SINKHORN DISCREPANCY FOR COUNTERFACTUAL GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating individual treatment effects from observational data is very challenging due to the existence of treatment selection bias. Most existing representation-based methods mitigate this issue by aligning distributions of different treatment groups in the representation space. However, they still suffer from two critical problems: (1) Mini-batch Sampling Effects (MSE), where the alignment easily fails due to the outcome imbalance or outliers in the batch; (2) Unobserved Confounder Effects (UCE), where the unobserved confounders damage the correct alignment. To tackle these problems, we propose a principled approach named **E**ntire **S**pace **C**ounter**F**actual **R**egression (ESCFR) based on a generalized sinkhorn discrepancy for distribution alignment within the stochastic optimal transport framework. Based on the framework, we propose a relaxed mass preserving regularizer to address the MSE issue and design a proximal factual outcome regularizer to handle the UCE issue. Extensive experiments demonstrate that our proposed ESCFR can successfully tackle the treatment selection bias and achieve significantly better performance than state-of-the-art methods.

1 INTRODUCTION

Estimating individual treatment effect (ITE) with randomized controlled trials (RCTs) is a common practice in causal inference, which has been widely used in e-commerce (Betlei et al., 2021), education (Cordero et al., 2018) and health care (Schwab et al., 2020). For example, application managers would impose a marketing strategy on the randomly sampled users to see its potential benefit on the click-through rate; drug developers would conduct clinical A/B tests to evaluate the drug effects. Although RCT is the golden standard for causal inference (Judea & Dana, 2018), it is always too expensive to conduct randomized experiments. Hence, observational data that can be acquired without intervention has been a tempting shortcut. For example, drug developers tend to assess drug effects with post-marketing monitoring reports instead of the clinical A/B trials. With the growing access to observational data, estimating ITE from observational data has attracted intense research interest.

Estimating ITE with observational data suffers from two primary issues: (1) missing counterfactuals, *i.e.*, we can only observe one factual outcome out of all potential outcomes; (2) treatment selection bias, *i.e.*, individuals have their preferences regarding treatment selection, making the population across different groups heterogeneous. To cope with missing counterfactuals, meta-learners (R et al., 2019) decompose the ITE estimation task into solvable subproblems. However, as shown in Section 2.1, the treatment selection bias makes it difficult to generalize the factual outcome estimators trained over the treated/untreated group to the entire population, and the ITE estimation is thus biased.

Representation-based methods mitigate this selection bias by minimizing the distribution discrepancy between groups in the representation space. In particular, Uri et al. (2017) proposed the CounterFactual Regression (CFR) framework, which optimizes the Precision in Estimation of Heterogeneous Effect (PEHE) by minimizing the estimation error of factual outcomes as well as the discrepancy of learned representations across groups. Liuyi et al. (2018) and Hassanpour & Greiner (2020) augmented CFR by considering local similarity and non-confounding factors, respectively.

However, two critical issues have long been circumvented, significantly hindering the representation-based methods from eliminating the treatment selection bias. The first problem is the Mini-batch Sampling Effects (MSE). Specifically, current representation-based methods (Uri et al., 2017; Liuyi

et al., 2018) compute distribution discrepancy within mini-batches instead of the entire data space, making it vulnerable to bad sampling cases. For example, given two aligned distributions, if an outlier exists in either distribution, the mini-batch discrepancy will be significant, making the training process noise-filled. The second problem is the Unobserved Confounder Effects (UCE). Specifically, current approaches directly assume *unconfoundedness*, while the unobserved confounders widely exist in real scenarios and make the resulting estimator biased.

Contributions and outline. In this paper, we propose an effective ITE estimator based on optimal transport, namely Entire Space Counterfactual Regression (ESCFR), which tackles both the MSE and UCE issues with a generalized sinkhorn discrepancy. Specifically, after necessary preliminaries in Section 2, we first reformulate the ITE estimation problem as a stochastic optimal transport problem in Section 3.1. We next showcase the MSE issue faced by existing approaches in Section 3.2 and propose a relaxed mass preserving regularizer to mitigate this issue. We further discuss the UCE issue in Section 3.3 and propose a proximal factual outcome regularizer to solve it. We finally formulate the architecture of ESCFR in Section 3.4, and report the experimental results in Section 4.

2 Preliminaries

2.1 Causal Inference from Observational Data

In this section, we formulate basic definitions and models in observational causal inference. We first define the critical elements following the general notation convention[1].

Definition 2.1. *Let X be the random variable of covariates, with support $\mathcal{X}$ and distribution $\mathbb{P}(x)$; Let R be the random variable of induced representations, with support $\mathcal{R}$ and distribution $\mathbb{P}(r)$; Let Y be the random variable of outcomes, with support $\mathcal{Y}$ and distribution $\mathbb{P}(y)$; Let T be the random variable of treatment indicator, with support $\mathcal{T} = \{0, 1\}$ and distribution $\mathbb{P}(T)$.*

Following the potential outcome framework (Rubin, 1974), an individual with covariates x has two potential outcomes, namely $Y_1(x)$ given it is treated and $Y_0(x)$ otherwise. The ground-truth individual treatment effect (ITE) is always formulated as the expected difference of potential outcome:

$$\tau(x) \coloneqq \mathbb{E}\left[Y_1 - Y_0 \mid x\right], \tag{1}$$

where one of these two outcomes is always being unobserved. To solve such missing counterfactuals, following meta-learners (R et al., 2019; Nie & Wager, 2020), the ITE task is commonly decomposed into several subproblems that are solvable with any supervised learning method.

For example, T-learner models the factual outcomes Y for units in the treated and untreated groups separately; S-learner models factual outcomes by regarding the treatment indicator T as one of the covariates X. The ITE estimate is the difference in predicted outcomes when T is set to treated and untreated. The advantages and limitations of both methods were discussed by R et al. (2019).

Definition 2.2. *Let $\psi : \mathcal{X} \to \mathcal{R}$ be a map from support $\mathcal{X}$ to $\mathcal{R}$. That is, $\forall x \in \mathcal{X}$, $\exists r = \psi(x) \in \mathcal{R}$. Let $\phi : \mathcal{R} \times \mathcal{T} \to \mathcal{Y}$ be a map from support $\mathcal{R} \times \mathcal{T}$ to $\mathcal{Y}$. That is, it maps the representations and treatment indicator to the corresponding factual outcome. For example, $Y_1 = \phi_1(R)$, $Y_0 = \phi_0(R)$, where we will always abbreviate $\phi(R, T = 1)$ and $\phi(R, T = 0)$ to $\phi_1(R)$ and $\phi_0(R)$, respectively, for brevity.*

TARNet (Uri et al., 2017) obtains better results by absorbing the advantages of both T-learner and S-learner, consisting of a representation map ψ and an outcome map ϕ as defined in Definition 2.2. As shown in Figure 3, for an individual with covariates X, TARNet estimates ITE as:

$$\hat{\tau}_{\psi,\phi}(X) = \hat{Y}_1 - \hat{Y}_0, \quad \text{where} \quad \hat{Y}_1 = \phi_1(\psi(X)), \quad \hat{Y}_0 = \phi_0(\psi(X)), \tag{2}$$

where ψ is trained over all individuals, ϕ_1 and ϕ_0 are trained over the treated and untreated group, respectively, to minimize the factual error. Finally, the performance of ITE estimators is mainly evaluated with the expected Precision in Estimation of Heterogeneous Effect (PEHE):

$$\epsilon_{\text{PEHE}}(\psi, \phi) = \int_{\mathcal{X}} \left(\hat{\tau}_{\psi,\phi}(x) - \tau(x)\right)^2 \hat{\mathbb{P}}(x)\, dx. \tag{3}$$

[1]We use uppercase letter, *e.g.*, X to denote a random variable and lowercase letter, *e.g.*, x to denote an associated specific value. Letters in calligraphic font, *e.g.*, $\mathcal{X}$ represent the support of the corresponding random variable, and $\mathbb{P}()$ represents the probability distribution of the random variable, *e.g.*, $\mathbb{P}(X)$.

However, as shown in Figure 2 (a), the treatment selection bias shifts covariate distributions across groups. As such, ϕ_1 and ϕ_0 would overfit the respective group's properties and thus cannot generalize well to the entire population. Therefore, the resulting $\hat{\tau}$ would be biased.

2.2 Discrete Optimal Transport and Sinkhorn Divergence

Optimal transport defines a powerful geometry to measure the distribution discrepancy. Monge first formulated optimal transport as a problem of finding an optimal mapping between two measures. However, this formulation cannot guarantee the existence and uniqueness of solutions. More applicable is Kantorovich's formulation, which can be seen as a generalized Monge problem.

Definition 2.3. *For empirical distributions α and β with n and m units, respectively, the Kantorovich problem aims to find a feasible plan $\pi \in \mathbb{R}_+^{n\times m}$ which transports α to β at minimum cost:*

$$\mathbb{W}(\alpha,\beta) := \min_{\boldsymbol{\pi}\in\Pi(\alpha,\beta)} \langle \mathbf{D}, \boldsymbol{\pi}\rangle, \quad \Pi(\alpha,\beta) := \left\{\boldsymbol{\pi} \in \mathbb{R}_+^{n\times m} : \boldsymbol{\pi}\mathbf{1}_m = \mathbf{a}, \boldsymbol{\pi}^{\mathrm{T}}\mathbf{1}_n = \mathbf{b}\right\}, \tag{4}$$

where $\mathbb{W}(\alpha,\beta) \in \mathbb{R}$ is the Wasserstein discrepancy between α and β; $\mathbf{D} \in \mathbb{R}_+^{n\times m}$ is the unit-wise distance[2] between α and β; $\mathbf{a}$ and $\mathbf{b}$ indicate the mass of units in α and β, and Π is the feasible transportation plan set which ensures the mass preserving constraint holds.

The exact solutions (Bonneel et al., 2011) to this problem always come with high computational costs. As such, researchers would always add an entropic regularization to the Kantorovich problem:

$$\mathbb{W}^{\epsilon}(\alpha,\beta) := \langle \mathbf{D}, \boldsymbol{\pi}^{\epsilon}\rangle, \quad \boldsymbol{\pi}^{\epsilon} := \operatorname*{arg\,min}_{\boldsymbol{\pi}\in\Pi(\alpha,\beta)} \langle \mathbf{D}, \boldsymbol{\pi}\rangle - \epsilon \mathrm{H}(\boldsymbol{\pi}), \quad \mathrm{H}(\boldsymbol{\pi}) := -\sum_{i,j} \boldsymbol{\pi}_{i,j}\left(\log(\boldsymbol{\pi}_{i,j}) - 1\right), \tag{5}$$

making the problem ϵ-convex and solvable with the Sinkhorn algorithm (Cuturi, 2013). The Sinkhorn algorithm only consists of matrix-vector products, making it suited to be accelerated with GPUs.

3 Proposed Method

This section proposes the ESCFR approach to tackle the treatment selection bias. It is built upon the stochastic optimal transport framework for distribution discrepancy minimization across treatment groups. Based on the framework, we propose a relaxed mass preserving regularizer to address the sampling effect, and a proximal factual outcome regularizer to handle the unobserved confounders. We finally summarize the model architecture, learning objective, and optimization procedure.

3.1 Stochastic Optimal Transport for Counterfactual Regression

Representation-based approaches mitigate the treatment selection bias by calculating the distribution discrepancy and then minimizing it. We select optimal transport to calculate the discrepancy among current discrepancy measures, as it has shown attractive advantages over its competitors. Specifically, it takes into account the distribution's geometry and thus works in the cases where ϕ-divergence (*e.g.*, Kullback-Leibler divergence) fails (Seguy et al., 2018). In addition, it does not require adversarial training and is, therefore, easier to optimize than adversarial-based measures (Kallus, 2020).

Optimal transport calculates the group discrepancy as $\mathbb{W}\left(\mathbb{P}_{\psi}^{\mathrm{T}=1}(r), \mathbb{P}_{\psi}^{\mathrm{T}=0}(r)\right)$, where $\mathbb{P}_{\psi}^{\mathrm{T}=1}(r)$ and $\mathbb{P}_{\psi}^{\mathrm{T}=0}(r)$ are the representation distributions under treated and untreated groups, respectively, induced by the map $r = \psi(x)$. The discrepancy could be minimized by updating the representation map ψ with gradient-based optimizers, as it is differentiable with respect to ψ (Flamary et al., 2021).

Definition 3.1. *Let $\hat{\mathbb{P}}^{T=1}(x)$ and $\hat{\mathbb{P}}^{T=0}(x)$ be the empirical distributions of covariates at a mini-batch level, which contain n treated units and m untreated units, respectively; $\hat{\mathbb{P}}_{\psi}^{T=1}(r)$ and $\hat{\mathbb{P}}_{\psi}^{T=0}(r)$ be that of representations induced by the representation map $r = \psi(x)$ defined in Definition 2.2.*

However, as neural network estimators mainly update parameters with stochastic gradient methods, only a subset of the representation's distribution is accessible within each iteration. As such, a shortcut (Liuyi et al., 2018) is to calculate the group discrepancy at a stochastic mini-batch level:

$$\hat{\mathbb{W}}_{\psi} := \mathbb{W}\left(\hat{\mathbb{P}}_{\psi}^{\mathrm{T}=1}(r), \hat{\mathbb{P}}_{\psi}^{\mathrm{T}=0}(r)\right). \tag{6}$$

[2]We calculate the unit-wise distance with the squared Euclidean metric following Courty et al. (2017b).

We further derive the upper bound of PEHE in the stochastic batch form as in Theorem 3.1 based on Uri et al. (2017), which demonstrates that the PEHE can be optimized by iteratively minimizing the factual outcome estimation error and the optimal transport discrepancy *at a mini-batch level*.

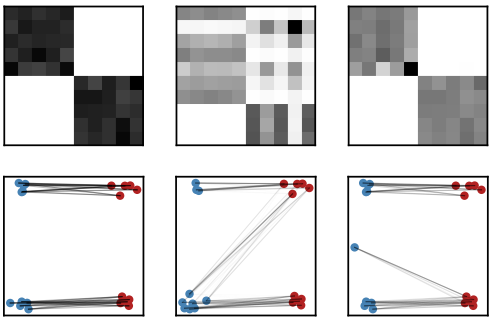
Theorem 3.1. *Let ψ and ϕ be the representation map and factual outcome map, respectively; $\hat{\mathbb{W}}_\psi$ be the group discrepancy at a mini-batch level. With the probability of at least $1-\delta$, we have:*

$$\epsilon_{\text{PEHE}}(\psi,\phi) \leq 2\left[\epsilon_{\text{F}}^{T=1}(\psi,\phi) + \epsilon_{\text{F}}^{T=0}(\psi,\phi) + B_\psi \hat{\mathbb{W}}_\psi - 2\sigma_Y^2 + \mathcal{O}(\frac{1}{\delta N})\right], \tag{7}$$

where $\epsilon_{\text{F}}^{T=1}$ and $\epsilon_{\text{F}}^{T=0}$ are the expected losses of factual outcome estimation over treated and untreated units, respectively. N is the batch size, σ_Y^2 is the variance of outcomes, B_ψ is some constant, and $\mathcal{O}(\cdot)$ is the sampling complexity term. (More details are provided in the appendix.)

3.2 Relaxed Mass Preserving Regularizer for Sampling Effect

Current representation-based methods fail to eliminate the treatment selection bias for circumventing mini-batch sampling effects. Although Theorem 3.1 guarantees that the stochastic optimal transport discrepancy bounds the PEHE, the sampling complexity term $\mathcal{O}(\cdot)$ inspires us to consider the potential risk of bad cases caused by stochastic sampling. Figure 1 (a) showcases the ideal sampled mini-batch where the transport strategy is reasonable and applicable; Figure 1 (b) showcases the mini-batch with outcome imbalance, where units with unrelated factual outcomes are incorrectly matched; Figure 1 (c) showcases the mini-batch with outliers, where the outlier significantly disturbs the transport plan of other units. Evidently, the traditional optimal transport performed at a minibatch level makes ITE estimation vulnerable to these bad sampling cases, which we summarize as the Mini-batch Sampling Effect (MSE) issue.



(a) Ideal (b) Imbalance (c) Outlier

Figure 1: Optimal transport plan (upper) and its geometric interpretation (down) in three cases, where the connection strength depicts the transported mass. Different colors (vertical positions) indicate different treatments (outcomes).

This issue stems mainly from the mass preservation constraint. Specifically, units acting as outliers at a mini-batch level would be forced to be transported by this marginal constraint, as depicted in Figure 1. It hinders the transport of normal units and makes the group discrepancy vulnerable to the sampling effect. Such issue would be exacerbated by the small batch size.

Definition 3.2. *For empirical distributions α and β with n and m units, respectively, optimal transport with relaxed mass preserving constraint seeks for a transport plan at minimum cost:*

$$\mathbb{W}^{\epsilon,\kappa}(\alpha,\beta) := \min_{\boldsymbol{\pi}} \langle \mathbf{D}, \boldsymbol{\pi} \rangle, \boldsymbol{\pi} := \arg\min_{\boldsymbol{\pi}} \langle \mathbf{D}, \boldsymbol{\pi} \rangle - \epsilon \mathrm{H}(\boldsymbol{\pi}) + \kappa(\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\pi}\mathbf{1}_m, \mathbf{a}) + \mathrm{D}_{\mathrm{KL}}(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{1}_n, \mathbf{b})) \tag{8}$$

where $\mathbf{D} \in \mathbb{R}_+^{n\times m}$ is the unit-wise distance, and $\mathbf{a}$ and $\mathbf{b}$ indicate the mass of units in α and β.

An intuitive approach to mitigate MSE is to relax the marginal constraint and allow for the creation and destruction of mass. To achieve this, we propose to leverage a relaxed mass preserving regularizer (RMPR) as inspired by Chizat et al. (2018), which replaces the hard marginal constraint with a soft penalty. As detailed in Definition 3.2, we calculate the stochastic discrepancy as:

$$\hat{\mathbb{W}}_\psi^{\epsilon,\kappa} := \mathbb{W}^{\epsilon,\kappa}\left(\hat{\mathbb{P}}_\psi^{\mathrm{T}=1}(r), \hat{\mathbb{P}}_\psi^{\mathrm{T}=0}(r)\right) \tag{9}$$

where the RMPR replaces the hard mass preservation constraint and thus mitigates the MSE issue. Inspired by Fatras et al. (2021), the robustness of RMPR to sampling effects is further formalized in Theorem 3.2, where the effect of sampled outliers is upper bounded by a constant.

Theorem 3.2. *For empirical distributions α, β with n and m units, respectively, adding an outlier a' to α and denote the disturbed distribution as α', we have*

$$\mathbb{W}^{0,\kappa}(\alpha',\beta) - \mathbb{W}^{0,\kappa}(\alpha,\beta) \leq 2\kappa(1 - e^{-\sum_{b\in\beta}(a'-b)^2/2\kappa})/n, \tag{10}$$

which is upper bounded by $2\kappa/n$. $\mathbb{W}^{0,\kappa}$ is the unbalanced discrepancy as per Definition 3.2.

In addition, compared with alternatives (Xu et al., 2020; Chapel et al., 2021) to relax the marginal constraint, the approach in Definition 3.2 has good metric properties (Séjourné et al., 2019) and can be accelerated via the generalized Sinkhorn algorithm (Chizat et al., 2018). It is differentiable *w.r.t.* ψ and thus can be minimized via stochastic gradient methods in an end-to-end manner.

3.3 Proximal Factual Outcome Regularizer for Unobserved Confounders

Existing representation-based methods fail to eliminate the treatment selection bias due to the Unobserved Confounder Effects (UCE). Specifically, as shown in Figure 2 (b), existing methods (Yao et al., 2019) just assume unconfoundedness in Assumption A.1 to circumvent the UCE issue. Specifically, given two units $r_i \in \mathbb{P}_{\psi}^{T=1}(r)$ and $r_j \in \mathbb{P}_{\psi}^{T=0}(r)$, the optimal transport in Definition 3.2 directly calculates the unit-wise distance as $\mathbf{D}_{ij} := (r_i - r_j)^2$. Given Assumption A.1 holds, this approach mitigates the treatment selection bias since it blocks the backdoor path $X \to T$ by balancing the distribution of the observed covariates in a latent space. However, Assumption A.1 is usually violated in real scenarios, which invalidates existing methods as the backdoor path $X' \to T$ is not blocked.

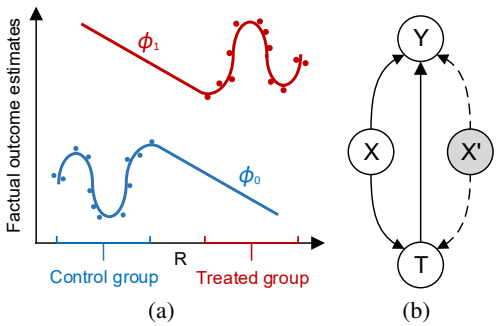


Figure 2: Visualization of (a) treatment selection bias; (b) unobserved confounder effect, where shaded X' is the unobserved confounder.

According to Figure 2 (b), given balanced X and identical T, the only variable reflecting the variation of X' is the outcome Y. As such, inspired by the joint distribution transport technology (Courty et al., 2017a), we propose to calibrate the unit-wise distance $\mathbf{D}$ with the potential outcomes as follow:

$$\mathbf{D}_{ij}^{\gamma} = \|r_i - r_j\|^2 + \gamma \cdot \left[\|y_i^{T=0} - y_j^{T=0}\|^2 + \|y_j^{T=1} - y_i^{T=1}\|^2\right], \tag{11}$$

where γ controls the strength of regularization. The underlying regularization is that units with similar observed and unobserved confounders ought to exhibit similar potential outcomes. As such, for a pair of units with similar observed covariates, *i.e.*, $\|r_i - r_j\|^2 \approx 0$, if the potential outcomes under the same treatment $t = \{0, 1\}$ differ significantly, *i.e.*, $\|y_i^t - y_j^t\| >> 0$, the unobserved confounders of the unit pair should also differ significantly. The optimal transport based on $\mathbf{D}_{ij} := (r_i - r_j)^2$ would incorrectly match this pair because $\|r_i - r_j\|^2 \approx 0$, and thus bias the updating of the representation map ψ. In contrast, the optimal transport with $\mathbf{D}_{ij}^{\gamma}$ would not match this pair as the difference of unobserved confounders is compensated with that of potential outcomes.

However, only factual outcomes are available owing to the missing counterfactuals, *i.e.*, counterfactual outcomes $y_i^{T=0}$ and $y_j^{T=1}$ are unavailable in the training data. As such, the proposed Proximal Factual Outcome Regularizer (PFOR) uses the estimates of these counterfactual outcomes instead:

$$\mathbf{D}_{ij}^{\gamma} = \|r_i - r_j\|^2 + \gamma \cdot \left[\|\hat{y}_i - y_j\|^2 + \|\hat{y}_j - y_i\|^2\right], \quad \hat{y}_i = \phi_0(r_i), \quad \hat{y}_j = \phi_1(r_j), \tag{12}$$

where $y_j := y_j^{T=0}$ and $y_i := y_i^{T=1}$. See more rigorous analysis in Appendix D.3.

3.4 Architecture of Entire Space Counterfactual Regression

The architecture of ESCFR is illustrated in Figure 3, where the covariate X is mapped to the representations R with $\psi(\cdot)$, then to the potential outcomes with $\phi(\cdot)$. The group discrepancy $\mathbb{W}$ is calculated in the representation space with the optimal transport equipping with RMPR and PFOR regularizers.

The learning objective seeks to minimize the risk of factual outcome estimation and the group discrepancy. Given mini-batch distributions $\hat{\mathbb{P}}^{T=1}(x)$ and $\hat{\mathbb{P}}^{T=0}(x)$ in Definition 3.1, the risk of factual

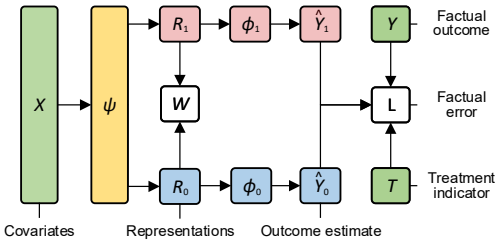


Figure 3: Architecture of ESCFR

outcome estimation following Shi et al. (2019) is:

$$\mathcal{L}_{\mathrm{F}}(\psi, \phi) := \mathbb{E}_{x_i \in \hat{\mathbb{P}}^{T=1}(x)} \left(\phi_1(\psi(x_i)) - y_i\right)^2 + \mathbb{E}_{x_j \in \hat{\mathbb{P}}^{T=0}(x)} \left(\phi_0(\psi(x_j)) - y_j\right)^2, \tag{13}$$

where y_i and y_j are the corresponding factual outcomes. Afterwards, the group discrepancy is:

$$\mathcal{L}_{\mathrm{D}}^{\epsilon,\kappa,\gamma}(\psi) := \mathbb{W}^{\epsilon,\kappa}\left(\hat{\mathbb{P}}_{\psi}^{\mathrm{T}=1}(r), \hat{\mathbb{P}}_{\psi}^{\mathrm{T}=0}(r)\right), \tag{14}$$

which in general follows the RMPR regularizer in Definition 3.2, except for the unit-wise distance $\mathbf{D}^{\gamma}$ calculated with the PFOR in Equation 12. Finally, we derive the overall learning objective:

$$\mathcal{L}_{\mathrm{ESCFR}}^{\epsilon,\kappa,\gamma,\lambda} := \mathcal{L}_{\mathrm{F}}(\psi, \phi) + \lambda \cdot \mathcal{L}_{\mathrm{D}}^{\epsilon,\kappa,\gamma}(\psi) \tag{15}$$

where λ controls the strength of distribution alignment, ϵ controls the entropic regularization in Equation 5, κ controls RMPR in Equation 8, and γ controls PFOR in Equation 12. These objectives coorperate to mitigate the selection bias following Theorem 3.1 as well as the MSE and UCE issues.

The optimization procedure consists of three steps. (1) compute $\boldsymbol{\pi}^{\epsilon,\kappa}$ by solving the linear programming problem in Definition 3.2 with Algorithm 2, where the unit-wise distance is calculated with $\mathbf{D}^{\gamma}$. (2) calculate the discrepancy in Equation 14 as $\langle \boldsymbol{\pi}^{\epsilon,\kappa}, \mathbf{D}^{\gamma} \rangle$, where $\mathbf{D}^{\gamma}$ is differentiable to ψ, making it feasible to minimize this discrepancy with its gradient w.r.t. ψ. (3) calculate the overall loss in Equation 15 and update ψ and ϕ with stochastic gradient methods. See Algorithm 3 for details.

4 Experiments

4.1 Experimental Setup

Datasets. Missing counterfactual makes it infeasible to evaluate the ground-truth PEHE over observational benchmarks. Following Liuyi et al. (2018); Uri et al. (2017), experiments are conducted on two semi-synthetic benchmarks. Specifically, the IHDP benchmark aims to estimate the effect of specialist home visits on infant's future cognitive scores, with 747 observations and 25 covariates; the ACIC dataset comes from the collaborative perinatal project (Niswander & Gordon, 1972), with 4802 observations and 58 covariates.

Baselines. We consider three groups of baselines. (1) Statistical estimators: least square regression with the treatment as covariates (OLS), random forest with the treatment as covariates (R.Forest), a single network with the treatment as covariates (S.learner by R et al. (2019)), separate neural regressors for each treatment group (T.learner by R et al. (2019)), TARNet (Uri et al., 2017); (2) Matching estimators: propensity score match with logistic regression (PSM by Rosenbaum & Rubin (1983b)), k-nearest neighbor (k-NN by Crump et al. (2008)), causal forest (C.Forest by Wager & Athey (2018)), orthogonal forest (O.Forest by Wager & Athey (2018)); (3) Representation-based methods: balancing neural network (BNN by Johansson et al. (2016)), counterfactual regression with MMD (CFR-MMD) and Wasserstein metric (CFR-WASS) by Uri et al. (2017).

Training protocol. A fully connected neural network with two hidden layers as 60-60 has been selected to realize the representation map ψ and the factual outcome map ϕ for ESCFR and other neural network based baselines. For fair comparison, all neural models are trained for 400 epochs with Adam Kingma & Ba (2015) optimizer, where both learning rate and weight decay are set to $1 \times e^{-3}$. Other settings of optimizers follow Kingma & Ba (2015). We search hyperparameters within the range in Figure 5, checkpoint the validation performance every 2 epochs, and export the best model to evaluate its performance on the test dataset.

Evaluation protocol. Following Liuyi et al. (2018), PEHE in Equation 3 is primarily used as the precision metric for performance evaluation. However, PEHE is not computable during the model selection stage as the counterfactual outcome is unavailable. As such, we use the Area Under the Uplift Curve (AUUC) (Betlei et al., 2021) to guide model selection, which reports the ranking performance and can be calculated with factual outcomes. We report the within-sample results on the training dataset and the out-of-sample results on the test dataset, where the factual outcome is available in the within-sample case following Uri et al. (2017).

Table 1: Performance (mean±std) on the PEHE and AUUC metrics. "*" marks the baseline estimators that ESCFR outperforms significantly at p-value < 0.05 over paired samples t-test.

Dataset	ACIC (PEHE)		IHDP (PEHE)		ACIC (AUUC)		IHDP (AUUC)	
Model	In-sample	Out-sample	In-sample	Out-sample	In-sample	Out-sample	In-sample	Out-sample
OLS	3.749±0.080*	4.340±0.117*	3.856±6.018	5.674±9.026	0.843±0.007	0.496±0.017*	0.652±0.050	0.492±0.032*
R.Forest	3.597±0.064*	3.399±0.165*	2.635±3.598	4.671±9.291	0.902±0.016	0.702±0.026*	0.736±0.142	0.661±0.259
S.Learner	3.572±0.269*	3.636±0.254*	1.706±1.600*	3.038±5.319	**0.905±0.041**	0.627±0.014*	0.633±0.183	0.702±0.330
T.Learner	3.429±0.142*	3.566±0.248*	1.567±1.136*	2.730±3.627	0.846±0.019	0.632±0.020*	0.651±0.179	0.707±0.333
TARNet	3.236±0.266*	3.254±0.150*	0.749±0.291	1.788±2.812	0.886±0.046	0.662±0.014*	0.654±0.184	0.711±0.329
C.Forest	3.449±0.101*	3.196±0.177*	4.018±5.602*	4.486±8.677	0.717±0.005*	0.709±0.018*	0.643±0.141	0.695±0.294
k-NN	5.605±0.168*	5.892±0.138*	2.208±2.233*	4.319±7.336	0.892±0.007*	0.507±0.034*	0.725±0.142	0.668±0.299
O.Forest	8.094±4.669*	4.148±2.224*	2.605±2.418*	3.136±5.642	0.744±0.013	0.699±0.022*	0.664±0.157	0.702±0.325
PSM	5.228±0.154*	5.094±0.301*	3.219±4.352*	4.634±8.574	0.884±0.010	0.745±0.021	**0.740±0.149**	0.681±0.253
BNN	3.345±0.233*	3.368±0.176*	0.709±0.330	1.806±2.837	0.882±0.033	0.645±0.013*	0.654±0.184	0.711±0.329
CFR-MMD	3.182±0.174*	3.357±0.321*	0.777±0.327	1.791±2.741	0.871±0.032	0.659±0.017*	0.655±0.183	0.710±0.329
CFR-WASS	3.128±0.263*	3.207±0.169*	0.657±0.673	1.704±3.115	0.873±0.029	0.669±0.018*	0.656±0.187	0.715±0.329
ESCFR	**2.252±0.297**	**2.316±0.613**	**0.607±0.328**	**1.257±0.677**	0.796±0.030	**0.754±0.021**	0.659±0.187	**0.734±0.329**

Table 2: Ablation study (mean±std) on the ACIC benchmark. "*" marks the variants that ESCFR outperforms significantly at p-value < 0.01 over paired samples t-test.

			In-sample		Out-sample	
SOT	RMPR	PFOR	PEHE	AUUC	PEHE	AUUC
✗	✗	✗	3.2367±0.2666*	**0.8862+0.0462**	3.2542+0.1505*	0.6624+0.0149*
✓	✗	✗	3.1284±0.2638*	0.8734+0.0291	3.2073+0.1699*	0.6698+0.0187*
✓	✓	✗	2.6459+0.2747*	0.8356+0.0286	2.7688±0.4009	0.7099+0.0157*
✓	✗	✓	2.5705±0.3403*	0.8270±0.0341	2.6330±0.4672	0.7110±0.0287*
✓	✓	✓	**2.2520±0.2975**	0.7968±0.0307	**2.3165+0.6136**	**0.7542±0.0202**

4.2 Performance Comparison

Table 1 reports the PEHE of ESCFR and its competitors over ten runs. We also report the AUUC score in Table 1 since it acts as the model selection criteria and reflects the ranking performance.

Statistical estimators exhibit competitive performance on the PEHE metric. In particular, neural network estimators outperform the linear and random forest methods because they can depict the nonlinearity in data. TARNet obtains better overall performance than other statistic estimators by absorbing the advantages (R et al., 2019) of both T-learner and S-learner. However, the treatment selection bias makes these estimators biased, leading to sub-optimal performance.

Matching methods, *e.g.*, PSM exhibit promising performance on ranking metric, which explains why they are favored by counterfactual ranking applications (Betlei et al., 2021) in practice. However, their poor performance on PEHE hinders their application in counterfactual estimation applications such as advertising systems where accuracy metrics are more critical.

Representation-based methods mitigate the treatment selection bias and enhance the overall performance. In particular, CFR-WASS reaches an out-of-sample PEHE of 3.207 on ACIC, significantly outperforming most statistical methods. It also gets an AUUC of 0.715 on IHDP, exceeding all other baselines. However, MSE and UCE issues hinder the mitigation of treatment selection bias by these methods. The proposed ESCFR achieves significant improvement over most metrics compared with various state-of-the-art baselines [3]. Combined with aforementioned comparisons, we attribute its superiority to our design of the RMPR and PFOR regularizers, making it robust to MSE and UCE. See Appendix D.5 for additional comparison results.

4.3 Ablation Study

To avoid the significant variance in the IHDP benchmark, we conduct ablation study on the ACIC benchmark, to evaluate the effectiveness of ESCFR's components and validate our claims in Section 3.

[3]An exception would be the within-sample AUUC, which is reported over training data and thus easy to be overfitted. This metric is not critical as the factual outcomes are generally unavailable in the inference phase. We mainly focus on out-of-sample AUUC instead, to assess the ranking performance and conduct model selection.

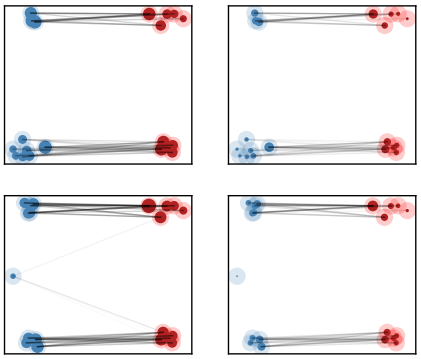
In Table 2, ESCFR firstly augments TARNet with stochastic optimal transport in Section 3.1, which effectively reduces the out-of-sample PEHE from 3.254 to 3.207. Afterwards, it mitigates the MSE issue with RMPR in Section 3.2 and the UCE issue with PFOR in Section 3.3, reducing the out-of-sample PEHE to 2.768 and 2.633, respectively. Finally, ESCFR combines the RMPR and PFOR in a unified framework in Section 3.4 and further reduces the out-of-sample PEHE to 2.316.

4.4 Analysis of Relaxed Mass Preserving Regularizer

Existing models fail to eliminate the effect of outliers at a mini-batch level due to the MSE issue, as indicated by the two bad cases in Figure 1 (b-c). Following the same settings, Figure 4 showcases the transport plan with RMPR, where RMPR alleviates MSE in both bad cases. Specifically, RMPR with $\kappa = 10$ avoids the incorrect matching of units with different outcomes; RMPR with $\kappa = 2$ gets robust to the outlier's interference and correctly matches the remaining units. The robustness to MSE comes from the relaxed mass preserving constraint, validating Section 3.2.

Additionally, RMPR does not transport all mass of a unit. The closer the unit is to the overlap zone in a batch, the greater the mass being transferred. That is, RMPR adaptively matches and pulls closer units that are close to the overlap region, ignoring outliers, which mitigates the bias of causal inference methods in cases where the positivity assumption does not strictly hold. Current approaches mainly achieve it by manually cleaning the data or dynamically weighting the units (Johansson et al., 2020), while RMPR naturally implements it via the soft penalty in Equation 8.

We further quantify the performance improvement of RMPR with different settings of batch size and κ in Appendix D.2, empirically backing up Section 3.2.



(a) $\kappa = 10$ (b) $\kappa = 2$

Figure 4: Geometric interpretation of optimal transport plan with RMPR under the outcome imbalance (upper) and outlier (down) settings. The dark area indicates the transported mass of a unit, *i.e.*, marginal of the transport matrix π. The light area indicates the total mass.

4.5 Parameter Sensitivity Study

We discuss four critical hyperparameters in ESCFR, *i.e.*, λ, ϵ, κ, γ, which are the weights in the learning objective and influence the final performance significantly.

We firstly vary λ to investigate the influence of stochastic optimal transport. Specifically, increasing λ consistently improves the precision of ITE estimates. For example, the out-of-sample PEHE reduces from 3.22 at $\lambda = 0$ to approximately 2.85 at $\lambda = 1.5$. However, assigning a higher focus on distribution balancing in a multitask learning framework leads to difficulties in factual outcome estimation, which in turn leads to sub-optimal ITE estimates. Afterwards, we vary ϵ to investigate the entropic regularizer. It does accelerate the computation of optimal transport discrepancy with large ϵ corresponding to faster computation speed (Flamary et al., 2021). However, it comes with a biased transport plan, evidenced by the out-of-sample PEHE increasing to 2.95 with fluctuations.

We further vary γ and κ to showcase the influence of PFOR and RMPR, respectively. PFOR benefits ITE estimation, while assigning large weight to the proximal outcome distance is detrimental to ITE estimates, because the unit-wise distance calculated with representations (covariates) would be neglected. Relaxing the mass preserving constraint via RMPR could significantly improve the ITE estimates; however, too-small κ is always detrimental as we cannot guarantee the representations across treatment groups are drawn closer together in the optimal transport approach.

5 Related Works

Current works mitigate the treatment selection bias by balancing the distributions of treated and untreated groups, which can be divided into three groups according to the balancing methods used.

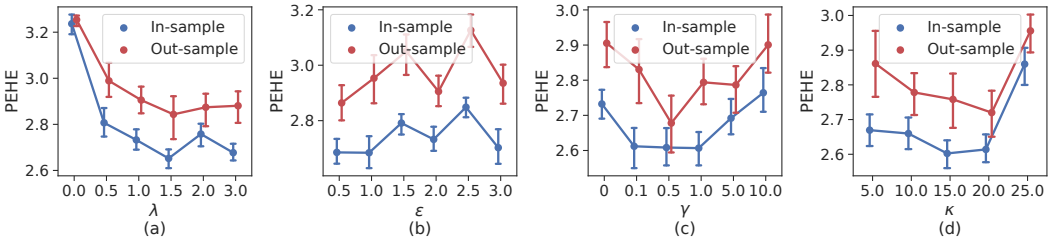


Figure 5: Parameter sensitivity study for critical hyper-parameters of ESCFR

Reweighting methods weight individuals with *balanced score* to obtain globally balanced distributions, represented by the inverse propensity score (IPS) approach (Rosenbaum & Rubin, 1983a) and its doubly robust variant (Robins et al., 1994). Imai & Ratkovic (2014) and Fong et al. (2018) propose to calculate the balancing score via an optimization problem. Kuang et al. (2017b) and Kuang et al. (2017a) further consider the non-confounding factors in covariates. However, these methods suffer from high variance and are vulnerable to non-overlapped units.

Matching methods match similar units from different groups to construct locally balanced distributions. Representative methods (Rosenbaum & Rubin, 1983b; Chang & Dy, 2017; Li et al., 2016) mainly differ in terms of the similarity measurements. Tree based methods (Wager & Athey, 2018) can also be considered as matching methods with adaptive neighborhood metrics. However, the computational cost hinders the application of these methods in large-scale scenarios.

Representation-based methods, *e.g.*, BNN (Johansson et al., 2016) and CFR (Uri et al., 2017), balance the distributions in the latent space. Liuyi et al. (2018) and Hassanpour & Greiner (2020) further augment CFR with local similarity and non-confounding factors, respectively. Kallus (2020) and Yoon et al. (2018) propose to balance the distributions of representations with adversarial training. Representation learning has been the primary approach to mitigate the treatment selection bias, owing to its avoidance of the high variance issue and the suitability for large-scale scenarios.

It is necessary to emphasize our difference with emerging causal inference approaches based on optimal transport. Dunipace (2021) augments the IPS method via a propensity score estimator based on optimal transport; however, it is limited by the aforementioned high variance issue. Torous et al. (2021) uses the push forward operator to improve change-in-change models; however, they are designed for multi-phase data which is not available in our case. Li et al. (2022) shares similar settings to us, while they focuses on variable decomposition in latent space and is identical to Uri et al. (2017) in terms of alignment technology. Our contribution lies in investigating the role and flexibility of optimal transport to augment CFR, mitigating the MSE and UCE issues that have been long circumvented in the literature as recent as this year.

6 Conclusion

Due to the effectiveness of mitigating treatment selection bias, representation learning has been the primary approach to estimating individual treatment effect. However, existing methods circumvent the mini-match sampling effects and unobserved confounders, which hinders them from eliminating the treatment selection bias. A principled approach named ESCFR is devised based on a generalized Sinkhorn discrepancy. Extensive experiments demonstrate that ESCFR can largely mitigate MSE and UCE issues, and achieve better performance compared with other baseline models.

There are two directions of future work that we intend to pursue. The first direction attempts to construct the representation map with normalizing flows (Chen et al., 2018), which is invertible and thus satisfies the assumption to the representation map by Uri et al. (2017). Another direction seeks to extend our methodology to industrial applications, *e.g.*, recommender debias (Wang et al., 2022), which has long been dominated by reweighting methods with the high variance issue.

REFERENCES

Jason M. Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *NeurIPS*, pp. 1964–1974, 2017.

Artem Betlei, Eustache Diemert, and Massih-Reza Amini. Uplift modeling with generalization guarantees. In *SIGKDD*, pp. 55–65, 2021.

Jose H. Blanchet, Arun Jambulapati, Carson Kent, and Aaron Sidford. Towards optimal running times for optimal transport. *CoRR*, abs/1810.07717, 2018.

François Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3): 541–593, 2007.

Nicolas Bonneel, Michiel van de Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. *ACM Trans. Graph.*, 30(6):158, 2011.

Yale Chang and Jennifer G. Dy. Informative subspace learning for counterfactual inference. In *AAAI*, pp. 1770–1776, 2017.

Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso. Unbalanced optimal transport through non-negative penalized linear regression. In *NeurIPS*, pp. 23270–23282, 2021.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *NeurIPS*, pp. 6572–6583, 2018.

Mingyuan Cheng, Xinru Liao, Quan Liu, Bin Ma, Jian Xu, and Bo Zheng. Learning disentangled representations for counterfactual regression via mutual information minimization. In *SIGIR*, pp. 1802–1806, 2022.

Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Math. Comput.*, 87(314):2563–2609, 2018.

José M Cordero, Víctor Cristóbal, and Daniel Santín. Causal inference on education policies: A survey of empirical studies using pisa, timss and pirls. *J. Econ. Surv.*, 32(3):878–915, 2018.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *NeurIPS*, pp. 3730–3739, 2017a.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1853–1865, 2017b.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Nonparametric tests for treatment effect heterogeneity. *Rev. Econ. Stat.*, 90(3):389–405, 2008.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, pp. 2292–2300, 2013.

Eric Dunipace. Optimal transport weights for causal inference. *CoRR*, abs/2109.01991, 2021.

Pavel E. Dvurechensky, Alexander V. Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *ICML*, volume 80, pp. 1366–1375, 2018.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *ICML*, volume 139, pp. 3186–3197, 2021.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: python optimal transport. *J. Mach. Learn. Res.*, 22:78:1–78:8, 2021.

Christian Fong, Chad Hazlett, and Kosuke Imai. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Ann. Appl. Stat.*, 12(1):156–177, 2018.

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *ICLR*. OpenReview.net, 2020.

Xiaofeng He, Guoqiang Xu, Cunxiang Yin, Zhongyu Wei, Yuncong Li, Yancheng He, and Jing Cai. Causal enhanced uplift model. In *PAKDD*, volume 13282, pp. 119–131, 2022.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *J. R. Stat. Soc. Series. B. Stat. Methodol.*, 76(1):243–263, 2014.

Arun Jambulapati, Aaron Sidford, and Kevin Tian. A direct $o(1/epsilon)$ iteration parallel algorithm for optimal transport. In *NeurIPS*, pp. 11355–11366, 2019.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *ICML*, pp. 1724–1732, 2017.

Fredrik D. Johansson, Uri Shalit, and David A. Sontag. Learning representations for counterfactual inference. In *ICML*, pp. 3020–3029, 2016.

Fredrik D. Johansson, Uri Shalit, Nathan Kallus, and David A. Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *CoRR*, abs/2001.07426, 2020.

Pearl Judea and Mackenzie Dana. *The book of why: the new science of cause and effect*. Basic books, 2018.

Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *ICML*, volume 119, pp. 5067–5077, 2020.

Wenwei Ke, Chuanren Liu, Xiangfu Shi, Yiqiao Dai, Philip S. Yu, and Xiaoqiang Zhu. Addressing exposure bias in uplift modeling for large-scale online advertising. In *ICDM*, pp. 1156–1161, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. Estimating treatment effect in the wild via differentiated confounder balancing. In *SIGKDD*, pp. 265–274, 2017a.

Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang, and Fei Wang. Treatment effect estimation with data-driven variable decomposition. In *AAAI*, pp. 140–146, 2017b.

Qian Li, Zhichao Wang, Shaowu Liu, Gang Li, and Guandong Xu. Deep treatment-adaptive network for causal inference. *VLDBJ*, pp. 1–16, 2022.

Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In *IJCAI*, pp. 3768–3774, 2016.

Tianyi Lin, Nhat Ho, and Michael I. Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *ICML*, volume 97, pp. 3982–3991, 2019.

Yao Liuyi, Li Sheng, Li Yaliang, Huai Mengdi, Gao Jing, and Zhang Aidong. Representation learning for treatment effect estimation from observational data. In *NeurIPS*, pp. 2638–2648, 2018.

X Nie and S Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2): 299–319, 2020.

Kenneth R Niswander and Myron Gordon. *The women and their pregnancies: the Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke*, volume 73. National Institute of Health, 1972.

Elizabeth L Ogburn and Tyler J VanderWeele. On the nondifferential misclassification of a binary confounder. *Epidemiology*, 23(3):433, 2012.

Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *ICCV*, pp. 460–467, 2009.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 11 (5-6):355–607, 2019.

Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *ICML*, volume 119, pp. 7673–7682, 2020.

Kunzel Soren R, Sekhon Jasjeet S, Bickel Peter J, and Yu Bin. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. U. S. A.*, 116(10): 4156–4165, 2019.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *PKDD*, volume 10535, pp. 737–753, 2017.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.*, 89(427):846–866, 1994.

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

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

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66(5):688, 1974.

Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M. Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *AAAI*, pp. 5612–5619, 2020.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large scale optimal transport and mapping estimation. In *ICLR*, 2018.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *CoRR*, abs/1910.12958, 2019.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. Faster unbalanced optimal transport: Translation invariant sinkhorn and 1-d frank-wolfe. In *AISTATS*, volume 151, pp. 4995–5021, 2022.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *NeurIPS*, pp. 2503–2513, 2019.

Tamar Sofer, David B Richardson, Elena Colicino, Joel Schwartz, and Eric J Tchetgen Tchetgen. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Stat. Sci.*, 31(3):348, 2016.

William Torous, Florian Gunsilius, and Philippe Rigollet. An optimal transport approach to causal inference. *CoRR*, abs/2108.05858, 2021.

Shalit Uri, Johansson Fredrik D, and Sontag David. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, pp. 3076–3085, 2017.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.*, 113(523):1228–1242, 2018.

Hao Wang, Tai-Wei Chang, Tianqiao Liu, Jianmin Huang, Zhichao Chen, Chao Yu, Ruopeng Li, and Wei Chu. Escm^2: Entire space counterfactual multi-task model for post-click conversion rate estimation. In *SIGIR*, 2022.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *ICLR*, 2020.

Renjun Xu, Pelen Liu, Yin Zhang, Fang Cai, Jindong Wang, Shuoying Liang, Heting Ying, and Jianwei Yin. Joint partial optimal transport for open set domain adaptation. In *IJCAI*, pp. 2540–2546, 2020.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. ACE: adaptively similarity-preserved representation learning for individual treatment effect estimation. In *ICDM*, pp. 1432–1437, 2019.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: estimation of individualized treatment effects using generative adversarial nets. In *ICLR*, 2018.

Jiajing Zheng, Alexander D'Amour, and Alexander Franks. Copula-based sensitivity analysis for multi-treatment causal inference with unobserved confounding. *CoRR*, abs/2102.09412, 2021.