nonlinear representations for treatment effects estimation," in *NIPS*, 2017.
[22] Z. Chu, S. L. Rathbun, and S. Li, "Matching in selective and balanced representation space for treatment effects estimation," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 205–214.
[23] R. Prentice, "Use of the logistic model in retrospective studies," *Biometrics*, pp. 599–606, 1976.
[24] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.
[25] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang, "Representation learning for treatment effect estimation from observational data," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
[26] J. Yoon, J. Jordon, and M. van der Schaar, "Ganite: Estimation of individualized treatment effects using generative adversarial nets," 2018.
[27] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
[28] D. B. Rubin, "Randomization analysis of experimental data: The fisher randomization test comment," *Journal of the American Statistical Association*, vol. 75, no. 371, pp. 591–593, 1980.
[29] J. Pearl, *Causality*. Cambridge university press, 2009.
[30] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.
[31] G. Peyré, M. Cuturi *et al.*, "Computational optimal transport," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.
[32] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, 1781.
[33] N. Courty, R. Flamary, and D. Tuia, "Domain adaptation with regularized optimal transport," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 274–289.
[34] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio, "Learning with a wasserstein loss," in *Advances in Neural Information Processing Systems*, 2015, pp. 2053–2061.
[35] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 9, pp. 1853–1865, 2016.
[36] S. Kim, R. Ma, D. Mesa, and T. P. Coleman, "Efficient bayesian inference methods via convex optimization and optimal transport," in *2013 IEEE International Symposium on Information Theory*. IEEE, 2013, pp. 2259–2263.
[37] D. S. Oliver, "Minimization for conditional simulation: Relationship to optimal transport," *Journal of Computational Physics*, vol. 265, pp. 1–15, 2014.
[38] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini, "An introduction to sampling via measure transport," *arXiv preprint arXiv:1602.05023*, 2016.
[39] B. Acciaio, J. Backhoff-Veraguas, and A. Zalashko, "Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization," *Stochastic Processes and their Applications*, vol. 130, no. 5, pp. 2918–2953, 2020.
[40] M. Schmidt, N. L. Roux, and F. R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Advances in neural information processing systems*, 2011, pp. 1458–1466.
[41] S. Si, D. Tao, and B. Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 7, pp. 929–942, 2009.
[42] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in neural information processing systems*, 2013, pp. 2292–2300.
[43] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, "Iterative bregman projections for regularized transportation problems," *SIAM Journal on Scientific Computing*, vol. 37, no. 2, pp. A1111–A1138, 2015.
[44] F. Bolley, A. Guillin, and C. Villani, "Quantitative concentration inequalities for empirical measures on non-compact spaces," *Probability Theory and Related Fields*, vol. 137, no. 3-4, pp. 541–593, 2007.
[45] A. S. Goldberger *et al.*, "Econometric theory." *Econometric theory.*, 1964.
[46] D. Almond, K. Y. Chay, and D. S. Lee, "The costs of low birth weight," *The Quarterly Journal of Economics*, vol. 120, no. 3, pp. 1031–1083, 2005.
[47] R. J. LaLonde, "Evaluating the econometric evaluations of training programs with experimental data," *The American economic review*, pp. 604–620, 1986.
[48] S. Alam, E. E. Moodie, and D. A. Stephens, "Should a propensity score model be super? the utility of ensemble procedures for causal adjustment," *Statistics in medicine*, vol. 38, no. 9, pp. 1690–1702, 2019.



Qian Li received her doctorate from Institute of Information Engineering, Chinese Academy of Science. She is currently a postdoctoral research fellow at University of Technology Sydney. She is interested in optimization algorithms for machine learning, topological data analysis and statistical causal analysis.

Zhichao Wang received his doctorate from Department of Automation, Tsinghua University. He is currently a postdoctoral research fellow at University of New South Wales. His research interests lie in the optimization for machine learning and stochastic modeling.

Shaowu Liu is a research fellow in the School of Computer Science, University of Technology Sydney. His current research interests include interpretable machine learning and causality learning. Besides research, he is also a data scientist with extensive experience in FinTech, renewable energy, and health.

Gang Li is IEEE senior member, associate professor in the school of Information Technology, Deakin University (Australia). His research interests include data mining, data privacy, causal discovery, and business intelligence.

Guandong Xu is a Full Professor and Program Leader at School of Computer Science and Advanced Analytics Institute, University of Technology Sydney, Australia and he received PhD degree in Computer Science from Victoria University, Australia. His research interests cover Data Science, Data Analytics, Recommender Systems, Web Mining, User Modelling, NLP, Social Network Analysis, and Social Media Mining. He has published three monographs in Springer and CRC press, and 250 + journal and conference papers including TOIS, TIST, TNNLS, TSC, TIFS, IEEE-IS, Inf. Sci., KAIS, WWWJ, KBS, Neurocomputing, ESWA, Inf. Retr., IJCAI, AAAI, WWW, ICDM, ICDE, CIKM. He is the assistant Editor-in-Chief of World Wide Web Journal, and served as guest editors for Pattern Recognition, IEEE Transactions on Computational Social Systems, Journal of Software and Systems, World Wide Web Journal, Multimedia Tools and Applications, and Online Information Review.

REVISION SUMMARY

Thank you for reviewing our submission to *IEEE Transactions on Neural Networks and Learning Systems*. We are grateful for your constructive and valuable comments, which help us further improve the overall quality of the paper. All those issues raised in the review reports have been addressed. **We highlight extensive changes in blue for the revised manuscript, the major changes made in this revision areas follows**:

- We reorganized the related work section by adding several important related work.
- We use the term *Counterfactual Loss* instead of *Counterfactual Regularization* to avoid ambiguity.
- We refined the deductions of some important formula for the sake of comprehensibility.
- We added two experiments to investigate the trade-off between the propensity measure transport and counterfactual loss on `IHDP` and `News` dataset.

We hope that this revision satisfies the standards of *IEEE Transactions on Neural Networks and Learning Systems*.

Response to Reviewer #1's Comments

Comments 1-1: This paper presents a causal optimal transport model (CausalOT) for treatment effect estimation, based on the optimal transport theory. In particular, the counterfactual outcomes are inferred by solving a regularized optimal transport problem. Also, a regularization term is designed to align the factual outcome distribution and the counterfactual outcome distribution. Experimental results on multiple benchmark datasets are reported and discussed. Pros.

- *Overall this paper is well organized and clearly written. The idea of using optimal transport theory for treatment effect estimation is well motivated.*
- *Theoretical analysis on optimal transport for counterfactual inference is provided.*
- *Experimental results show that the proposed method outperforms several representative baselines.*

Response 1-1: Thank your for your encouragement and acknowledge of the contribution of this work. In this work, we have further polished the work to highlight those aspects.

REVISION SUMMARY

Comments 1-2: The proposed CausalOT method is still dependent on the propensity scores. It is possible that the misspecification of propensity score estimation may negatively affect the performance of CausalOT. The authors may add some discussions regarding this issue.

Response 1-2: Thank you for your constructive comments. We add the following paragraph at the end of experimental part.

"The misspecification of propensity score indeed affects the performance of CausalOT. Following the pioneer work [5] in causal inference, we model the propensity score using logistic regression due to its simplicity and robustness. The parameters of the logistic regression can be tuned to achieve high accuracy. For instance, the trained logistic regression achieves accuracy of 0.901 for estimating the propensity scores on IHDP dataset, which results in $\sqrt{\epsilon_{\text{PEHE}}} = 2.22 \pm 0.2$ for treatment effect estimation. Inspired by [48], we train an advanced machine learning method (i.e., super-learner) to estimate the propensity score. We run the trained super-learner 100 times and achieve average accuracy of 0.931 on IHDP. Compared with logistic regression, the treatment effect estimation under the super-learner achieves very limited improvements with $\sqrt{\epsilon_{\text{PEHE}}} = 2.20 \pm 0.3$. Following this, we resort to the common choice of logistic regression for the estimation of propensity score."

Comments 1-3: The name, causal optimal transport, has been proposed in literature [a]. Although the technical approaches are different, the authors may clarify the differences to avoid confusions. [a] B. Acciaio, J. Backhoff-Veraguas, and A. Zalashko. "Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization". In: Stochastic Processes and their Applications (2019).

Response 1-3: Thank you for your concern. Actually the method proposed in this work shares the same name but it is essentially different from the work. In this revision, we explicitly clarify this issue in the Section III-B by adding the following paragraph.

"Note that Acciaio et al. [39] use a same name as our method, i.e., causal optimal transport. In fact, our method is totally different from their work with regarding to both the aim and technical methodology. Particularly, Acciaio et al. [39] define an optimal transport over causal couplings to addresses the stochastic analysis problem of filtrations enlargement. By contrast, our paper defines transportation plan in the propensity measure space to improve the causal effect estimation."

Comments 1-4: In the experiments, the authors may add ablation studies to evaluate the contribution of the regularization term.

Response 1-4: Thank you for your constructive suggestion. Instead of using the term *counterfactual regularization*, it is more appropriate to refer $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ as the *counterfactual loss*. Following your suggestion, we revised this term in the revised version to avoid misunderstandings. Namely, $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ can not be removed from the objective function, because $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ serves to estimate the counterfactual outcome predictor $h(\boldsymbol{v}))$, which is the main task of our method. Inspired by your insights, we add experiments to investigate how $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ affects the performance on treatment effect estimation. The discussion of the added experiment is given as follows:

"As defined in objective function (11), the parameter $\lambda > 0$ balances the contributions of transport loss $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$ and counterfactual loss $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$. fig:lambda plots the performance of CausalOT under various λ on datasets `IHDP` and `News`. When $\lambda \to +\infty$, the objective function $\mathcal{L}$ is mainly dominated by $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$ and thus is less affected by $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$. fig:lambda indicates that relying mostly $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$ leads to a deteriorate accuracy of treatment effect estimation. Increasing the contribution of $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ (e.g., decreasing $\lambda = 10^0$ to $\lambda = 10^{-5}$) allows a performance gain, i.e., the errors of PEHE and $\tau_{\mathbf{ATE}}$ are reduced. Namely, counterfactual loss $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ contributes more than transport loss $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$ to improving the performance on treatment effect estimation. Using λ being in the range of $(10^{-3}, 10^{-1})$ leads to the best performance on datasets `IHDP` and `News`."

Comments 1-5: Missing references on treatment effect estimation, such as [b-d]. [b] Matching in Selective and Balanced Representation Space for Treatment Effects Estimation. ACM CIKM, 2020. [c] Representation learning for treatment effect estimation from observational data, NeurIPS, 2018. [d] Matching on balanced nonlinear representations for treatment effects estimation, NIPS, 2017.

Response 1-5: Thank you for raising these issues, we re-organized related work and updated the latest references [b-d] accordingly.

Comments 1-6: Typo. Page 3: "counterfactual outcomes to for→ counterfactual outcomes for"

Response 1-6: Thank you for pointing out this typo, we deleted "to" and have further proofread the revised paper.

Response to Reviewer #2's Comments

Comments 2-1: The idea of formulating the treatment effects estimation in an optimal transport problem is interesting and novel. The manuscript is overall well organized and clearly describe the contributions of the work.

REVISION SUMMARY

Response 2-1: Thank your for your encouragement and acknowledge of the contributions of this work. In this work, we have further polished the work to highlight those aspects.

Comments 2-2: In regard to the probability of $h(\cdot)$, it was assumed to be uniform for simplicity. However, it is of importance for validity of the assumption in optimizing the loss in Eq. (11). Thus, it would be good to better justify the rationale of assuming the uniform distribution of $h(\cdot)$.

Response 2-1: Thank you for pointing out this issue. The probability of $h(\cdot)$ in fact refers to $p(h(\boldsymbol{v}) \mid \boldsymbol{v})$ in Eq. (9), i.e., the probability of counterfactual outcomes $h(\boldsymbol{v})$ given $\boldsymbol{v}$. Consider the fact that no prior knowledge is available about $h(\boldsymbol{v})$, we simply assume a uniform distribution in which all counterfactual outcomes equally occur given $\boldsymbol{v}$.

Comments 2-3: The authors need to check Eq. (21) carefully. It is not clear the rewriting from terms in the second '=' to those in the third '='. Especially, why the two ξ^2 terms are disappeared.

Response 2-2: Thank you for pointing out this issue. We corrected Eq. (21). There is one ξ^2 term after the second and third "=", respectively.

Comments 2-4: page 5, left-column, line 22: $p(h|x,t) \rightarrow p(h|x,1-t)$?

Response 2-4: Thank you for pointing out this issue. We revised $p(h|x,t)$ in line 22 as $p(h|x,1-t)$.

Comments 2-5: page 6, left-column, line 49: $l(v) \rightarrow l(h(v), f(v))$?

Response 2-5: Thank you for pointing out this problem. We revised the sentence as " where the loss function $l(h(\boldsymbol{v}), f(\boldsymbol{v})) = (h(\boldsymbol{v}) - f(\boldsymbol{v}))^2$ is denoted as $l(\boldsymbol{v})$ for short in the following text.".

Comments 2-6: It is not clear how the Eq. (26) was derived from Eq. (23) with reorganization?

Response 2-6: Thank you for pointing out this issue. We had a typo in Eq.(26), i.e., ϵ should be α that is a trade-off parameter. We added the following details in Eq. (25) and Eq. (26) to explain how to derive Eq. (26) from Eq. (23):

"Substituting Bregman divergence (24) into proximal point iteration

(23), with simplex constraints, we have

$$\boldsymbol{\gamma}^{(l+1)} = \arg\min_{\boldsymbol{\gamma}} \left\langle \nabla\mathcal{L}(\boldsymbol{\gamma}^{(l)}), \boldsymbol{\gamma} \right\rangle + \alpha \sum_{i,j}^{n} \boldsymbol{\gamma}_{i,j} \left(\log \boldsymbol{\gamma}_{i,j} - \log \boldsymbol{\gamma}_{i,j}^{(l)} \right) - \alpha \sum_{i,j}^{n} \boldsymbol{\gamma}_{i,j} + \alpha \sum_{i=1}^{n} \boldsymbol{\gamma}_{i,j}^{(l)} \quad (25)$$

We define $H(\boldsymbol{\gamma}) = \sum_{i,j}^{n} \boldsymbol{\gamma}_{i,j}(\log \boldsymbol{\gamma}_{i,j} - 1)$ and $\boldsymbol{\gamma}^{(l+1)}$ is reformulated (26) as

$$\begin{aligned} \boldsymbol{\gamma}^{(l+1)} &= \arg\min_{\boldsymbol{\gamma}} \left\langle \nabla\mathcal{L}(\boldsymbol{\gamma}^{(l)}) - \alpha \log \boldsymbol{\gamma}^{(l)}, \boldsymbol{\gamma} \right\rangle + \alpha H(\boldsymbol{\gamma}) + \alpha \sum_{i=1}^{n} \boldsymbol{\gamma}_{i,j}^{(l)} \\ &= \arg\min_{\boldsymbol{\gamma}} \left\langle \nabla\mathcal{L}(\boldsymbol{\gamma}^{(l)}) - \alpha \log \boldsymbol{\gamma}^{(l)}, \boldsymbol{\gamma} \right\rangle + \alpha H(\boldsymbol{\gamma}) \quad (26) \end{aligned}$$

where $\boldsymbol{\gamma}_{i,j}^{(l)}$ is a fixed value that is irrelevant to the optimization variable $\boldsymbol{\gamma}$. According to [42], $H(\boldsymbol{\gamma})$ is an entropy that allows Eq. (26) to have a closed-form solution and speed up the optimization."

Comments 2-7: What is the meaning of ϵ in Eq. (26) and Eq. (28)?

Response 2-7: Thank you for pointing out this problem. ϵ should be α that is the same value in Eq.(23).

Comments 2-8: As one of the main contributions in the work is the introduction of a regularization term. Thus, it is required to do an ablation study to see its effect, e.g., the comparison between with and without the regularization.

Response 2-8: Thank you for your constructive suggestion. Thank you for your constructive suggestion. Instead of using the term *counterfactual regularization*, it is more appropriate to refer $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ as the *counterfactual loss*. Following your suggestion, we revised this term in the revised version to avoid misunderstandings. Namely, $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ can not be removed from the objective function, because $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ serves to estimate the counterfactual outcome predictor $h(\boldsymbol{v}))$ which is the main task of our method.
Inspired by your insights, we add experiments to investigate how $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ affects the performance on treatment effect estimation. The discussion of the added experiment is given as follows:

"As defined in objective function (11), the parameter $\lambda > 0$ balances the contributions of transport loss $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$ and counterfactual loss $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$. fig:lambda plots the performance of CausalOT under various λ on datasets `IHDP` and `News`. When $\lambda \to +\infty$, the objective function $\mathcal{L}$ is mainly dominated by $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$ and thus is less affected by $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$. fig:lambda indicates that relying mostly $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$ leads to a deteriorate accuracy of treatment effect estimation. Increasing the contribution of $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ (e.g., decreasing $\lambda = 10^0$ to $\lambda = 10^{-5}$) allows a performance gain, i.e., the errors of PEHE and $\tau_{\mathbf{ATE}}$ are reduced. Namely, counterfactual loss $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ contributes more than transport loss $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$ to improving the

performance on treatment effect estimation. Using λ being in the range of $(10^{-3}, 10^{-1})$ leads to the best performance on datasets `IHDP` and `News`."

Comments 2-9: It is questionable not to have a weighting coefficient for the regularization term in Eq. (11).

Response 2-9: Thank you for pointing out this issue. As we discussed in response 2-8, $\mathcal{R}(\boldsymbol{y}, h(\boldsymbol{v}))$ is in fact a *counterfactual loss* rather than a *counterfactual regularization*. In Eq.(11), we added λ as the weighting coefficient for the transport loss $\mathcal{C}(\boldsymbol{u}, \boldsymbol{v})$.