SINKHORN DISCREPANCY FOR COUNTERFACTUAL GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

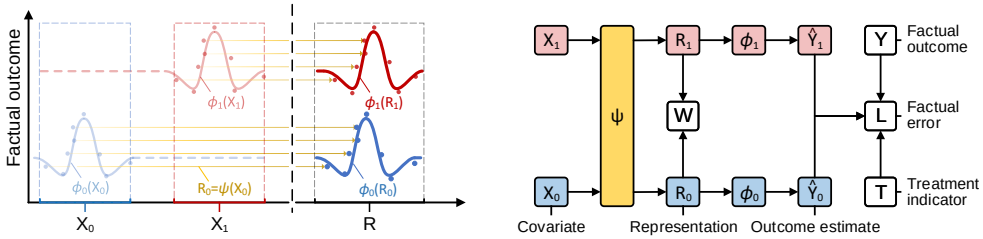
Estimating individual treatment effects from observational data is highly challenging due to the existence of treatment selection bias. Most prevalent approaches mitigate this issue by aligning distributions of different treatment groups in the representation space. However, there are two critical problems circumvented: (1) mini-batch sampling effects (MSE), where the alignment easily fails due to the outcome imbalance or outliers at a mini-batch level; (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 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, drug developers would conduct clinical A/B tests to evaluate the drug effects. Although randomized controlled trials are the gold standard (Pearl & Mackenzie, 2018) for causal inference, it is often prohibitively expensive to conduct such 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 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 has two main challenges: (1) missing counterfactuals, *i.e.*, only one factual outcome out of all potential outcomes can be observed; (2) treatment selection bias, *i.e.*, individuals have their preferences for treatment selection, making units in different treatment groups heterogeneous. To handle missing counterfactuals, meta-learners (Künzel et al., 2019) decompose the ITE estimation task into solvable factual outcome estimation subproblems. However, the treatment selection bias makes it difficult to generalize the factual outcome estimators trained within respective treatment groups to the entire population; consequently, the derived ITE estimator is biased.

Beginning with counterfactual regression (Shalit et al., 2017) and its revolutionary performance, most prevalent methods handle the selection bias by minimizing the distribution discrepancy between groups in the representation space (see Liuyi et al., 2018; Hassanpour & Greiner, 2020; Cheng et al., 2022). However, two critical issues with these methods have long been neglected, which significantly impedes them from handling the treatment selection bias. The first problem is the mini-batch sampling effects (MSE). Specifically, current representation-based methods (Shalit 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 a mini-batch outlier exists in the sampled distribution, the mini-batch discrepancy will be significant, making the training process noise-filled. The second problem is the unobserved confounder effects



(a) Handling treatment selection bias with $\psi(\cdot)$. (b) Architecture of ESCFR.

Figure 1: Overview of handling treatment selection bias with ESCFR. Red (blue) indicates treated (untreated) group. (a) Treatment selection bias causes the shift between X_1 and X_0, impeding ϕ_1 and ϕ_0 to generalize beyond the respective group's properties. Scatters and curves indicate the units and fitted outcome mappings, respectively. (b) ESCFR handles this issue by mapping covariates to an overlapped representation space with $R = \psi(X)$ where ϕ_1 and ϕ_0 are mutually generalizable.

(UCE). Specifically, current approaches directly assume *unconfoundedness* Ma et al. (2022), while the unobserved confounders widely exist in real scenarios and make the resulting estimators biased.

Contributions and outline. In this paper, we propose an effective ITE estimator based on optimal transport, Entire Space CounterFactual Regression (ESCFR), which tackles both the MSE and UCE issues with a generalized sinkhorn discrepancy. Specifically, after 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 investigate the UCE issue in Section 3.3 and propose a proximal factual outcome regularizer to solve it. We finally formulate the architecture and learning objectives of ESCFR in Section 3.4, and report the experimental results in Section 4.

2 Preliminaries

2.1 Causal Inference from Observational Data

This section formulates basic definitions and models in observational causal inference. We first formalize the fundamental elements in Definition 2.1 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 outcomes:

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

where one of these two outcomes is always unobserved. To address such missing counterfactuals, the ITE estimation task is commonly decomposed into potential outcome estimation subproblems that are solvable with any supervised learning method (Künzel et al., 2019). For example, T-learner models the factual outcomes Y for units in the treated and untreated groups separately; S-learner regards the treatment indicator T as one of the covariates X, and models Y for all units simultaneously. The ITE estimate is then the difference of the estimated outcomes when T is set to treated and untreated.

Definition 2.2. *Let $\psi : \mathcal{X} \to \mathcal{R}$ be a mapping from support $\mathcal{X}$ to $\mathcal{R}$, i.e., $\forall x \in \mathcal{X}$, $\exists r = \psi(x) \in \mathcal{R}$. Let $\phi : \mathcal{R} \times \mathcal{T} \to \mathcal{Y}$ be a mapping from support $\mathcal{R} \times \mathcal{T}$ to $\mathcal{Y}$, i.e., 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 abbreviate $\phi(R, T = 1)$ and $\phi(R, T = 0)$ to $\phi_1(R)$ and $\phi_0(R)$, respectively, for brevity.*

[1]We use uppercase letters, *e.g.*, X to denote a random variable, and lowercase letters, *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.*, $\mathbb{P}(X)$.

TARNet (Shalit et al., 2017) obtains better performance by absorbing the advantages of both T-learner and S-learner, consisting of a representation mapping ψ and an outcome mapping ϕ as defined in Definition 2.2. 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 precision in estimation of heterogeneous effect (PEHE):

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

However, according to Figure 1(a), the treatment selection bias causes a distribution shift of covariates across groups, which misleads ϕ_1 and ϕ_0 to overfit their respective group's properties and generalize poorly to the entire population. Therefore, the ITE estimate $\hat{\tau}$ by these methods would be biased.

2.2 Discrete Optimal Transport and Sinkhorn Divergence

Optimal transport (OT) instantiates distribution discrepancy as the minimum transport cost, which provides a grip for quantifying the treatment selection bias in Figure 1(a). Monge (1781) first formulated OT as finding an optimal mapping between two distributions. However, this formulation cannot guarantee the existence and uniqueness of solutions. Kantorovich (2006) proposed a more applicable formulation in Definition 2.3, which can be seen as a generalization of 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.

However, exact solutions (Bonneel et al., 2011) to (4) 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

In this section, we present the proposed Entire Space CounterFactual Regression (ESCFR) approach based on optimal transport to tackle the treatment selection bias. We first illustrate the stochastic optimal transport framework for distribution discrepancy minimization across treatment groups. Based on the framework, we then 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 objectives, and optimization algorithm.

3.1 Stochastic Optimal Transport for Counterfactual Regression

Representation-based approaches mitigate the treatment selection bias by calculating the distribution discrepancy in the representation space and then minimizing it. We select optimal transport to compute the discrepancy since it has shown compelling advantages over its competitors. Specifically, it accounts for 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, the calculated discrepancy can

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

be optimized with the traditional supervised learning framework instead of the adversarial learning framework, and is therefore easier to optimize than adversarial-based methods (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 distributions of representations in treated and untreated groups, respectively, induced by the mapping $r = \psi(x)$. The discrepancy is differentiable with respect to ψ (Flamary et al., 2021), thus can be minimized by updating the representation mapping ψ with gradient-based optimizers.

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 mapping $r = \psi(x)$ defined in Definition 2.2.*

However, since prevalent neural estimators mainly update parameters with stochastic gradient methods, only a fraction of the units is accessible within each iteration. A shortcut in this context is to calculate the group discrepancy at a stochastic mini-batch level:

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

To further investigate the effectiveness of this shortcut, Theorem 3.1 demonstrates that PEHE can be optimized by iteratively minimizing the factual outcome estimation error and the mini-batch group discrepancy (6). The proof of the theorem can be found in Appendix A.3.

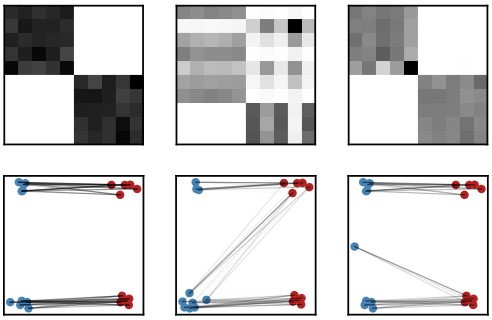
Theorem 3.1. *Let ψ and ϕ be the representation mapping and factual outcome mapping, respectively; $\hat{\mathbb{W}}_{\psi}$ be the group discrepancy at a mini-batch level. With the probability of at least $1-\delta$, we have:*

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

where $\epsilon_{\mathrm{F}}^{T=1}$ and $\epsilon_{\mathrm{F}}^{T=0}$ are the expected errors of factual outcome estimation, N is the batch size, σ_Y^2 is the variance of outcomes, B_{ψ} is a constant term, and $\mathcal{O}(\cdot)$ is a sampling complexity term.

3.2 Relaxed Mass-Preserving Regularizer for Sampling Effect

Although Theorem 3.1 guarantees that the empirical OT discrepancy (6) bounds the PEHE, the sampling complexity term $\mathcal{O}(\cdot)$ inspires us to investigate the potential risks of bad cases caused by stochastic sampling. Precisely, $\mathcal{O}(\cdot)$ results from the discrepancy between the entire population and the sampled mini-batch units (see (30) and (32) in Appendix A) which is highly dependent on the uncontrollable sampling quality. Therefore, the discrepancy measure should be robust to bad sampling cases, otherwise the resulting huge variance will impede it from computing and reducing the actual discrepancy.



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

Figure 2: 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).

The OT discrepancy in (6) can be easily disturbed by many sampling cases, except for the ideal case in Figure 2(a) where the transport strategy is reasonable and applicable. For example, Figure 2(b) demonstrates the mini-batch with outcome imbalance, where units with unrelated factual outcomes are falsely matched; For example, according to Figure 2(b), it falsely matches units with unrelated factual outcomes in the mini-batch where the outcomes between groups are imbalanced; according to Figure 2(c), it falsely matches the mini-batch outliers to normal units, causing a substantial disruption of the transportation strategy. In short, the vanilla OT technique in (6) fails to quantify the group discrepancy for producing erroneous transport strategies in non-ideal mini-batches, and thus misleads the update of the representation mapping ψ. We summarize this phenomenon as the mini-batch sampling effect (MSE) issue [3].

[3] Apart from OT, most prevalent methods (Ma et al., 2022; Liuyi et al., 2018) fail to handle the treatment selection bias for neglecting the MSE issue. The ability to formalize the MSE issue through the mass-preserving constraint is an important advantage of OT over other techniques, as it provides a grip for handling it.

This issue is attributed to the mass-preservation constraint in (5), which requires that all units in both groups match each other, regardless of the actual situation. Mini-batch outliers, for instance, would be compelled to be transported according to Figure 2, which impedes the transport of normal units and the computation of the actual group discrepancy. A small batch size would exacerbate this defect.

Definition 3.2. *For empirical distributions α and β with n and m units, respectively, optimal transport with relaxed mass-preserving constraint seeks the transport strategy $\boldsymbol{\pi}$ at the minimum cost:*

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

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

An intuitive approach to mitigate MSE is to relax the marginal constraint and allow for the creation and destruction of unit's mass. To this end, a relaxed mass-preserving regularizer (RMPR) is devised in Definition 3.2, which replaces the hard marginal constraint in (4) with a soft penalty in (8) for deriving transport strategy. In this context, the stochastic discrepancy is calculated as

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

where the hard mass-preservation constraint is removed to mitigate the MSE issue. Inspired by Fatras et al. (2021), the robustness of RMPR to sampling effects can be further theoretically investigated in Theorem 3.2, where the effect of mini-batch 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 denoting the disturbed distribution as α', we have*

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

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 better 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). Beginning with CFR (Shalit et al., 2017), the unconfoundedness assumption A.1 (see Appendix A) is often taken to circumvent the UCE issue (Ma et al., 2022). Given two units $r_i \in \mathbb{P}_{\psi}^{T=1}(r)$ and $r_j \in \mathbb{P}_{\psi}^{T=0}(r)$, for instance, optimal transport in Definition 3.2 calculates the unit-wise distance as $\mathbf{D}_{ij} := \|r_i - r_j\|^2$. If Assumption A.1 holds, this approach mitigates the treatment selection bias since it blocks the backdoor path $X \to T$ in Figure 3(a) by balancing the confounders across groups in a latent space. However, Assumption A.1 is usually violated in practice as per Figure 3(b), which hinder existing methods including OT from handling treatment selection bias since the backdoor path $X' \to T$ is not blocked.

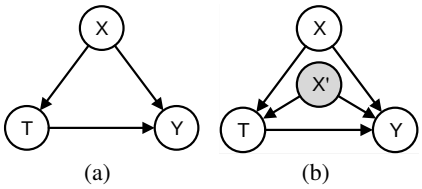


Figure 3: Causal graphs with (a) and w/o (b) the unconfoundedness assumption. The shaded node indicates the hidden confounder X'.

According to Figure 3(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 technique (Courty et al., 2017a), we propose to calibrate the unit-wise distance $\mathbf{D}$ with the potential outcomes as follow:

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

where γ controls the strength of regularization. The underlying regularization is: units with similar (both observed and unobserved) confounders should have similar potential outcomes. As such, for a pair of units with similar observed covariates, *i.e.*, $\|r_i - r_j\|^2 \approx 0$, if their potential outcomes given the same treatment $t = \{0, 1\}$ differ greatly, *i.e.*, $\|y_i^t - y_j^t\| >> 0$, their unobserved confounders should likewise differ significantly. The vanilla OT technique in (6) with $\mathbf{D}_{ij} = \|r_i - r_j\|^2$ would incorrectly

match this pair because $\|r_i - r_j\|^2 \approx 0$, generate a false transport strategy, and consequently bias the update of the representation mapping ψ. In contrast, OT based on $\mathbf{D}_{ij}^{\gamma}$ would not match this pair as the difference of unobserved confounders is compensated with that of potential outcomes.

Moreover, since $y_i^{T=0}$ and $y_j^{T=1}$ in (11) are unavailable due to the missing counterfactual outcomes, the proposed proximal factual outcome regularizer (PFOR) uses their estimates instead. Specifically, let $\hat{y}_i$ and $\hat{y}_j$ be the estimates of $y_i^{T=0}$ and $y_j^{T=1}$, respectively, PFOR refines (11) as

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

Additional justifications, assumptions and limitations of PFOR are discussed in Appendix D.3.

3.4 Architecture of Entire Space Counterfactual Regression

The architecture of ESCFR is presented in Figure 1(b), where the covariate X is first mapped to the representations R with $\psi(\cdot)$, and then to the potential outcomes with $\phi(\cdot)$. The group discrepancy $\mathbb{W}$ is calculated with the optimal transport equipping with the RMPR in (8) and PFOR in (12).

The learning objective is 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 outcome estimation following Shi et al. (2019) can be formulated as

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

where y_i and y_j are the factual outcomes for the corresponding treatment groups. The 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 is in general the optimal transport with RMPR in Definition 3.2, except for the unit-wise distance $\mathbf{D}^{\gamma}$ calculated with the PFOR in (12). Finally, the overall learning objective of ESCFR is

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

where λ controls the strength of distribution alignment, ϵ controls the entropic regularization in (5), κ controls RMPR in (8), and γ controls PFOR in (12). The learning objective (15) mitigates the selection bias following Theorem 3.1 and handles the MSE and UCE issues.

The optimization procedure consists of three steps as summarized in Algorithm 3. First, 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 $\mathbf{D}^{\gamma}$. Second, compute the discrepancy in (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. Finally, calculate the overall loss in (15) and update ψ and ϕ with stochastic gradient methods.

4 Experiments

4.1 Experimental Setup

Datasets. Missing counterfactuals impedes the evaluation of PEHE with observational benchmarks. Following Liuyi et al. (2018); Shalit 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 infants' potential cognitive scores, with 747 observations and 25 covariates; the ACIC dataset comes from the collaborative perinatal project, 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 Künzel et al. (2019)), separate neural regressors for each treatment group (T.learner by Künzel et al. (2019)), TARNet (Shalit 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 Shalit 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**

Training protocol. A fully connected neural network with two 60-dimensional hidden layers is selected to instantiate the representation mapping ψ and the factual outcome mapping ϕ for ESCFR and other neural network based baselines. To ensure a fair comparison, all neural models are trained for 400 epochs with Adam optimizer, with the learning rate and weight decay both set to 0.001. Other settings of optimizers follow Kingma & Ba (2015). We fine-tune hyperparameters within the range in Figure 5, validate performance every two epochs, and save the optimal model for test.

Evaluation protocol. Following Liuyi et al. (2018), PEHE in (3) is used as the precision metric for performance evaluation. However, it is unavailable in the model selection phase due to missing counterfactuals. As such, we use the area under the uplift curve (AUUC) (Betlei et al., 2021) to guide model selection, which measures the counterfactual ranking performance of the model and can be computed without counterfactual outcomes. The within-sample and out-of-sample results are reported on the training and test data, respectively, following Shalit et al. (2017).

4.2 Performance Comparison

Table 1 reports the performance of ESCFR and its competitors over ten runs. Statistical estimators exhibit competitive performance on the PEHE metric. Due to the superiority to depict non-linearity, neural estimators outperform linear and random forest methods. In particular, TARNet that absorbs the advantage of T-learner and S-learner achieves the best overall performance in statistic estimators. However, the circumvention to treatment selection bias leads to inferior performance.

Matching methods, *e.g.*, PSM exhibit compelling ranking performance, which explains why they are favored in counterfactual ranking practice (Betlei et al., 2021). However, their poor performance on PEHE hinders their application in counterfactual estimation applications such as advertising systems that place more emphasis on the accuracy of treatment effect estimation.

Representation-based methods mitigate the treatment selection bias and enhance overall performance. In particular, CFR-WASS reaches an out-of-sample PEHE of 3.207 on ACIC, significantly outperforming most statistical methods. However, the MSE and UCE issues impede these methods from solving the treatment selection bias. The proposed ESCFR achieves significant improvement over most metrics compared with various prevalent baselines [4]. Combined with the comparisons above, we attribute its superiority to the proposed RMPR and PFOR regularizers, which makes it robust to MSE and UCE. See Appendix D.5 for additional comparison results.

4.3 Ablation Study

To verify the effectiveness of individual components, an ablation study is conducted on the ACIC benchmark in Table 2. Specifically, ESCFR first augments TARNet with stochastic optimal transport in Section 3.1, which effectively reduces the out-of-sample PEHE from 3.254 to 3.207. Then, it

[4] 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 typically unavailable in the inference phase. We mainly rely on out-of-sample AUUC instead to evaluate the ranking performance and perform model selection.

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**

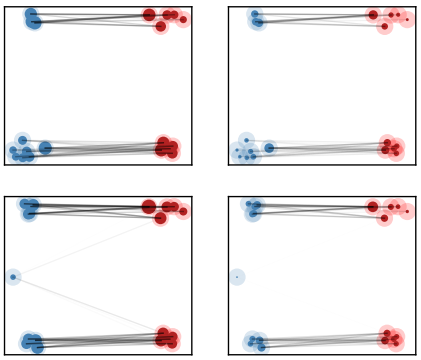
mitigates the MSE issue with RMPR in Section 3.2 and the UCE issue with PFOR in Section 3.3, reducing the out-of-sample PEHE to 2.768 and 2.633, respectively. Finally, ESCFR combines the RMPR and PFOR in a unified framework in Section 3.4, reducing the out-of-sample PEHE to 2.316.

4.4 Analysis of Relaxed Mass-Preserving Regularizer

Most prevalent methods fail to cope with the label imbalance and mini-batch outliers in Figure 2(b-c). Figure 4 shows the transport plan generated with RMPR in the same situations, where RMPR alleviates the MSE issue in both bad cases. RMPR with $\kappa = 10$, for instance, 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. We attribute the success to the relaxed mass-preserving constraint in Section 3.2.

Notably, RMPR does not transport all mass of a unit. The closer the unit is to the overlapping zone in a batch, the greater the mass is transferred. That is, RMPR adaptively matches and pulls closer units that are close to the overlapping 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 (8).

We further investigate the performance of RMPR under different batch sizes and κ in Appendix D.2 to verify the effectiveness of RMPR more thoroughly.



(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 multi-task learning framework leads to difficulties in factual outcome estimation, leading to sub-optimal ITE estimates. Then, we vary ϵ to investigate the entropic regularizer. It does accelerate the computation of optimal transport discrepancy with a large ϵ corresponding to a 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 considerable 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 that the representations across treatment groups are drawn closer together in the optimal transport approach.

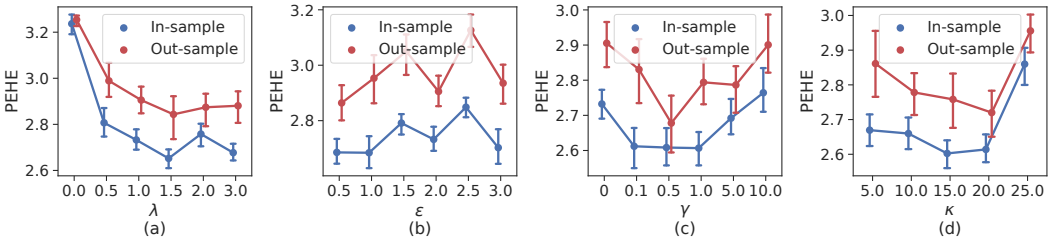


Figure 5: Parameter sensitivity study for critical hyper-parameters of ESCFR

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 methods used for balancing.

Reweighting-based methods weight individuals with *balanced scores* to achieve 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 calculating the balancing score by solving an optimization problem. Kuang et al. (2017b) and Kuang et al. (2017a) consider additional non-confounding factors in covariates. However, these methods are susceptible to non-overlapping units and suffer from a high variance issue.

Matching-based 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 similarity measures. Notably, 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.

Begining with BNN (Johansson et al., 2016) and CFR (Shalit et al., 2017), representation-based methods minimize the group discrepancy 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. Due to its scalability and avoidance of the high variance issue, representation-based methods have been predominant for handling the treatment selection bias.

It is necessary to distinguish ourselves from emerging OT-based causal inference approaches. Dunipace (2021) augments the IPS method with a propensity score estimator based on OT; 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) has a similar setup to to us, while it focuses on the decomposition of latent variables and is identical to Shalit et al. (2017) in terms of alignment technology. Our work is a new take on OT under the CFR framework, alleviating the MSE and UCE issues that have been long neglected by the causal inference community until 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 neglect the mini-match sampling effects and unobserved confounders, which hinders them from handling 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 mapping with normalizing flows (Chen et al., 2018), which is invertible and thus satisfies the assumption to the representation mapping by Shalit et al. (2017). Another direction seeks to extend our methodology to industrial applications, *e.g.*, debias in recommenders (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.

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

Leonid V Kantorovich. On the translocation of masses. *J. Math. Sci.*, 133(4):1381–1382, 2006.

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.

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

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.

Jing Ma, Mengting Wan, Longqi Yang, Jundong Li, Brent Hecht, and Jaime Teevan. Learning causal effects on hypergraphs. In *SIGKDD*, pp. 1202–1212, 2022.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pp. 666–704, 1781.

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

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

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

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.

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 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 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 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.

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

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.

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.