

Proximity Matters: Local Proximity Enhanced Balancing for Treatment Effect Estimation

Hao Wang
College of Control Science and
Engineering, Zhejiang University
Hangzhou, China

Zhichao Chen
College of Control Science and
Engineering, Zhejiang University
Hangzhou, China

Zhaoran Liu
College of Control Science and
Engineering, Zhejiang University
Hangzhou, China

Xu Chen
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China

Haoxuan Li*
Center for Data Science
Peking University
Beijing, China

Zhouchen Lin*[†]
School of Intelligence Science and
Technology, Peking University
Beijing, China

Abstract

Heterogeneous treatment effect (HTE) estimation from observational data poses significant challenges due to treatment selection bias. Existing methods address this bias by minimizing distribution discrepancies between treatment groups in latent space, focusing on global alignment. However, the fruitful aspect of local proximity, where *similar units exhibit similar outcomes*, is often overlooked. In this study, we propose **Proximity-enhanced CounterFactual Regression (CFR-Pro)** to exploit proximity for enhancing representation balancing within the HTE estimation context. Specifically, we introduce a pair-wise proximity regularizer based on optimal transport to incorporate the local proximity in discrepancy calculation. However, the curse of dimensionality renders the proximity measure and discrepancy estimation ineffective—exacerbated by limited data availability for HTE estimation. To handle this problem, we further develop an informative subspace projector, which trades off minimal distance precision for improved sample complexity. Extensive experiments demonstrate that CFR-Pro accurately matches units across different treatment groups, effectively mitigates treatment selection bias, and significantly outperforms competitors. Code is available at <https://github.com/HowardZJU/CFR-Pro>.

CCS Concepts

• **Computing methodologies** → **Machine learning**.

Keywords

Selection Bias, Causal Inference, Optimal Transport, Treatment Effect Estimation

*Haoxuan Li and Zhouchen Lin are the corresponding authors.

[†]Zhouchen Lin is now also working at State Key Lab of General AI & Institute for AI, Peking University, and Pazhou Lab (Huangpu), Guangzhou, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '25, Toronto, ON, Canada

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1454-2/2025/08

<https://doi.org/10.1145/3711896.3737092>

ACM Reference Format:

Hao Wang, Zhichao Chen, Zhaoran Liu, Xu Chen, Haoxuan Li, and Zhouchen Lin. 2025. Proximity Matters: Local Proximity Enhanced Balancing for Treatment Effect Estimation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25)*, August 3–7, 2025, Toronto, ON, Canada. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3711896.3737092>

1 Introduction

Estimating heterogeneous treatment effect (HTE) through randomized controlled trials is fundamental in causal inference, widely applied in various domains such as healthcare [37, 48], e-commerce [1, 45, 54], and education [7]. While randomized controlled trials are considered as the golden standard in HTE estimation [32], their availability is often limited by significant financial and ethical constraints [21, 22]. Consequently, there is increasing reliance on observational data for HTE estimation, driven by its broader availability and the feasibility of post-marketing surveillance as a cost-effective alternative to clinical trials [23, 53].

Estimating HTE from observational data is challenging primarily due to: (1) the absence of counterfactuals, where only one potential outcome is observable; and (2) treatment selection bias, where non-random treatment assignments cause covariate shifts between treated and untreated groups, thereby affecting the generalizability of outcome estimators [50, 52]. Traditional meta-learners address the counterfactual problem by segmenting HTE estimation into tasks focused on factual outcomes [19]. However, these methods often struggle with treatment selection bias, resulting in biased HTE estimations.

Recent methods, such as counterfactual regression, have shown potential for mitigating selection bias by minimizing distribution discrepancies in the representation space [5, 14, 38, 58, 59]. However, current methods for discrepancy calculation overlook two critical issues. First, they emphasize a global perspective in calculating distribution discrepancies, neglecting the local proximity between treatment units. Local proximity—where similar units likely exhibit similar outcomes—is a pivotal factor in accurate HTE estimation [9, 34, 43]. Ignoring this aspect can lead to misleading discrepancy estimates and consequently erroneous updates to the HTE estimator. The second challenge pertains to the curse of dimensionality, where a substantial number of units is required to reliably estimate treatment effects. Often, acquiring a sufficiently

large sample of treated units is impractical in real-world settings, rendering the discrepancy estimation unreliable. Addressing these limitations is essential for advancing the precision and applicability of HTE estimations from observational data.

In this work, we propose an effective HTE estimator, namely **Proximity-enhanced CounterFactual Regression (CFR-Pro)**, which tackles both local proximity and dimensionality issues through a generalized optimal transport problem. Specifically, to incorporate local proximity, CFR-Pro incorporates a pairwise proximity regularizer (PPR) in the optimal transport formulation to explicitly maintain local proximity in discrepancy calculation. To mitigate the curse of dimensionality, CFR-Pro innovatively introduces an informative subspace projector (ISP), which seeks for an informative subspace to calculate distribution discrepancy with minimal precision loss. The architecture and computation workflow of CFR-Pro are detailed in Section 3.3. Extensive experimental results demonstrate that CFR-Pro accurately matches units across varying treatment groups, effectively mitigates treatment selection bias, and significantly outperforms its competitors.

Contributions. The contributions are summarized as follows.

- We innovatively investigate the local proximity preservation and curse of dimensionality issues for causal balancing, which historically limits the performance of HTE estimation based on representation learning.
- We propose CFR-Pro, a streamlined approach that employs a pairwise proximity regularizer and an informative subspace projector within a unified optimal transport framework to overcome the above issues.
- Through comprehensive experiments on open benchmarks, we demonstrate that CFR-Pro outperforms a range of competitors. We further substantiate its effectiveness via extensive hyperparameter tuning and ablative studies.

2 Preliminaries

2.1 HTE estimation with observational data

This section outlines the HTE estimation task within the potential outcome framework [36] and the challenge of treatment selection bias. The fundamental is encapsulated in Definition 2.1¹. Specifically, a unit characterized by covariates x possesses two potential outcomes: Y_1 if treated and Y_0 if untreated. The expected difference between these potential outcomes given covariates, represented as $\tau(x) = \mathbb{E}[Y_1 - Y_0 | x]$, is termed as the conditional average treatment effect (CATE), and its expectation over all units is termed as the average treatment effect (ATE).

Definition 2.1. Suppose X , R , Y , and T are random variables with probability density function ρ_* and support \mathcal{S}_* . Typically, X represents covariates with the probability density function ρ_X and support $\mathcal{S}_X = \{0, 1\}$. R represents induced representations, Y denotes outcomes, and T denotes treatment indicators.

Definition 2.2. Suppose $\psi : \mathcal{S}_X \rightarrow \mathcal{S}_R$ is a representation mapping with $R = \psi(X)$. Define $\phi_T : \mathcal{R} \times \mathcal{T} \rightarrow \mathcal{Y}$ as an outcome

¹We use uppercase letters, e.g., X , to denote a random variable, and lowercase letters, e.g., x , to denote a specific value. Letters in calligraphic font, e.g., \mathcal{X} , represent the empirical distribution, and $\mathbb{P}(\cdot)$ represents the probability distribution function, e.g., $\mathbb{P}(X)$.

mapping that maps the representations and treatment to the corresponding factual outcome: $Y_1 = \phi_1(R)$ and $Y_0 = \phi_0(R)$.

Definition 2.3. Suppose \mathcal{X} , \mathcal{R} , and \mathcal{Y} are the empirical distributions of X , R , and Y at a minibatch level, respectively. Let $\mathcal{X}^{T=1}$ and $\mathcal{X}^{T=0}$ be the covariates of treated and untreated units, respectively, with $\mathcal{R}_{\psi}^{T=1}$ and $\mathcal{R}_{\psi}^{T=0}$ as the corresponding representations induced by $R = \psi(X)$.

The estimation of CATE is the cornerstone in HTE estimation. Since one of these outcomes is always unobserved in a dataset, effective CATE estimation typically involves decomposing it into factual outcome estimation subproblems solvable with supervised regression methods [19]. An exemplary approach TARNet [38] employs the representation mapping ψ and outcome mapping ϕ from Definition 2.2, which estimates the CATE as

$$\begin{aligned} \hat{\tau}_{\psi, \phi}(x) &= \hat{Y}(x, 1) - \hat{Y}(x, 0), \\ \hat{Y}(x, 1) &= \phi_1(\psi(x)), \quad \hat{Y}(x, 0) = \phi_0(\psi(x)), \end{aligned} \quad (1)$$

where ψ is trained across all units, while ϕ_1 and ϕ_0 are trained separately on treated and untreated groups. The training objective is to minimize the factual outcome estimation error:

$$\mathcal{L}^{(F)}(\psi, \phi) := \sum_{i=1}^N \left\| \phi_1(\mathcal{R}_{\psi, i}^{T=1}) - \mathcal{Y}_i^{T=1} \right\|_2^2 + \sum_{j=1}^M \left\| \phi_0(\mathcal{R}_{\psi, j}^{T=0}) - \mathcal{Y}_j^{T=0} \right\|_2^2, \quad (2)$$

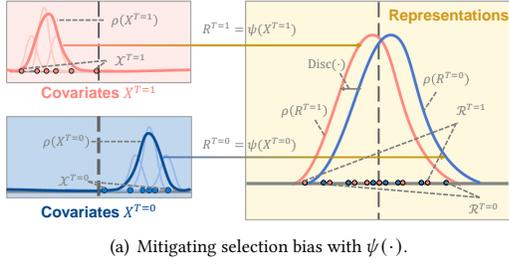
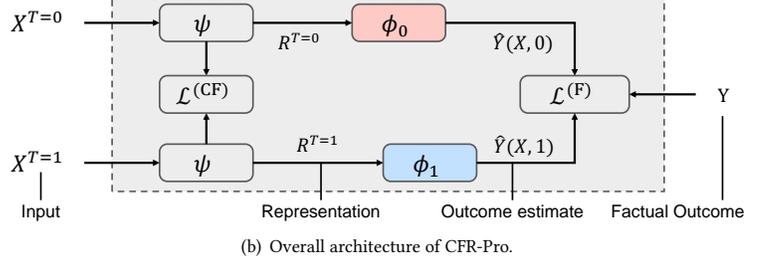
where \mathcal{R}_{ψ} and \mathcal{Y} are the empirical distributions of representations and outcomes as defined in Definition 2.3, and i and j are sample indices in the associated empirical distribution. CATE estimators are evaluated using the precision in estimation of heterogeneous effect (PEHE) metric:

$$\epsilon_{\text{PEHE}}(\psi, \phi) := \int \left(\hat{\tau}_{\psi, \phi}(x) - \tau(x) \right)^2 \rho(x) dx. \quad (3)$$

Selection bias. As illustrated in Figure 1(a), treatment selection bias introduces a distribution shift of covariates between groups, which causes ϕ_1 and ϕ_0 to overfit to their respective group's characteristics and generalize poorly across the entire population. To mitigate selection bias, seminal works starting from CFR [38] augment the learning objective with a *distribution discrepancy* term as $\mathcal{L}^{(F)} + \text{Disc}(\mathcal{R}_{\psi}^{T=1}, \mathcal{R}_{\psi}^{T=0})$. This adjustment reduces distribution shift in the representation space, thereby enabling ϕ_1 and ϕ_0 to generalize to both treated and untreated groups.

2.2 Local proximity for HTE estimation

Local proximity, quantified as the mutual distance between units within a distribution, encapsulates the geometric properties of distributions. The assumption that *similar units have similar outcomes* [58, 59] highlights its critical role in HTE estimation. This principle is central to various HTE estimators—such as matching techniques (e.g., KNN [9], propensity score matching [34]) and stratification methods [43]—that leverage proximity to enhance estimation accuracy. Despite its acknowledged importance, modern HTE approaches, particularly those based on the Counterfactual Representation (CFR) paradigm, primarily focus on minimizing a global discrepancy metric, denoted as $\text{Disc}(\cdot)$, while neglecting the nuances of local proximity that can be pivotal for precise causal inference.

(a) Mitigating selection bias with $\psi(\cdot)$.

(b) Overall architecture of CFR-Pro.

Figure 1: Overview of handling treatment selection bias with CFR-Pro. The red and blue colors signify the treated and untreated groups, respectively. (a) The treatment selection bias is illustrated through a distribution shift between treated (X_1) and untreated (X_0) units. The curves and scatters indicate the probability density functions and associated empirical distributions, respectively. (b) CFR-Pro reduces selection bias by aligning units from both treatment groups within a common representation space, denoted as $R = \psi(X)$. This alignment facilitates the generalization of the outcome mappings ϕ_1 and ϕ_0 across different groups.

One notable exception is the SITE model [58], which incorporates the PDDM metric [15] to measure proximity. SITE differs fundamentally from our work in two respects. Firstly, SITE employs PDDM merely to align latent and covariate spaces while using a simplistic middle-point distance as its discrepancy measure, thereby not integrating local proximity into the discrepancy for balancing treated and untreated samples. In contrast, we focus on enhancing the discrepancy by comprehensively incorporating local proximity. Secondly, SITE quantifies proximity using only six pre-selected anchor units, which may limit its ability to capture broader contextual information. Our approach transcends this limitation by leveraging an optimal transport methodology to construct a more inclusive and nuanced representation of local proximity.

2.3 Discrete optimal transport

Optimal transport (OT) quantifies distribution discrepancy as the minimum transport cost [8, 47, 51], offering a tool to quantify the treatment selection bias in Figure 1(a). An applicable formulation proposed by Kantorovich [17] is present in Definition 2.4, which can be seen as a linear programming problem.

Definition 2.4. 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 the minimum cost:

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

where $\mathbb{W}(\alpha, \beta) \in \mathbb{R}$ is the OT discrepancy; $\mathbf{D} \in \mathbb{R}_+^{n \times m}$ is the unit-wise Euclidean distance between α and β ; \mathbf{a} and \mathbf{b} are the mass of units in α and β , respectively; Π is the feasible plan set where the mass-preserving constraint holds.

3 Proposed method

In this section, we present the Proximity-enhanced Counterfactual Regression (CFR-Pro) approach, which leverages OT to tackle the treatment selection bias. We first illustrate the pair-wise proximity regularizer (PPR) for measuring and maintaining local proximity

in different treatment groups, and demonstrate its efficacy for improving HTE estimation. Subsequently, we propose an informative subspace projector (ISP) to reduce the sampling complexity and handle the curse of dimensionality in calculating OT. We finally open a new thread to summarize the model architecture, learning objectives, and optimization algorithm.

3.1 Pair-wise proximity regularizer for the preservation of local proximity

To mitigate treatment selection bias, representation-based methods align treated and untreated groups in the representation space, the core of which is the quantification of the distribution discrepancy $\text{Disc}(\cdot)$ between treatment groups. It is plausible to quantify the discrepancy with OT due to its numerical advantages and flexibility over competitors [48]. However, standard OT overlooks local proximity, a crucial aspect in HTE estimation. The treated and untreated units with similar neighbors for instance should have a higher probability of matching together since similar units have similar outcomes [34, 43].

An extension of OT that encodes local proximity is the Gromov-Wasserstein measure, primarily applied to matching objects with geometric structures [30, 55]. On the basis, inspired by [41], the PPR fuses the Gromov Wasserstein measure and restates the transport problem in the representation space as:

$$\mathbb{F}(\mathcal{R}_{\psi}^{T=0}, \mathcal{R}_{\psi}^{T=1}) := \min_{\pi \in \Pi(\alpha, \beta)} \left(\kappa \langle \pi, \mathbf{D} \rangle + (1 - \kappa) \sum_{i,j,k,l} \mathbf{P}_{i,j,k,l} \pi_{i,j} \pi_{k,l} \right), \quad (5)$$

where $0 \leq \kappa \leq 1$ controls the relative strength. The first term, following the standard OT formulation in (4), measures the global discrepancy between treatment groups with $\mathbf{D}_{i,j} = \left\| \mathcal{R}_{\psi,i}^{T=0} - \mathcal{R}_{\psi,j}^{T=1} \right\|_2^2$. The second term measures local proximity within each treatment group as $\mathbf{D}_{i,j}^t = \left\| \mathcal{R}_{\psi,i}^{T=t} - \mathcal{R}_{\psi,j}^{T=t} \right\|_2^2$, and incorporates such local proximity via $\mathbf{P}_{i,j,k,l} = \left\| \mathbf{D}_{i,k}^{T=0} - \mathbf{D}_{j,l}^{T=1} \right\|_2^2$. Specifically, if the distance between $\mathcal{R}_{\psi,i}^{T=0}$ and $\mathcal{R}_{\psi,k}^{T=0}$ is close to that between $\mathcal{R}_{\psi,j}^{T=1}$ and $\mathcal{R}_{\psi,l}^{T=1}$ (i.e., $\left\| \mathbf{D}_{i,k}^{T=0} - \mathbf{D}_{j,l}^{T=1} \right\|_2^2 \rightarrow 0$), a higher volume of mass will be matched,

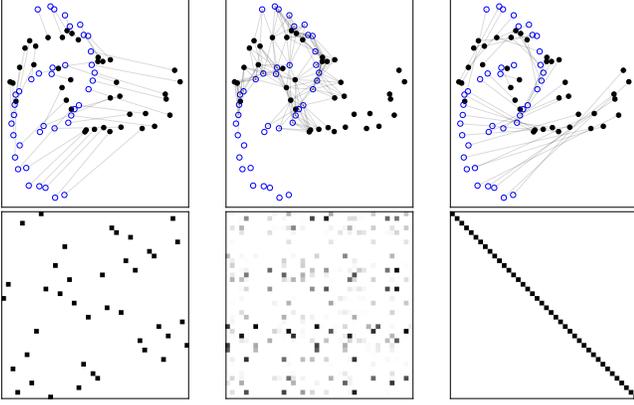


Figure 2: Overview of the transport strategies in three HTE estimators: CFR [38] (left), ESCFR [48] (center) and Ours (right). The upper panels visualize the sample locations of two different treatment groups, where different scatter colors indicate different treatments. The generated transport strategies are marked with gray lines. The down panels elaborate on the generated transport strategies by visualizing the transport matrices π . The darkness indicates the mass transported.

indicated by a larger $\pi_{i,j}\pi_{k,l}$. Conversely, less mass will be transported. The derived transport plan encourages matching units with similar neighbors, preserving local proximity. Therefore, \mathbb{F} quantifies the discrepancy between treatment groups while accommodating the preservation of local proximity.

Case study. To showcase the importance of preserving local proximity, a toy dataset is provided in Figure 2. The untreated group units are simulated from a two-moon distribution (black circles) and treated group units are generated by rotating these positions 90 degrees (blue circles) and adding a small horizontal shift. This setup mimics shifts in both mean and covariance across treatment groups. The ideal transport strategy should align pre- and post-rotation samples, manifested by a primarily diagonal transport matrix. We compare the strategies used in CFR [38], ESCFR [48] and the PPR-enhanced CFR. Key observations are presented below.

- The canonical Wasserstein discrepancy in CFR [38] targets global alignment between groups. However, it could cause erroneous matching as it does not account for local proximity, which that can mislead the update of HTE estimators.
- The unbalanced Wasserstein discrepancy in ESCFR [48] discards skewed samples and concentrates on aligning units that overlap between treatment groups. Despite improvements, this strategy can still produce a noisy and blurred transportation of units.
- Our method integrates PPR to exploit local proximity, effectively rectifying the transport strategy and ensuring precise matching of all samples. This property is essential for generating accurate gradient signals to update representation mappings.

Theoretical investigation. Although PPR has shown practical efficacy, questions remain regarding its contribution to minimizing PEHE. Theorem 1 (refer to Appendix A.2 for proof) offers a theoretical bound, which demonstrates that PEHE can be optimized by

minimizing the estimation error of factual outcomes and the group discrepancy with PPR term. Notably, the integration of PPR slightly expands the theoretical bound compared to that of canonical CFR, yet it is promising to trade off some tightness of the bound for the preservation of local proximity, due to its importance to produce a viable transport strategy, as demonstrated in Figure 2.

THEOREM 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:*

$$\epsilon_{\text{PEHE}}(\psi, \phi) \leq 2[\epsilon_{\mathbb{F}}^{T=1}(\psi, \phi) + \epsilon_{\mathbb{F}}^{T=0}(\psi, \phi) + B_{\psi, \kappa} \mathbb{F}(\mathcal{R}_{\psi}^{T=0}, \mathcal{R}_{\psi}^{T=1})] - 4\sigma_Y^2 + O(N^{-\frac{2}{d}}), \quad (6)$$

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

3.2 Informative subspace projector for the curse of dimensionality

The curse of dimensionality refers to the phenomenon where the Euclidean distance between data points tend to be identical [12]. It renders Euclidean distance difficult to model the proximity in high dimensional settings due to diminished discrimination.

From a computational view, the diminishing discrimination necessitates more samples for estimating Euclidean-based discrepancy [29], which is a pivotal component in state-of-the-art HTE estimators such as MMD in [38], PDDM in [58], and EMD in [48]. A well-known result states, for instance, that the sample complexity of EMD can grow exponentially with dimension [6]. Similarly, the sample complexity of \mathbb{F} reaches $O(N^{-\frac{2}{d}})$, forming the complexity term in Theorem 3.1. Such large sample complexity necessitates many treated and untreated units to faithfully estimate the true discrepancy. However, a large number of treated units is often difficult to acquire in real-world experiments, underscoring the adverse impact of the curse of dimensionality on HTE estimation.

To handle the curse of dimensionality, a common strategy involves reducing the dimension of the computational space. A naive approach is reducing the hidden dimension of the representation mapping ψ . However, it can be counterproductive as it may limit its capacity, preventing the mapping from capturing complex patterns and nonlinearities in data. In this study, we propose identifying an informative subspace and calculating pairwise distances within this subspace, thus mitigating the curse of dimensionality while preserving the full capacity of ψ . Specifically, given a projector $U \in \mathbb{R}^{d \times k}$ to transform the data from a high-dimensional space d to a lower-dimensional subspace k . The distance between two unit representations \mathcal{R}_i and \mathcal{R}_j can be computed as $D_{i,j}^P = \|\mathcal{R}_i U - \mathcal{R}_j U\|$. This approach alleviates the curse of dimensionality but introduces the risk of losing significant information, potentially leading to overly optimistic discrepancy estimations.

The central problem is how to find the informative subspace projector. To reduce the information loss caused by naive dimension reduction, it is feasible to determine the projector in an adversarial

manner as follow:

$$\min_{\pi \in \Pi(\mathcal{R}_{\psi}^{T=0}, \mathcal{R}_{\psi}^{T=1})} \max_{UU^{\top}=I} \left(\kappa \cdot \langle \pi, \mathbf{D}^U \rangle + (1 - \kappa) \cdot \sum_{i,j,k,l} \mathbf{P}_{i,j,k,l}^U \pi_{i,j} \pi_{k,l} \right),$$

where $\mathbf{D}_{i,j}^U = \left\| \mathcal{R}_{\psi,i}^{T=0} U - \mathcal{R}_{\psi,j}^{T=1} U \right\|_2^2$ is the distance measured in the reduced k -dimensional space; $\mathbf{P}_{i,j,k,l}^U = \left\| \mathbf{D}_{i,k}^{U,T=0} - \mathbf{D}_{j,l}^{U,T=1} \right\|_2$. However, this optimization problem proves difficult to solve [29]. An effective compromise involves projecting the data into a k -dimensional subspace through an ISP module that maximally preserves the information, and then compute discrepancy in the subspace. Built upon this idea, the transport problem modified with the ISP module is formulated in Definition 3.1.

Definition 3.1. Suppose $U^* \in \mathbb{R}^{d \times k}$ is an informative subspace projector that is obtained by:

$$U^* = \arg \min_{UU^{\top}=I} \left\| \mathcal{R} - \mathcal{R}UU^{\top} \right\|_2^2, \quad (7)$$

where $P = k/d$ denotes the ratio of dimensionality reduction. The distribution discrepancy equipped with PPR and ISP modules is formulated as

$$\mathbb{P}_{\kappa,P}(\psi) := \min_{\pi \in \Pi(\mathcal{R}_{\psi}^{T=0}, \mathcal{R}_{\psi}^{T=1})} \left(\kappa \cdot \langle \pi, \mathbf{D}^{U^*} \rangle + (1 - \kappa) \cdot \sum_{i,j,k,l} \mathbf{P}_{i,j,k,l}^{U^*} \pi_{i,j} \pi_{k,l} \right). \quad (8)$$

3.3 Overall workflow of CFR-Pro

The architecture of CFR-Pro 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 learning objective is to minimize the risk of factual outcome estimation and the group discrepancy. Given mini-batch distributions $\mathcal{X}^{T=1}$ and $\mathcal{X}^{T=0}$, the risk of factual outcome estimation can be estimated as (2). Afterwards, the group discrepancy is calculated as $\mathbb{P}_{\kappa,P}(\psi)$. Finally, the overall learning objective of CFR-Pro is

$$\mathcal{L}_{\lambda,\kappa,P}^{(\text{CFR-Pro})} := \mathcal{L}^{(\text{F})}(\psi, \phi) + \lambda \cdot \mathbb{P}_{\kappa,P}(\psi), \quad (9)$$

where λ controls the strength of distribution alignment, κ controls PPR in (5), and P controls the ratio of dimension reduction in Definition 3.1. The learning objective above mitigates the selection bias following Theorem 1 while handling the issues of local proximity and curse of dimensionality.

The optimization procedure of CFR-Pro is encapsulated in Algorithm 1. First, we compute the latent space representations using the representation mapping ψ . Second, we determine the informative subspace projector U^* by solving the dimension reduction problem in (7), which can be solved via the well-established principal component analysis. Then, we calculate the pair-wise distance matrix \mathbf{D}^U and \mathbf{P}^U in the subspace induced by U^* . Subsequently, compute the distribution discrepancy term \mathbb{P} by solving the OT problem in Definition 3.1. The solution process is available in [41]. Finally, we compute the overall loss in (9) and update ψ and ϕ with stochastic gradient methods.

4 Experiments

We validate CFR-Pro by investigating the aspects as follows:

Algorithm 1

The computation workflow of CFR-Pro.
Input: covariates \mathcal{X} ; factual outcomes \mathcal{Y} ; outcome mapping ϕ ; treatments \mathcal{T} ; representation mapping ψ .

Parameter: λ : strength of discrepancy alignment; κ : strength of proximity preservation in PPR; P : ratio of dimension reduction. B : batch size

Output: $\mathcal{L}_{\lambda,\kappa,P}^{(\text{CFR-Pro})}$: the learning objective of CFR-Pro.

- 1: $\mathcal{R} \leftarrow \psi(\mathcal{X})$.
 - 2: $U^* = \arg \min_{UU^{\top}=I} \left\| \mathcal{R} - \mathcal{R}UU^{\top} \right\|_2^2$.
 - 3: $\mathbf{D}_{i,j}^{U^*} \leftarrow \left\| \mathcal{R}_i^{T=0} U^* - \mathcal{R}_j^{T=1} U^* \right\|_2^2$ for $1 \leq i, j \leq B$.
 - 4: $\mathbf{D}_{i,k}^{U^*,T=0} \leftarrow \left\| \mathcal{R}_i^{T=0} U^* - \mathcal{R}_k^{T=0} U^* \right\|_2^2$ for $1 \leq i, k \leq B$.
 - 5: $\mathbf{D}_{j,l}^{U^*,T=1} \leftarrow \left\| \mathcal{R}_j^{T=1} U^* - \mathcal{R}_l^{T=1} U^* \right\|_2^2$ for $1 \leq j, l \leq B$.
 - 6: $\mathbf{P}_{i,j,k,l}^{U^*} \leftarrow \left\| \mathbf{D}_{i,k}^{U^*,T=0} - \mathbf{D}_{j,l}^{U^*,T=1} \right\|_2$ for $1 \leq i, j, k, l \leq B$.
 - 7: Calculate $\mathbb{P}_{\kappa,P}(\psi)$ following Eq. (8).
 - 8: $\mathcal{L}^{(\text{F})}(\psi, \phi) \leftarrow \left\| \phi(\mathcal{R}, \mathcal{T}) - \mathcal{Y} \right\|_2^2$.
 - 9: $\mathcal{L}_{\lambda,\kappa,P}^{(\text{CFR-Pro})} \leftarrow \mathcal{L}^{(\text{F})}(\psi, \phi) + \lambda \cdot \mathbb{P}_{\kappa,P}(\psi)$.
-

- (1) **Performance:** *Does CFR-Pro work?* Section 4.2 compares CFR-Pro against established CATE estimators on real-world public benchmarks, where CFR-Pro achieves the best performance.
- (2) **Efficacy:** *How does it work?* Section 4.3 conducts an ablative study to investigate the contribution of PPR and ISP, where both components are beneficial to improve canonical CFR and cooperate well.
- (3) **Sensitivity:** *Is it sensitive to hyperparameters?* Section 4.4 conducts a sensitivity study on the hyperparameters introduced by PPR and ISP, respectively, and give further insights on the rational they work.

4.1 Experimental setup

Datasets. The evaluation of PEHE is challenged by the absence of counterfactuals in observational data. To address this challenge, experiments are conducted using two semi-synthetic datasets: the Infant Health and Development Program (IHDP) and the Atlantic Causal Inference Conference (ACIC) competition data [38, 58]. The IHDP dataset evaluates the effect of specialist home visits on infants' cognitive development, comprising 747 observations with 25 covariates. The ACIC dataset, derived from the Collaborative Perinatal Project [31], includes 4802 observations and 58 covariates. All datasets are randomly shuffled and partitioned into training, validation, and test sets in a 0.7:0.15:0.15 ratio, maintaining the same proportion of treated units across all splits to ensure numerical reliability. To increase the distinguishability of results, we omit dataset scaling to heighten the impact of selection bias.

Baselines. The baselines can be categorized into three groups.

- (1) **Direct estimators:** R.Forest [44], S.learner [19], T.learner [19], TARNet [38], DESCN [63];
- (2) **Matching estimators:** PSM [34], k -NN [9], O.Forest [43];
- (3) **Representation-based estimators:** CFR-MMD [38], CFR-WASS [38], ESCFR [48], and SITE [58].

Implementation details. CFR-Pro is implemented with a fully connected neural network architecture comprising two hidden layers with 16-16 and 32-32 nodes, respectively. It is trained using

Table 1: Out-of-sample performance (mean±std) on the ACIC and IHDP datasets. “*” marks the baseline estimators that CFR-Pro outperforms significantly at p-value < 0.05 over paired samples t-test.

Dataset	ACIC			IHDP		
	€PEHE	€ATE	€ATT	€PEHE	€ATE	€ATT
R.Forest	3.3908±0.1811*	0.8347±0.3635	0.7785±0.3816	4.6697±9.2920	0.4544±0.8308	0.8353±1.2413
S.Learner	4.8835±0.7933*	3.0913±0.7731*	3.1213±0.6372*	4.7408±3.9688*	2.5785±1.8521*	2.7951±1.6036*
T.Learner	4.2749±0.6793*	2.2176±1.2131*	2.3940±1.1964*	2.5257±3.3643*	0.5818±1.2177	0.6642±1.2197
TARNet	3.5331±0.9556*	1.5308±1.0469*	1.6601±1.0670*	1.7781±3.4467	0.2814±0.3193	0.3338±0.3349
DESCN	2.6420±0.2614*	0.4548±0.1693	0.4987±0.1881	4.0128±6.1409*	1.2219±1.7453	0.7917±1.3185
k-NN	5.8977±0.1400*	1.5773±0.3075*	1.9068±0.2870*	4.3191±7.3361	0.8316±1.6911	1.8118±3.2342
O.Forest	2.7451±0.3379*	0.6003±0.1879	0.6597±0.2013	3.1888±5.6657	0.3150±0.3696	0.6539±0.5370*
PSM	5.1014±0.2987*	0.6468±0.3478	0.6231±0.3465	4.6347±8.5748	0.2129±0.3362	0.9353±2.7094
CFR-MMD	3.8514±0.4558*	1.7379±0.9133*	1.9060±0.9290*	1.9398±2.9029	0.5870±1.2231	0.6678±1.2207
CFR-WASS	3.3187±0.7622*	1.3581±1.0325*	1.4682±1.0636*	1.9252±2.9323	0.5578±1.2455	0.6532±1.2544
SITE	3.4910±0.7799*	1.3425±1.1929	1.5443±1.2128*	1.7339±3.1709	0.2271±0.3140	0.2525±0.2805
ESCFR	2.6780±0.6566	1.1468±0.8146*	1.2365±0.8689*	1.6299±3.0344	0.2135±0.3788	0.2319±0.2488
CFR-Pro	2.0413 ±0.6646	0.4551 ±0.3845	0.5034 ±0.4221	1.4601 ±2.6607	0.1079 ±0.1087	0.2224 ±0.2472

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

Model	PPR	ISP	In-sample			Out-sample		
			€PEHE	€ATE	€ATT	€PEHE	€ATE	€ATT
CFR	✗	✗	3.4288±0.3952*	1.1796±0.6443*	1.9186±0.8632*	3.3187±0.7622*	1.3581±1.0325*	1.4682±1.0636*
CFR [†]	✓	✗	2.9668±0.9142	0.9162±0.5930	1.3961±0.9425	2.5193±0.7771	0.9164±0.8203	1.0020±0.8903
CFR [‡]	✗	✓	2.9341±0.7583	0.7825±0.5363	1.2473±0.7007	2.5983±0.7378*	0.8303±0.7695	0.9080±0.8328
CFR-Pro	✓	✓	2.6091 ±0.7673	0.5384 ±0.3932	1.0313 ±0.7206	2.0413 ±0.6640	0.4551 ±0.3845	0.5034 ±0.4221

the Adam optimizer for a maximum of 400 epochs, with an early stopping patience set to 30 epochs. The learning rate and weight decay parameters are set at $1e^{-3}$ and $1e^{-4}$, respectively. Other optimization settings follow Kingma and Ba [18]. Hyper-parameters are tuned on the validation set within the ranges in Section 4.4, with model performance validation conducted every epoch.

4.2 Overall performance

Table 1 provides a comprehensive comparison of the CFR-Pro framework with various baseline methodologies. Key observations from this comparative analysis are outlined below:

- Direct estimators exhibit strong performance on the PEHE metric. Neural network-based estimators particularly excel, surpassing linear models and random forests due to their enhanced ability to capture nonlinear relationships. Among them, TARNet, which integrates the strengths of both T-learner and S-learner, achieves superior and stable performance on both datasets. However, the limitations in addressing treatment selection bias hamper their performance in certain scenarios.
- Matching methods such as PSM and O.Forest show robust capabilities in estimating average treatment effects, which contributes to their widespread adoption in policy evaluation contexts. However, their efficacy diminishes on the PEHE metric, restricting

their suitability for applications requiring personalized treatments (e.g., advertising).

- Representation-based methods effectively address treatment selection bias and enhance HTE estimation performance. However, their oversight of local proximity and curse of dimensionality restricts their effectiveness in overcoming selection bias.
- CFR-Pro surpasses other prevalent baselines with significant improvements across most metrics. This superiority is attributed to the innovative PPR and ISP modules. These components cooperate to enable CFR-Pro to adeptly harness the local proximity and mitigate the curse of dimensionality, which facilitates more accurate alignment of treatment groups and thereby handling of selection bias.

4.3 Ablation study

In Table 2, we examine the contributions of individual components of CFR-Pro on the ACIC benchmark. Our study builds upon the CFR-Wass model [38], a canonical approach that employs the Wasserstein discrepancy to align treatment groups within the representation space.

- In CFR[†], we enhance CFR by involving PPR, where the Wasserstein discrepancy \mathbb{W} is replaced by the fused Gromov Wasserstein discrepancy in (5). A significant performance improvement is observed, where the out-of-sample ϵ_{PEHE} decreases from 3.3187

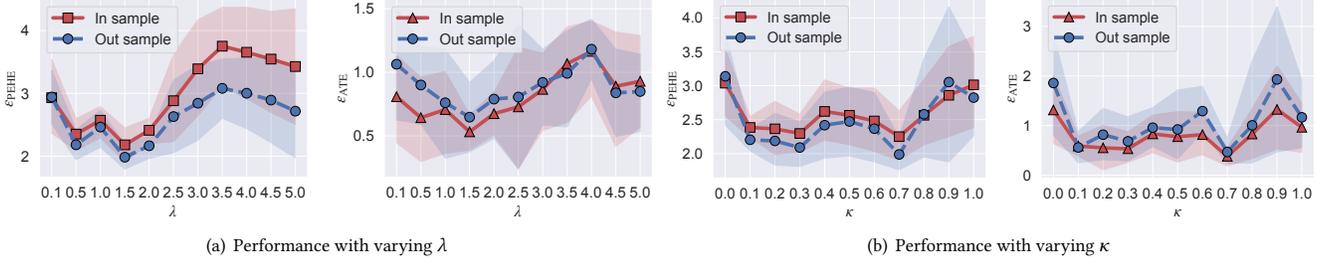


Figure 3: Parameter sensitivity of the PPR module on the ACIC dataset, with focus on λ and κ . The lines and shaded areas indicate the mean values and 90% confidence intervals, respectively.

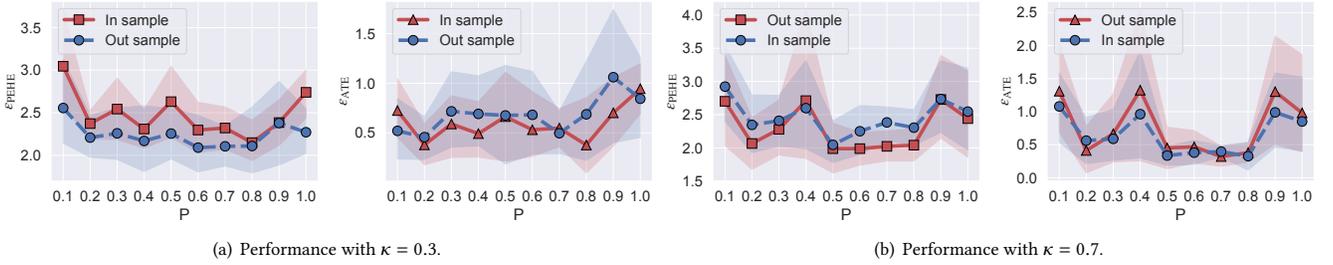


Figure 4: Parameter sensitivity of the ISP module on the ACIC dataset, where P stands for the ratio of dimensionality reduction. The lines and shaded areas indicate the mean values and 90% confidence intervals, respectively.

to 2.5193, underscoring the utility of local proximity when constructing balanced representations.

- In CFR[‡], we enhance CFR by incorporating ISP. Similarly, a huge performance improvement is observed. For instance, the out-of-sample ϵ_{PEHE} decreases from 3.3187 to 2.5983, ϵ_{ATE} decreases from 1.3581 to 0.4551. These performance improvements showcase the utility of ISP to mitigate the curse of dimensionality and generate reliable discrepancy estimates.
- In CFR-Pro, we synthesize PPR and ISP into a unified framework. It maintains the advantages of each individual component and achieves the best overall performance compared to other variants.

4.4 In-depth analysis

Analysis on PPR. We analyze the performance of the PPR component focusing on its two main hyperparameters: λ and κ , which respectively control the strength of distribution alignment and proximity preservation as detailed in (5). The results of a sensitivity study for these parameters are depicted in Figure 3, and the key findings are summarized below.

- Increasing the value of λ from 0.1 to 1.5 leads to a notable decrease in the out-of-sample ϵ_{PEHE} , from approximately 3.0 to 2.2. However, further increases in λ result in a rise in estimation error. The phenomenon indicates that proper distribution alignment is effective to enhance the performance of HTE estimators. However, overly emphasizing distribution balancing within a multi-task learning framework can degrade the accuracy of factual outcomes and, consequently, treatment effect estimates within a multi-task learning framework.

- A similar trend is observed for the strength of proximity preservation. Increasing κ from 0 to 0.1 renders a huge performance improvement. The performance gain is consistent in a broad range from 0.1 to 0.7. However, further increasing κ beyond this range leads to performance reduction. It results from an excessive emphasis on the local proximity in OT formulation, which reduces the focus on global discrepancy and hinders the reduction of selection bias.

Analysis on ISP. The characteristics of ISP are governed by the hyperparameter P , which controls the extent of dimensionality reduction. We conduct a hyperparameter study for P in Figure 4, and the key observations are summarized below.

- Dimensionality reduction effectively enhances model performance. Specifically, as P is decreased from 1 to 0.7, there is a notable improvement in the estimation accuracy, with the out-of-sample ϵ_{ATE} diminishing from approximately 0.8 to about 0.5. This improvement is primarily due to effective handling of the curse of dimensionality, which in turn facilitates a more accurate estimation of discrepancies using minibatch samples. However, excessive reduction in dimensionality can lead to substantial information loss, and thereby suboptimal estimates.
- Interestingly, there is a relationship between the weight of PPR (κ) and the optimal setting for P . As κ increases, which shifts the discrepancy measure \mathbb{P} closer to the Gromov-Wasserstein discrepancy, the curse of dimensionality becomes more pronounced: the Gromov term relies heavily on unit-wise distances to compute local proximity, making dimensionality reduction increasingly crucial. Consequently, the optimal P value tends to decrease with

larger κ , underscoring the need to balance dimensionality reduction against the risk of significant information loss.

5 Related works

5.1 Overview of HTE estimation

The central challenge of HTE estimation is mitigating treatment selection bias by balancing treated and untreated groups. Current solutions include three main categories: reweighting-based, matching-based, and representation-based methods [25].

Reweighting-based methods primarily utilize propensity scores to achieve global balance between groups, which entails the estimation of propensity scores and the construction of unbiased estimators [24, 52]. Propensity scores are typically estimated using logistic regression [4, 10, 20, 61], with improvements via feature selection [39, 49, 50], joint optimization [21, 61, 62], and alternative training techniques [64]. The inverse propensity score method exemplifies the construction of unbiased estimator [35]. However, it suffers from high variance at low propensity scores and bias with incorrect estimates [46]. To handle these issues, doubly robust estimators and variance reduction techniques have been developed [13, 33].

Matching-based methods construct locally balanced distributions by matching comparable units from different groups. These methods mainly differ in terms of the incorporated similarity measures. Propensity score matching for instance uses estimated propensity scores for calculating unit (dis)similarity [34]. Tree-based methods, such as CausalForest [43], can also viewed as a matching-based method employing adaptive similarity measures. Despite their effectiveness, the computational intensity of these methods restricts their scalability in large-scale applications [2, 28, 52].

Representation-based methods seeks for a mapping to a latent space where distributional discrepancies are minimized. Initial approaches advocate maximum mean discrepancy and vanilla Wasserstein discrepancy [16, 38]. Enhancements have been made by integrating feature selection [5, 14], representation decomposition [14, 53], and adversarial training [60]. However, local proximity is a critical yet scarcely investigated aspect to facilitate learning representation. A notable pioneer is SITE [58], which employs the PDDM metric [15] to depict local proximity.

5.2 Optimal transport for HTE estimation

Recent advances in optimal transport (OT) have significantly impacted causality studies, leading to the development of innovative HTE estimators [40]. One research direction involves using OT to enhance reweighting [11, 56, 57] and matching [3] strategies. A related work proposed by Yan et al. [56] uses Gromov discrepancy to adjust the transport matrix for HTE estimation. However, they focuses on reweighting samples within the covariate space, which contrasts with our approach that focuses on aligning representations in the latent space. Additionally, the issue of curse of dimensionality remains a limitation of [56] but handled in our work. These factors differentiate our work with Yan et al. [56].

Another line of works, which are typically related to us, advocates building balanced representations with OT [26, 27, 48]. Li et al. [27] for instance use OT to align factual and counterfactual

distributions; Wang et al. [48] apply OT to achieve HTE estimation while mitigating noise and unobserved confounding effects. Despite this progress, these approaches generally adhere to the traditional Kantorovich problem, similar to [38], focusing on global alignment while often neglecting both local proximity and the curse of dimensionality. Therefore, developing OT formulations for HTE estimation remains a fruitful avenue for future research.

6 Conclusion

Representation learning has emerged as a pivotal approach for HTE estimation. However, current methods often overlook crucial aspects such as local proximity and the curse of dimensionality, which are essential for adequately addressing treatment selection bias. To bridge this gap, a principled approach known as CFR-Pro, based on a generalized OT problem, has been developed. Extensive experiments validate that CFR-Pro handles both problems effectively and outperforms prevalent baseline models.

There are two promising research avenues for further investigation. The first involves the integration of normalizing flows for representation mapping, since their invertibility effectively aligns with the foundational assumptions of counterfactual regression [38]. The second avenue focuses on the practical application of our methodology to industrial contexts, specifically for bias mitigation in recommendation systems [45].

Acknowledgements

This work is supported by National Key R&D Program of China (2022ZD0160300) and the NSF China (No. 62276004, 623B2002).

References

- [1] Artem Betlei, Eustache Diemert, and Massih-Reza Amini. 2021. Uplift Modeling with Generalization Guarantees. In *SIGKDD*. 55–65.
- [2] Yale Chang and Jennifer G. Dy. 2017. Informative Subspace Learning for Counterfactual Inference. In *AAAI*. 1770–1776.
- [3] Arthur Charpentier, Emmanuel Flachaire, and Ewen Gallic. 2023. Optimal Transport for Counterfactual Estimation: A Method for Causal Inference. *arXiv preprint arXiv:2301.07755* (2023).
- [4] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. 2021. Autodebias: Learning to debias for recommendation. In *SIGIR*. 21–30.
- [5] Mingyuan Cheng, Xinru Liao, Quan Liu, Bin Ma, Jian Xu, and Bo Zheng. 2022. Learning Disentangled Representations for Counterfactual Regression via Mutual Information Minimization. In *SIGIR*. 1802–1806.
- [6] L enaic Chizat, Pierre Roussillon, Flavien L eger, Fran ois-Xavier Vialard, and Gabriel Peyr e. 2020. Faster Wasserstein Distance Estimation with the Sinkhorn Divergence. In *NeurIPS*.
- [7] Jos e M Cordero, Victor Crist obal, and Daniel Santin. 2018. Causal inference on education policies: A survey of empirical studies using PISA, TIMSS and PIRLS. *J. Econ. Surv.* 32, 3 (2018), 878–915.
- [8] Nicolas Courty, R emi Flamary, Devis Tuia, and Alain Rakotomamonjy. 2017. Optimal Transport for Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 9 (2017), 1853–1865.
- [9] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2008. Nonparametric tests for treatment effect heterogeneity. *Rev. Econ. Stat.* 90, 3 (2008), 389–405.
- [10] Quanyu Dai, Haoxuan Li, Peng Wu, Zhenhua Dong, Xiao-Hua Zhou, Rui Zhang, Rui Zhang, and Jie Sun. 2022. A generalized doubly robust learning framework for debiasing post-click conversion rate prediction. In *SIGKDD*. 252–262.
- [11] Eric Dunipace. 2021. Optimal transport weights for causal inference. *CoRR* abs/2109.01991 (2021).
- [12] Damien Fran ois, Vincent Wertz, and Michel Verleysen. 2007. The Concentration of Fractional Distances. *IEEE Trans. Knowl. Data Eng.* 19, 7 (2007), 873–886.
- [13] Siyuan Guo, Lixin Zou, Yiding Liu, Wenwen Ye, Suqi Cheng, Shuaiqiang Wang, Hechang Chen, Dawei Yin, and Yi Chang. 2021. Enhanced Doubly Robust Learning for Debiasing Post-Click Conversion Rate Estimation. In *SIGIR*. 275–284.

- [14] Negar Hassanpour and Russell Greiner. 2020. Learning Disentangled Representations for Counterfactual Regression. In *ICLR*.
- [15] Chen Huang, Chen Change Loy, and Xiaou Tang. 2016. Local similarity-aware deep feature embedding. *NeurIPS* 29 (2016).
- [16] Fredrik D. Johansson, Uri Shalit, and David A. Sontag. 2016. Learning Representations for Counterfactual Inference. In *ICML*. 3020–3029.
- [17] Leonid V Kantorovich. 2006. On the translocation of masses. *J. Math. Sci.* 133, 4 (2006), 1381–1382.
- [18] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.
- [19] Sören Reinhold Künzel, Jasjeet Sekhon, Peter Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. U. S. A.* 116, 10 (2019), 4156–4165.
- [20] Jae-won Lee, Seongmin Park, and Jongwuk Lee. 2021. Dual Unbiased Recommender Learning for Implicit Feedback. In *SIGIR*. 1647–1651.
- [21] Haoxuan Li, Kunhan Wu, Chunyuan Zheng, Yanghao Xiao, Hao Wang, Zhi Geng, Fuli Feng, Xiangnan He, and Peng Wu. 2024. Removing hidden confounding in recommendation: a unified multi-task learning approach. *NeurIPS* 36 (2024).
- [22] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, and Peng Wu. 2022. Balancing Unobserved Confounding with a Few Unbiased Ratings in Debaised Recommendations. In *WWW*. 305–315.
- [23] Haoxuan Li, Chunyuan Zheng, Yixiao Cao, Zhi Geng, Yue Liu, and Peng Wu. 2023. Trustworthy Policy Learning under the Counterfactual No-Harm Criterion. In *ICML*, Vol. 202. 20575–20598.
- [24] Haoxuan Li, Chunyuan Zheng, Shuyi Wang, Kunhan Wu, Eric Wang, Peng Wu, Zhi Geng, Xu Chen, and Xiao-Hua Zhou. [n. d.]. Relaxing the Accurate Imputation Assumption in Doubly Robust Learning for Debaised Collaborative Filtering. In *ICML*.
- [25] Haoxuan Li, Chunyuan Zheng, Wenjie Wang, Hao Wang, Fuli Feng, and Xiao-Hua Zhou. 2024. Debaised Recommendation with Noisy Feedback. In *SIGKDD*.
- [26] Qian Li, Zhichao Wang, Shaowu Liu, Gang Li, and Guandong Xu. 2022. Deep treatment-adaptive network for causal inference. *VLDBJ* (2022), 1–16.
- [27] Qian Li, Zhichao Wang, Shaowu Liu, Gang Li, and Guandong Xu. 2023. Causal Optimal Transport for Treatment Effect Estimation. *IEEE Trans. Neural Networks Learn. Syst.* 34, 8 (2023), 4083–4095.
- [28] Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. 2016. Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns. In *IJCAI*. 3768–3774.
- [29] Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael Jordan. 2020. Projection robust Wasserstein distance and Riemannian optimization. *NeurIPS* 33 (2020), 9383–9397.
- [30] Facundo Mémoli. 2011. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics* 11 (2011), 417–487.
- [31] Kenneth R Niswander and Myron Gordon. 1972. *The women and their pregnancies: the Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke*. Vol. 73. National Institute of Health.
- [32] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.
- [33] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* 89, 427 (1994), 846–866.
- [34] Paul Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.
- [35] Paul Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.
- [36] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 5 (1974), 688.
- [37] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim Buhmann, and Walter Karlen. 2020. Learning Counterfactual Representations for Estimating Individual Dose-Response Curves. In *AAAI*. 5612–5619.
- [38] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*. 3076–3085.
- [39] Susan M Shortreed and Ashkan Ertefaie. 2017. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics* 73, 4 (2017), 1111–1122.
- [40] Ruibo Tu, Kun Zhang, Hedvig Kjellström, and Cheng Zhang. 2022. Optimal Transport for Causal Discovery. In *ICLR*.
- [41] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. 2020. Fused Gromov–Wasserstein distance for structured objects. *Algorithms* 13, 9 (2020), 212.
- [42] Cédric Villani. 2009. *Optimal transport: old and new*. Vol. 338. Springer.
- [43] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113, 523 (2018), 1228–1242.
- [44] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113, 523 (2018), 1228–1242.
- [45] Hao Wang, Tai-Wei Chang, Tianqiao Liu, Jianmin Huang, Zhichao Chen, Chao Yu, Ruopeng Li, and Wei Chu. 2022. ESCM²: Entire Space Counterfactual Multi-Task Model for Post-Click Conversion Rate Estimation. In *SIGIR*.
- [46] Hao Wang, Zhichao Chen, Zhaoran Liu, Haozhe Li, Degui Yang, Xinggao Liu, and Haoxuan Li. 2024. Entire space counterfactual learning for reliable content recommendations. *IEEE Trans. Inf. Forensics Security* (2024).
- [47] Hao Wang, Zhichao Chen, Honglei Zhang, Zhengnan Li, Licheng Pan, Haoxuan Li, and Mingming Gong. 2025. Debaised Recommendation via Wasserstein Causal Balancing. *ACM Trans. Inf. Syst.* (2025).
- [48] Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. 2024. Optimal transport for treatment effect estimation. *NeurIPS* 36 (2024).
- [49] Haotian Wang, Kun Kuang, Haoang Chi, Longqi Yang, Mingyang Geng, Wanrong Huang, and Wenjing Yang. 2023. Treatment Effect Estimation with Adjustment Feature Selection. In *SIGKDD*. ACM, 2290–2301.
- [50] Haotian Wang, Kun Kuang, Long Lan, Zige Wang, Wanrong Huang, Fei Wu, and Wenjing Yang. 2023. Out-of-distribution Generalization with Causal Feature Separation. *IEEE Trans. Knowl. Data Eng.* (2023).
- [51] Hao Wang, Haoxuan Li, Xu Chen, Mingming Gong, Zhichao Chen, et al. 2025. Optimal Transport for Time Series Imputation. In *ICLR*.
- [52] Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, and Fei Wu. 2023. Stable Estimation of Heterogeneous Treatment Effects. In *ICML*, Vol. 202. 37496–37510.
- [53] Anpeng Wu, Junkun Yuan, Kun Kuang, Bo Li, Runze Wu, Qiang Zhu, Yueting Zhuang, and Fei Wu. 2023. Learning Decomposed Representations for Treatment Effect Estimation. *IEEE Trans. Knowl. Data Eng.* 35, 5 (2023), 4989–5001.
- [54] Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. 2022. On the opportunity of causal learning in recommendation systems: Foundation, estimation, prediction and challenges. In *IJCAI*. 23–29.
- [55] Hongteng Xu, Dixin Luo, and Lawrence Carin. 2019. Scalable Gromov–Wasserstein learning for graph partitioning and matching. *NeurIPS* 32 (2019).
- [56] Yuguang Yan, Zeqin Yang, Weilin Chen, Ruichu Cai, Zhifeng Hao, and Michael Kwok-Po Ng. 2024. Exploiting Geometry for Treatment Effect Estimation via Optimal Transport. In *AAAI*, Vol. 38. 16290–16298.
- [57] Hao Yang, Zexu Sun, Hongteng Xu, and Xu Chen. 2024. Revisiting Counterfactual Regression through the Lens of Gromov–Wasserstein Information Bottleneck. *arXiv preprint arXiv:2405.15505* (2024).
- [58] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation learning for treatment effect estimation from observational data. In *NeurIPS*. 2638–2648.
- [59] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2019. ACE: Adaptively Similarity-Preserved Representation Learning for Individual Treatment Effect Estimation. In *ICDM*. 1432–1437.
- [60] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. In *ICLR*.
- [61] Wenhao Zhang, Wentian Bao, Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, and Ramin Ramezani. 2020. Large-scale Causal Approaches to Debiasing Post-click Conversion Rate Estimation with Multi-task Learning. In *WWW*. 2775–2781.
- [62] Yang Zhang, Dong Wang, Qiang Li, Yue Shen, Ziqi Liu, Xiaodong Zeng, Zhiqiang Zhang, Jinjie Gu, and Derek F Wong. 2021. User Retention: A Causal Approach with Triple Task Modeling. In *IJCAI*.
- [63] Kailiang Zhong, Fengtong Xiao, Yan Ren, Yaorong Liang, Wenqing Yao, Xiaofeng Yang, and Ling Cen. 2022. Descn: Deep entire space cross networks for individual treatment effect estimation. In *SIGKDD*. 4612–4620.
- [64] Ziwei Zhu, Yun He, Yin Zhang, and James Caverlee. 2020. Unbiased implicit recommendation and propensity estimation via combinatorial joint learning. In *RecSys*. 551–556.

A Theoretical Justification

A.1 Notations and preliminaries on HTE estimation

Here, we formalize the definitions, assumptions, and pertinent lemmas in the domain of HTE estimation from observational data. Building on the notations introduced in Section 2.1, consider an individual with covariates x exhibiting two potential outcomes: $Y_1(x)$ if treated and $Y_0(x)$ otherwise; CATE is defined as the difference between these outcomes, interpreted as $\tau(x) := \mathbb{E}[Y_1 - Y_0 | x]$.

Definition A.1. Let $\psi : \mathcal{S}_X \rightarrow \mathcal{S}_R$ denote a representation mapping that transforms covariates X into a representation space $R = \psi(X)$. Define $\phi_T : \mathcal{R} \times \mathcal{T} \rightarrow \mathcal{Y}$ as an outcome mapping that correlates these representations and treatment states to their respective factual outcomes, with $Y_1 = \phi_1(R)$ and $Y_0 = \phi_0(R)$.

Definition A.2. The expected loss for the units with covariates x and treatment indicator t is: $l_{\psi, \phi}(x, t) := \int (Y_t - \phi_t(\psi(x)))^2 \cdot \rho(Y_t | x) dY_t$. Then, the expected factual outcome estimation error for treated, untreated and all units are:

$$\epsilon_F^{T=1}(\psi, \phi) := \int l_{\psi, \phi}(x, 1) \cdot \rho^{T=1}(x) dx, \quad \epsilon_F^{T=0}(\psi, \phi) := \int l_{\psi, \phi}(x, 0) \cdot \rho^{T=0}(x) dx, \quad \epsilon_F(\psi, \phi) := \int l_{\psi, \phi}(x, t) \cdot \rho(x, t) dx dt. \quad (10)$$

An exemplar approach to CATE estimation is TARNet [38]. Based on the elements in A.1, it involves a representation mapping ψ that is shared across the treated and untreated units, and outcome mappings ϕ_1 and ϕ_0 for treated and untreated units, respectively. It estimates the CATE as $\hat{\tau}_{\psi, \phi}(X) := \hat{Y}_1 - \hat{Y}_0$, where $\hat{Y}_1 = \phi_1(\psi(X))$, $\hat{Y}_0 = \phi_0(\psi(X))$. The quality of CATE estimation is evaluated via the PEHE metric

$$\epsilon_{\text{PEHE}}(\psi, \phi) := \int \left(\hat{\tau}_{\psi, \phi}(x) - \tau(x) \right)^2 \rho(x) dx. \quad (11)$$

During training, the factual error $\epsilon_F(\phi, \psi)$ in Definition A.2 is optimized. However, treatment selection bias results in covariate distribution differences between treated and untreated groups, impeding model generalization across these groups. For example, an estimator ϕ_1 trained solely on treated units may yield biased estimates of $\hat{\tau}$ when applied to untreated units, as shown in Figure 1(a). To mitigate this bias, representation-based methods [16, 38] advocate for minimizing distribution discrepancies in the representation space and construct generalization bounds (see Theorem 2). However, the IPM term in Theorem 2 is intractable for complex distributions. A common approach is to re-express IPM as the Wasserstein distance, as detailed in Lemma 1.

Definition A.3. Let $\rho^{T=1}(x)$ and $\rho^{T=0}(x)$ denote the covariate distributions for the treated and untreated groups, respectively. Define $\rho_{\psi}^{T=1}(r)$ and $\rho_{\psi}^{T=0}(r)$ as the distributions of the representations $r = \psi(x)$, where ψ is the representation mapping detailed in Definition 2.2.

Definition A.4. Given two distribution functions $\rho^{T=1}(x)$ and $\rho^{T=0}(x)$ supported over \mathcal{X} , and a sufficiently large function family \mathcal{F} , the Integral Probability Metric (IPM) induced by \mathcal{F} is defined as: $\text{IPM}_{\mathcal{F}}(\rho^{T=1}, \rho^{T=0}) = \sup_{f \in \mathcal{F}} \left| \int f(x) (\rho^{T=1}(x) - \rho^{T=0}(x)) dx \right|$.

THEOREM 2. Suppose \mathcal{F} is a function family sufficiently large to include $\frac{1}{B_{\psi}} \cdot l_{\psi, \phi}(x, t)$ for $t \in \{0, 1\}$, Shalit et al. [38] demonstrate that:

$$\epsilon_{\text{PEHE}}(\psi, \phi) \leq 2 \left(\epsilon_F^{T=0}(\psi, \phi) + \epsilon_F^{T=1}(\psi, \phi) + B_{\psi} \text{IPM}_{\mathcal{F}}(\rho_{\psi}^{T=1}, \rho_{\psi}^{T=0}) - 2\sigma_Y^2 \right), \quad (12)$$

where $\epsilon_F^{T=0}$ and $\epsilon_F^{T=1}$ are defined according to Definition A.2, and $\rho_{\psi}^{T=1}(r)$ and $\rho_{\psi}^{T=0}(r)$ are specified in Definition A.3.

LEMMA 1. Given two distribution functions $\rho_1(x)$ and $\rho_2(x)$ supported over \mathcal{X} , and letting \mathcal{F} be the family of 1-Lipschitz functions, we have $\text{IPM}_{\mathcal{F}}(\rho_1, \rho_2) = \mathbb{W}(\rho_1, \rho_2)$, i.e., the IPM induced by \mathcal{F} is equivalent to the Wasserstein distance \mathbb{W} [42].

A.2 Theoretical results

Theorem 2 assumes access to the entire populations of treated and untreated groups to calculate the distribution discrepancy. However, in training neural networks, parameters are typically updated using stochastic gradient methods on mini-batches rather than the full dataset. This raises concerns about the validity of Theorem 2 when applied at the mini-batch level. Recent studies have investigated the sample complexity of various discrepancy measures, such as the Wasserstein distance (see Lemma 2) and Gromov discrepancy (see Lemma 3). Building on these insights, we propose Theorem 3, which extends Theorem 2 to the specific Fused Gromov-Wasserstein (FGW) discrepancy used in this work. This theorem examines the sample complexity of FGW when only a small mini-batch sample is available.

LEMMA 2. Consider two measures α and β with compact supports $\mathcal{S}_{\alpha} \in \mathbb{R}^d$ and $\mathcal{S}_{\beta} \in \mathbb{R}^d$. Let $C = \text{diam}(\mathcal{S}_{\alpha}) \vee \text{diam}(\mathcal{S}_{\beta})$, considering the case where $d > 4$, we have:

$$\mathbb{E} \left[\left| \mathbb{W}(\alpha, \beta)^2 - \mathbb{W}(\hat{\alpha}_n, \hat{\beta}_n)^2 \right| \right] \lesssim n^{-\frac{2}{d}}, \quad (13)$$

where the notation \lesssim hides constants that is independent to the number of samples n . α_n and β_n are empirical distributions of α and β with n i.i.d. samples.

LEMMA 3. Consider two measures α and β with compact supports $\mathcal{S}_\alpha \in \mathbb{R}^d$ and $\mathcal{S}_\beta \in \mathbb{R}^d$. Let $C = \text{diam}(\mathcal{S}_\alpha) \vee \text{diam}(\mathcal{S}_\beta)$, considering the case where $d > 4$, we have:

$$\mathbb{E} \left[\left| \mathbb{G}(\alpha, \beta)^2 - \mathbb{G}(\hat{\alpha}_n, \hat{\beta}_n)^2 \right| \right] \lesssim \frac{C^4}{\sqrt{n}} + (1 + C^4) n^{-\frac{2}{d}}, \quad (14)$$

where the notation \lesssim hides constants that is independent to the number of samples n . α_n and β_n are empirical distributions of α and β with n i.i.d. samples.

THEOREM 3. Let ψ and ϕ be the representation mapping and factual outcome mapping, respectively; \mathbb{W}_ψ be the group discrepancy at a mini-batch level. With the probability of at least $1 - \delta$, we have:

$$\epsilon_{\text{PEHE}}(\psi, \phi) \leq 2[\epsilon_{\mathbb{F}}^{T=1}(\psi, \phi) + \epsilon_{\mathbb{F}}^{T=0}(\psi, \phi) + B_{\psi, \kappa} \mathbb{F}(\mathcal{R}_\psi^{T=0}, \mathcal{R}_\psi^{T=1}) - 2\sigma_Y^2 + \mathcal{O}(N^{-\frac{2}{d}})], \quad (15)$$

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

PROOF. According to Theorem 2 we have:

$$\epsilon_{\text{PEHE}}(\psi, \phi) \leq 2 \left(\epsilon_{\mathbb{F}}^{T=0}(\psi, \phi) + \epsilon_{\mathbb{F}}^{T=1}(\psi, \phi) + B_{\psi} \text{IPM}_{\mathcal{F}}(\rho_{\psi}^{T=1}, \rho_{\psi}^{T=0}) - 2\sigma_Y^2 \right). \quad (16)$$

Assuming that there exists a constant $B_{\psi} > 0$, such that for $t \in \{0, 1\}$, $\frac{1}{B_{\psi}} \cdot l_{\psi, \phi}(x, t)$ belongs to the family of 1-Lipschitz functions. According to Lemma 1, we have

$$\epsilon_{\text{PEHE}}(\psi, \phi) \leq 2 \left(\epsilon_{\mathbb{F}}^{T=0}(\psi, \phi) + \epsilon_{\mathbb{F}}^{T=1}(\psi, \phi) + B_{\psi} \mathbb{W}(\rho_{\psi}^{T=1}, \rho_{\psi}^{T=0}) - 2\sigma_Y^2 \right). \quad (17)$$

Following Definition 2.3, let $\mathcal{R}_\psi^{T=1}$ and $\mathcal{R}_\psi^{T=0}$ be the empirical distributions of representations at a mini-batch level, both containing n units. Then, according to Lemma 2 and 3, we have:

$$\begin{aligned} \mathbb{W}(\rho_{\psi}^{T=1}, \rho_{\psi}^{T=0}) &\leq \frac{1}{\kappa} \left(\kappa * \mathbb{W}(\rho_{\psi}^{T=1}, \rho_{\psi}^{T=0}) + (1 - \kappa) * \mathbb{G}(\rho_{\psi}^{T=1}, \rho_{\psi}^{T=0}) \right) \\ &\leq \frac{1}{\kappa} \left(\kappa * \mathbb{W}(\mathcal{R}_\psi^{T=1}, \mathcal{R}_\psi^{T=0}) + (1 - \kappa) * \mathbb{G}(\mathcal{R}_\psi^{T=1}, \mathcal{R}_\psi^{T=0}) + \frac{C^4}{\sqrt{n}} + (2 + C^4) n^{-\frac{2}{d}} \right) \\ &\leq \frac{1}{\kappa} \left(\mathbb{F}(\mathcal{R}_\psi^{T=1}, \mathcal{R}_\psi^{T=0}) + \frac{C^4}{\sqrt{n}} + (2 + C^4) n^{-\frac{2}{d}} \right), \end{aligned} \quad (18)$$

where \mathbb{W} denotes the Wasserstein discrepancy, \mathbb{G} denotes the Gromov discrepancy, \mathbb{F} denotes the fused Wasserstein discrepancy. These discrepancies will be introduced in the next section. Notably, there are two terms in the cost function of \mathbb{F} as per (5), which corresponding to the cost functions of \mathbb{W} and \mathbb{G} , respectively. Therefore, \mathbb{W} and \mathbb{G} can be viewed as minimizing the two terms of the cost function of \mathbb{F} individually, which often yields smaller values.

Denote $B_{\psi, \kappa} = B_{\psi} / \kappa$. Combing (18) and (17), we have

$$\epsilon_{\text{PEHE}}(\psi, \phi) \leq 2[\epsilon_{\mathbb{F}}^{T=1}(\psi, \phi) + \epsilon_{\mathbb{F}}^{T=0}(\psi, \phi) + B_{\psi, \kappa} \mathbb{F}(\mathcal{R}_\psi^{T=0}, \mathcal{R}_\psi^{T=1}) - 2\sigma_Y^2 + \mathcal{O}(n^{-\frac{2}{d}})], \quad (19)$$

where $\mathcal{O}(n^{-\frac{2}{d}}) = \frac{1}{\kappa} \left(\frac{C^4}{\sqrt{n}} + (2 + C^4) n^{-\frac{2}{d}} \right)$. The proof is completed. \square