DIGNet: Learning Decomposed Patterns in Representation Balancing for Treatment Effect Estimation

Anonymous authors
Paper under double-blind review

Abstract

Estimating treatment effects from observational data is often subject to a covariate shift problem incurred by selection bias. Recent research has sought to mitigate this problem by leveraging representation balancing methods that aim to extract balancing patterns from observational data and utilize them for outcome prediction. The underlying theoretical rationale is that minimizing the unobserved counterfactual error can be achieved through two principles: (I) reducing the risk associated with predicting factual outcomes and (II) mitigating the distributional discrepancy between the treated and controlled samples. However, an inherent trade-off between the two principles can lead to a potential over-balancing issue, resulting in the loss of valuable information for factual outcome predictions and, consequently, deteriorating treatment effect estimations. To overcome this challenge, we propose a novel representation balancing model, DIGNet, for treatment effect estimation. DIGNet incorporates two key components, PDIG and PPBR, which effectively mitigate the trade-off problem by improving one aforementioned principle without sacrificing the other. Specifically, PDIG captures more effective balancing patterns (Principle II) without affecting factual outcome predictions (Principle I), while PPBR enhances factual outcome prediction (Principle I) without affecting the learning of balancing patterns (Principle II). Our comprehensive ablation studies confirm the effectiveness of PDIG and PPBR in improving treatment effect estimation, and experimental results on benchmark datasets demonstrate the superior performance of our DIGNet model compared to baseline models.

1 Introduction

In the context of the ubiquity of personalized decision-making, causal inference has sparked a surge of research exploring causal machine learning in many disciplines, including economics and statistics (Wager & Athey, 2018; Athey & Wager, 2019; Farrell, 2015; Chernozhukov et al., 2018; Huang et al., 2021), healthcare (Qian et al., 2021; Bica et al., 2021a;b), and commercial applications (Guo et al., 2020b;c; Chu et al., 2021). The core of causal inference is to estimate *treatment effects*, which is closely related to the *factual outcomes* (observed outcomes) and *counterfactual outcomes*. The concept of the counterfactual outcome is closely linked to a fundamental hypothetical question: What would the outcome be if an alternative treatment were received? Answering this question is challenging because counterfactual outcomes are unobservable in reality, making it impossible to directly access ground-truth treatment effects from observational data. Consequently, an increasing amount of recent research has focused on developing innovative machine learning models that aim to enhance the estimation of counterfactual outcomes to obtain more accurate treatment effect estimates.

The major challenge of estimating counterfactual outcomes lies in the *covariate shift* problem incurred by *selection bias* inherent in observational data (Guo et al., 2020a; Zhang et al., 2020; Yao et al., 2021). Selection bias refers to the non-random treatment assignment, where the decision of whether to receive treatment (such as whether to administer vaccination) is typically influenced by covariates (such as age) that also impact the outcome (such as infection rate) (Huang et al., 2022b). Usually, individuals who received treatment are referred to as *treated samples* or *treatment samples*, while those who did not receive treatment are referred to as *controlled samples* or *control samples*. The probability of one receiving treatment is commonly referred to as the *propensity score*, and the differences in propensity scores within the population can naturally give rise to the covariate shift problem, i.e., the distribution of covariates in the treated group significantly differs from that in the control group. This covariate shift issue compounds the difficulty in estimating counterfactual outcomes from observational data (Yao et al., 2018; Hassanpour & Greiner, 2019a).

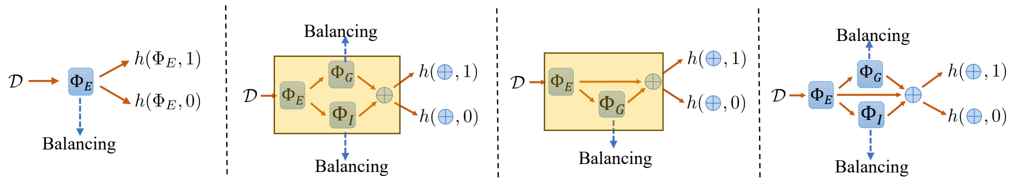
To alleviate the covariate shift problem, recent advancements in representation balancing research have explored the representation learning model, such as CounterFactual Regression Network (CFRNet) (Shalit et al., 2017), to estimate individual treatment effects (ITEs). These representation balancing models aim to extract balancing patterns from observational data and utilize these patterns to predict outcomes. The corresponding objective function is typically concerned with minimizing the empirical risk of factual outcomes while concurrently minimizing the distributional distance between the treatment and control groups in the representation space (Shalit et al., 2017; Johansson et al., 2022). The underlying theoretical logic behind these studies is that minimizing counterfactual error can be achieved by two principles in the representation space: ***(Principle I) minimizing the risk associated with predicting factual outcomes***, and ***(Principle II) reducing the distributional discrepancy between the treated and controlled samples***. The theoretical foundation and the classic CFRNet structure proposed in Shalit et al. (2017) have inspired many subsequent studies on representation balancing methods for treatment effect estimation, including Yao et al. (2018); Shi et al. (2019); Zhang et al. (2020); Hassanpour & Greiner (2019a); Assaad et al. (2021); Huang et al. (2022a).

While the representation balancing framework provides a powerful tool to tackle the covariate shift issue, models that rely on the classic CFRNet structure still encounters a critical hurdle: *Enforcing models to learn merely balancing patterns can undermine the predictive power of the outcome function.* This problem is referred to as the *over-balancing issue* due to the trade-off between Principle I and Principle II, and it may inadvertently harm the treatment effect estimation (Zhang et al., 2020; Assaad et al., 2021; Huang et al., 2022a). To better understand this phenomenon, we present a motivating example below.

Motivating Example. Consider two individuals who are identical in every aspect except for their age. One person is older and is designated as the treatment (T) group, while the other person is younger and serves as the control (C) group. Age is used as a covariate to distinguish between T and C. If it is known that the older person is more susceptible to a certain disease, the age information (covariate) can be used to predict the likelihood of one developing the disease (outcome). However, suppose the age information of each individual is mapped to some representations such that the representations of T and C are highly-balanced or even identical. In that case, it may be difficult to differentiate between T and C based on these representations. Consequently, these over-balanced representations may lose information to accurately predict the likelihood of each individual developing the disease.

Classic representation balancing models (Figure 1(a)) may encounter a potential over-balancing issue due to the inherent trade-off between Principle I and Principle II, because the learned balancing patterns Φ_E also serve as the outcome predictors. Once the learned patterns are over-balancing, leading to the loss of outcome-related information, the treatment effect estimation tends to deteriorate. This motivates us to ponder an important question: considering the inherent trade-off between the two principles, ***is it possible to explore a scheme that enhances one principle without compromising the other?*** More specifically, can we explore improving treatment effect estimation through the following two paths: ***(Path I) learning more balanced patterns without affecting factual outcome prediction*** and ***(Path II) enhancing factual outcome prediction without affecting the learning of balancing patterns***?

In this paper, we propose a novel representation balancing model, **DIGNet** (Section 4.2.2), which is a neural **Net**work that incorporates ***D****ecomposed patterns* with ***I****ndividual propensity confusion* and ***G****roup distance minimization.* The term of decomposed patterns denotes distinct components disentangled from some specific representations in DIGNet (Section 4.2). The individual propensity confusion aspect of DIGNet aims to learn representations that are difficult to utilize for characterizing the propensity of each individual being treated or controlled (Section 4.1.2), and the corresponding theoretical foundation is based on our derived $\mathcal{H}$-divergence guided counterfactual and ITE error bounds (Section 3.2). The group distance minimization aspect of DIGNet focuses on learning representations that minimize the distance between the treated and controlled groups (Section 4.1.1), and the corresponding theoretical foundation is supported by previous



(a) Classic model (b) PDIG (yellow) (c) PPBR (yellow) (d) DIGNet

Figure 1: (a): The classic model (e.g., GNet in Section 4.1.1 and INet in Section 4.1.2) maps the original data $\mathcal{D}$ into representations Φ_E to achieve representation balancing. The balanced representations are referred to as *balancing patterns*. These balancing patterns are also used for outcome prediction. (b): The PDIG (Section 4.2.1) is illustrated as the yellow part, where balancing patterns are decomposed into two distinct components, Φ_G and Φ_I. Φ_G serves for *group distance minimization* (Section 4.1.1) and Φ_I serves for *individual propensity confusion* (Section 4.1.2). The balancing patterns Φ_G and Φ_I are concatenated for predicting outcomes. (c): The PPBR (Section 4.2.1) is represented by the yellow section, where Φ_E is used for feature extraction and Φ_G is used for representation balancing. Here representations are decomposed into *pre-balancing patterns* Φ_E and balancing patterns Φ_G. Φ_E and Φ_G are concatenated for predicting outcomes. (d): The proposed model DIGNet (Section 4.2.2) integrates both PDIG and PPBR. Specifically, DIGNet decomposes balancing patterns into two distinct components, Φ_G and Φ_I. The outcome predictors are further formed by concatenating Φ_G, Φ_I, and pre-balancing patterns Φ_E.

work (Shalit et al., 2017) on Wasserstein distance guided counterfactual and ITE error bounds (Section 3.1). To illustrate and explain these introduced concepts, we provide Figure 1 which visually depicts the proposed components and their relationships.

Contributions. Our main contributions are summarized as follows:

1. We derive theoretical upper bounds for counterfactual error and ITE error based on $\mathcal{H}$-divergence (Section 3.2). In particular, this theoretical foundation highlights the important role of propensity score for representation balancing models, connecting the representation balancing with the concept of individual propensity confusion.

2. We suggest learning decomposed patterns in representation balancing models (Section 4.2.1). First, we propose a ***PDIG*** method (Figure 1(b)), which aims to learn **P**atterns **D**ecomposed with **I**ndividual propensity confusion and **G**roup distance minimization to improve treatment effect estimation through Path I. Second, we propose a ***PPBR*** method (Figure 1(c)), which aims to learn **P**atterns of **P**re-balancing and **B**alancing **R**epresentations to improve treatment effect estimation through Path II.

3. Building upon PDIG and PPBR, we propose a novel representation balancing model, DIGNet (Figure 1(d)), for treatment effect estimation. In Section 5, ablation studies verify the efficacy of PDIG and PPBR in improving ITE estimation through Path I and Path II, respectively. Furthermore, experimental results on benchmark datasets demonstrate that DIGNet surpasses the performance of baseline models in terms of treatment effect estimation.

1.1 Related Work

The presence of a covariate shift problem stimulates the line of representation balancing works (Johansson et al., 2016; Shalit et al., 2017; Johansson et al., 2022). These works aim to balance the distributions of representations between treated and controlled groups and simultaneously try to maintain representations predictive of factual outcomes. This idea is closely connected with domain adaptation. In particular, the ITE error bound based on Wasserstein distance is similar to the generalization bound in Ben-David et al.

(2010); Long et al. (2014); Shen et al. (2018). In addition to Wasserstein distance based model, this paper derives a new ITE error bound based on $\mathcal{H}$-divergence (Ben-David et al., 2006; 2010; Ganin et al., 2016).

Another recent line of causal representation learning literature investigates efficient neural network structures for treatment effect estimation. Kuang et al. (2017); Hassanpour & Greiner (2019b) extract the original covariates into treatment-specific factors, outcome-specific factors, and confounding factors; X-learner (Künzel et al., 2019) and R-learner (Nie & Wager, 2021) are developed beyond the classic S-learner and T-learner; Curth & van der Schaar (2021) leverage structures for end-to-end learners to counteract the inductive bias towards treatment effect estimation, which is motivated by Makar et al. (2020).

Our DIGNet model incoporates the PDIG and PPBR methods. The PDIG method is motivated by multi-task learning, where we design a framework incorporating two specific balancing patterns that share the same pre-balancing representations. The PPBR approach is motivated the over-balancing problem (Zhang et al., 2020; Assaad et al., 2021; Huang et al., 2022a), where the researchers argue that improperly balanced representations can be detrimental predictors for outcome modeling, since such representations can lose the original information that contributes to outcome prediction. Other representation learning methods relevant to treatment effect estimation include Louizos et al. (2017); Yao et al. (2018); Yoon et al. (2018); Shi et al. (2019); Du et al. (2021).

2 Preliminaries

Notations. Suppose there are the N i.i.d. random variables $\mathcal{D} = \{(\mathbf{X}_i, T_i, Y_i)\}_{i=1}^N$ with observed realizations $\{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^N$, where there are N_1 treated units and N_0 controlled units. For each unit i, $\mathbf{X}_i \in \mathcal{X} \subset \mathbb{R}^d$ denotes d-dimensional covariates and $T_i \in \{0,1\}$ denotes the binary treatment, with $e(\mathbf{x}_i) := p(T_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i)$ defined as the propensity score (Rosenbaum & Rubin, 1983). Potential outcome framework (Rubin, 2005) defines the potential outcomes $Y^1, Y^0 \in \mathcal{Y} \subset \mathbb{R}$ for treatment $T = 1$ and $T = 0$, respectively. We let the observed outcome (factual outcome) be $Y = T \cdot Y^1 + (1-T) \cdot Y^0$, and the unobserved outcome (counterfactual outcome) be $Y = T \cdot Y^0 + (1-T) \cdot Y^1$. For $t \in \{0,1\}$, let $\tau^t(\mathbf{x}) := \mathbb{E}[Y^t \mid \mathbf{X} = \mathbf{x}]$ be a function of Y^t w.r.t. $\mathbf{X}$, then our goal is to estimate the individual treatment effect (ITE) $\tau(\mathbf{x}) := \mathbb{E}\left[Y^1 - Y^0 \mid \mathbf{X} = \mathbf{x}\right] = \tau^1(\mathbf{x}) - \tau^0(\mathbf{x})$ [1], and the average treatment effect (ATE) $\tau_{ATE} := \mathbb{E}\left[Y^1 - Y^0\right] = \int_{\mathcal{X}} \tau(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$. The introduced concepts PPBR and PDIG are illustrated in Figure 1, and the necessary representation functions Φ_E, Φ_G and Φ_I, as well as different model structures, are illustrated in Figure 2. Throughout the paper, we refer to patterns as meaningful representations. For instance, decomposed patterns are distinct components disentangled from some specific representations.

2.1 Problem setup

In causal representation balancing works, we denote representation space by $\mathcal{R} \subset \mathbb{R}^d$, and $\Phi : \mathcal{X} \to \mathcal{R}$ is assumed to be a twice-differentiable, one-to-one and invertible function with its inverse $\Psi : \mathcal{R} \to \mathcal{X}$ such that $\Psi(\Phi(\mathbf{x})) = \mathbf{x}$. The densities of the treated and controlled covariates are denoted by $p_{\mathbf{x}}^{T=1} = p^{T=1}(\mathbf{x}) := p(\mathbf{x} \mid T = 1)$ and $p_{\mathbf{x}}^{T=0} = p^{T=0}(\mathbf{x}) := p(\mathbf{x} \mid T = 0)$, respectively. Correspondingly, the densities of the treated and controlled covariates in the representation space are denoted by $p_{\Phi}^{T=1} = p_{\Phi}^{T=1}(\mathbf{r}) := p_{\Phi}(\mathbf{r} \mid T = 1)$ and $p_{\Phi}^{T=0} = p_{\Phi}^{T=0}(\mathbf{r}) := p_{\Phi}(\mathbf{r} \mid T = 0)$, respectively.

Our study is based on the potential outcome framework (Rubin, 2005). Assumption 1 states standard and necessary assumptions to ensure treatment effects are identifiable. Before proceeding with theoretical analysis, we also present some necessary terms and definitions in Definition 1.

Assumption 1 (Consistency, Overlap, and Unconfoundedness)**.** *Consistency: If the treatment is t, then the observed outcome equals Y^t. Overlap: The propensity score is bounded away from 0 to 1, i.e., $0 < e(\mathbf{x}) < 1$. Unconfoundedness: $Y^t \perp\!\!\!\perp T \mid \mathbf{X}$, $\forall t \in \{0,1\}$.*

---

[1] The term $\mathbb{E}\left[Y^1 - Y^0 \mid \mathbf{X} = \mathbf{x}\right]$ is commonly known as the Conditional Average Treatment Effect (CATE). In order to maintain consistency with the notion used in the existing causal representation balancing literature, e.g., Shalit et al. (2017), we refer to this term as ITE throughout this paper. Note that the original definition of ITE for the i-th individual is commonly expressed as the difference between their potential outcomes, represented as $Y_i^1 - Y_i^0$.

Definition 1. *Let* $h : \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ *be an hypothesis defined over the representation space* $\mathcal{R}$ *such that* $h(\Phi(\mathbf{x}), t)$ *estimates* y^t*, and* $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ *be a loss function (e.g., the squared loss* $L(y, y') = (y - y')^2$ *or the absolute loss* $L(y, y') = |y - y'|$*). If we define the expected loss for* $(\mathbf{x}, t)$ *as* $\ell_{h,\Phi}(\mathbf{x}, t) = \int_{\mathcal{Y}} L(y^t, h(\Phi(\mathbf{x}), t))p(y^t|\mathbf{x})dy^t$*, we then have factual and counterfactual errors, as well as them on the treated and controlled:*

$$\epsilon_F(h, \Phi) = \int_{\mathcal{X}\times\{0,1\}} \ell_{h,\Phi}(\mathbf{x}, t)p(\mathbf{x}, t)d\mathbf{x}dt, \qquad \epsilon_{CF}(h, \Phi) = \int_{\mathcal{X}\times\{0,1\}} \ell_{h,\Phi}(\mathbf{x}, t)p(\mathbf{x}, 1-t)d\mathbf{x}dt,$$

$$\epsilon_F^{T=1}(h, \Phi) = \int_{\mathcal{X}} \ell_{h,\Phi}(\mathbf{x}, 1)p^{T=1}(\mathbf{x})d\mathbf{x}, \qquad \epsilon_F^{T=0}(h, \Phi) = \int_{\mathcal{X}} \ell_{h,\Phi}(\mathbf{x}, 0)p^{T=0}(\mathbf{x})d\mathbf{x},$$

$$\epsilon_{CF}^{T=1}(h, \Phi) = \int_{\mathcal{X}} \ell_{h,\Phi}(\mathbf{x}, 1)p^{T=0}(\mathbf{x})d\mathbf{x}, \qquad \epsilon_{CF}^{T=0}(h, \Phi) = \int_{\mathcal{X}} \ell_{h,\Phi}(\mathbf{x}, 0)p^{T=1}(\mathbf{x})d\mathbf{x}.$$

If we let $f(\mathbf{x}, t)$ be $h(\Phi(\mathbf{x}), t)$, where $f : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$ is a prediction function for outcome, then the estimated ITE over f is defined as $\hat{\tau}_f(\mathbf{x}) := f(\mathbf{x}, 1) - f(\mathbf{x}, 0)$. We can measure the error in ITE estimation with the metric, Precision in the expected Estimation of Heterogeneous Effect (PEHE):

$$\epsilon_{PEHE}(f) = \int_{\mathcal{X}} L(\hat{\tau}_f(\mathbf{x}), \tau(\mathbf{x}))p(\mathbf{x})d\mathbf{x}. \tag{1}$$

Here, $\epsilon_{PEHE}(f)$ can also be denoted by $\epsilon_{PEHE}(h, \Phi)$ if we let $f(\mathbf{x}, t)$ be $h(\Phi(\mathbf{x}), t)$.

3 Theoretical Results

In this section, we first prove ϵ_{PEHE} is bounded by ϵ_F and ϵ_{CF} in Lemma 1. Next, we revisit the upper bound for Wasserstein distance guided representation balancing method in Section 3.1. Furthermore, we state the new theoretical results concerning $\mathcal{H}$-divergence guided representation balancing method in Section 3.2.

Lemma 1. *Let functions* h *and* Φ *be as defined in Definition 1. Recall that* $\tau^t(\mathbf{x}) = \mathbb{E}[Y^t \mid \mathbf{X} = \mathbf{x}]$*. Define* $\sigma_y^2 = \min\{\sigma_{y^t}^2(p(\mathbf{x}, t)), \sigma_{y^t}^2(p(\mathbf{x}, 1-t))\}$ *and* $A_y = \max\{A_{y^t}(p(\mathbf{x}, t)), A_{y^t}(p(\mathbf{x}, 1-t))\}$ $\forall t \in \{0,1\}$*, where* $\sigma_{y^t}^2(p(\mathbf{x}, t)) = \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (y^t - \tau^t(\mathbf{x}))^2 p(y^t|\mathbf{x})p(\mathbf{x}, t)dy^t d\mathbf{x}dt$ *and* $A_{y^t}(p(\mathbf{x}, t)) = \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} |y^t - \tau^t(\mathbf{x})|p(y^t|\mathbf{x})p(\mathbf{x}, t)dy^t d\mathbf{x}dt$ $\forall t \in \{0,1\}$*.*
Let loss function L *be the squared loss. Then we have:*

$$\epsilon_{PEHE}(h, \Phi) \le 2(\epsilon_{CF}(h, \Phi) + \epsilon_F(h, \Phi) - 2\sigma_y^2). \tag{2}$$

Let loss function L *be the absolute loss. Then we have:*

$$\epsilon_{PEHE}(h, \Phi) \le \epsilon_{CF}(h, \Phi) + \epsilon_F(h, \Phi) + 2A_y. \tag{3}$$

Lemma 1 reveals that the ITE error ϵ_{PEHE} is closely connected with the factual error ϵ_F and counterfactual ϵ_{CF}. The proof of Lemma 1 is deferred to Section A.1. Note that equation (2) corresponds to the result presented in Shalit et al. (2017), while equation (3) is our new result, which supplements the case when L denotes the absolute loss.

3.1 Wasserstein Distance Guided Error Bounds

Previous causal learning models commonly adopt the Wasserstein distance guided approach to seek representation balancing. In this subsection, we first give the definition of Wasserstein distance (Cuturi & Doucet, 2014) by introducing the Integral Probability Metric (IPM) (Sriperumbudur et al., 2012) defined in Definition 2. Then we state the theorem regarding the upper bounds for counterfactual error ϵ_{CF} and ITE error ϵ_{PEHE} using Wasserstein distance in Theorem 1.

Definition 2. *Let* $\mathcal{G}$ *be a function family consisting of functions* $g : \mathcal{S} \to \mathbb{R}$*. For a pair of distributions* p_1*,* p_2 *over* $\mathcal{S}$*, the Integral Probability Metric is defined as*

$$IPM_{\mathcal{G}}(p_1, p_2) := \sup_{g\in\mathcal{G}} |\int_{\mathcal{S}} g(s)(p_1(s) - p_2(s))ds|.$$

If $\mathcal{G}$ is the family of 1-Lipschitz functions, we can obtain the so-called 1-Wasserstein distance, denoted by $Wass(p_1, p_2)$. Next, we present the bounds for counterfactual error ϵ_{CF} and ITE error ϵ_{PEHE} using Wasserstein distance in Theorem 1.

Theorem 1. *Let $\Phi : \mathcal{X} \to \mathcal{R}$ be an invertible representation with Ψ being its inverse. Define $\sigma_y^2 = \min\{\sigma_{y^t}^2(p(\mathbf{x}, t)), \sigma_{y^t}^2(p(\mathbf{x}, 1-t))\}$ and $A_y = \max\{A_{y^t}(p(\mathbf{x}, t)), A_{y^t}(p(\mathbf{x}, 1-t))\}$ $\forall t \in \{0,1\}$, where $\sigma_{y^t}^2(p(\mathbf{x},t)) = \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (y^t - \tau^t(\mathbf{x}))^2 p(y^t|\mathbf{x}) p(\mathbf{x},t) dy^t d\mathbf{x} dt$ and $A_{y^t}(p(\mathbf{x},t)) = \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} |y^t - \tau^t(\mathbf{x})| p(y^t|\mathbf{x}) p(\mathbf{x},t) dy^t d\mathbf{x} dt$ $\forall t \in \{0,1\}$. Let $p_\Phi^{T=1}(\mathbf{r})$, $p_\Phi^{T=0}(\mathbf{r})$ be as defined before, $h : \mathcal{R} \times \{0,1\} \to \mathcal{Y}$, $u := Pr(T=1)$ and $\mathcal{G}$ be the family of 1-Lipschitz functions. Assume there exists a constant $B_\Phi \geq 0$, such that for $t \in \{0,1\}$, the function $g_{\Phi,h}(\mathbf{r},t) := \frac{1}{B_\Phi} \cdot \ell_{h,\Phi}(\Psi(\mathbf{r}), t) \in \mathcal{G}$. Given a loss function L, we have*

$$\epsilon_{CF}(h,\Phi) \leq (1-u)\cdot \epsilon_F^{T=1}(h,\Phi) + u \cdot \epsilon_F^{T=0}(h,\Phi) + B_\Phi \cdot Wass(p_\Phi^{T=1}, p_\Phi^{T=0}). \tag{4}$$

Let loss function L be the squared loss. Then we have:

$$\epsilon_{PEHE}(h,\Phi) \leq 2(\epsilon_F^{T=1}(h,\Phi) + \epsilon_F^{T=0}(h,\Phi) + B_\Phi \cdot Wass(p_\Phi^{T=1}, p_\Phi^{T=0}) - 2\sigma_y^2). \tag{5}$$

Let loss function L be the absolute loss. Then we have:

$$\epsilon_{PEHE}(h,\Phi) \leq \epsilon_F^{T=1}(h,\Phi) + \epsilon_F^{T=0}(h,\Phi) + B_\Phi \cdot Wass(p_\Phi^{T=1}, p_\Phi^{T=0}) + 2A_y. \tag{6}$$

Theorem 1 reveals that the ITE error is closely tied to the factual error ϵ_F and the Wasserstein distance between treated and controlled groups in the representation space. This theorem provides a theoretical foundation for representation balancing models based on group distance minimization (Section 4.1.1). The proof of Theorem 1 is deferred to Section A.2. Note that equation (5) corresponds to the result presented in Shalit et al. (2017), while equation (6) is our new result, which supplements the case when L denotes the absolute loss.

3.2 $\mathcal{H}$-divergence Guided Error Bounds

In most representation balancing literature, the models mainly rely on Wasserstein distance guided error bounds as discussed in Section 3.1. In this subsection, we will focus on establishing $\mathcal{H}$-divergence guided error bounds for counterfactual and ITE estimations in representation balancing approach. We first give the definition of $\mathcal{H}$-divergence (Ben-David et al., 2006) in Definition 3. Then we state the theorem regarding the upper bounds for counterfactual error ϵ_{CF} and ITE error ϵ_{PEHE} using $\mathcal{H}$-divergence in Theorem 2.

Definition 3. *Given a pair of distributions p_1, p_2 over $\mathcal{S}$, and a hypothesis binary function class $\mathcal{H}$, the $\mathcal{H}$-divergence between p_1 and p_2 is defined as*

$$d_{\mathcal{H}}(p_1,p_2) := 2 \sup_{\eta\in\mathcal{H}} \left| Pr_{p_1}[\eta(s)=1] - Pr_{p_2}[\eta(s)=1] \right|. \tag{7}$$

Theorem 2. *Let $\Phi : \mathcal{X} \to \mathcal{R}$ be an invertible representation with Ψ being its inverse. Define $\sigma_y^2 = \min\{\sigma_{y^t}^2(p(\mathbf{x}, t)), \sigma_{y^t}^2(p(\mathbf{x}, 1-t))\}$ and $A_y = \max\{A_{y^t}(p(\mathbf{x}, t)), A_{y^t}(p(\mathbf{x}, 1-t))\}$ $\forall t \in \{0,1\}$, where $\sigma_{y^t}^2(p(\mathbf{x},t)) = \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (y^t - \tau^t(\mathbf{x}))^2 p(y^t|\mathbf{x}) p(\mathbf{x},t) dy^t d\mathbf{x} dt$ and $A_{y^t}(p(\mathbf{x},t)) = \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} |y^t - \tau^t(\mathbf{x})| p(y^t|\mathbf{x}) p(\mathbf{x},t) dy^t d\mathbf{x} dt$ $\forall t \in \{0,1\}$. Let $p_\Phi^{T=1}(\mathbf{r})$, $p_\Phi^{T=0}(\mathbf{r})$ be as defined before, $h : \mathcal{R} \times \{0,1\} \to \mathcal{Y}$, $u := Pr(T=1)$ and $\mathcal{H}$ be the family of binary functions. Assume that there exists a constant $K \geq 0$ such that $\int_{\mathcal{Y}} L(y,y')dy \leq K$ $\forall y' \in \mathcal{Y}$. Given a loss function L, we have*

$$\epsilon_{CF}(h,\Phi) \leq (1-u)\cdot \epsilon_F^{T=1}(h,\Phi) + u \cdot \epsilon_F^{T=0}(h,\Phi) + \frac{K}{2} d_{\mathcal{H}}(p_\Phi^{T=1}, p_\Phi^{T=0}). \tag{8}$$

Let loss function L be the squared loss. Then we have:

$$\epsilon_{PEHE}(h,\Phi) \leq 2(\epsilon_F^{T=1}(h,\Phi) + \epsilon_F^{T=0}(h,\Phi) + \frac{K}{2} d_{\mathcal{H}}(p_\Phi^{T=1}, p_\Phi^{T=0}) - 2\sigma_y^2). \tag{9}$$

Let loss function L be the absolute loss. Then we have:

$$\epsilon_{PEHE}(h,\Phi) \leq \epsilon_F^{T=1}(h,\Phi) + \epsilon_F^{T=0}(h,\Phi) + \frac{K}{2} d_{\mathcal{H}}(p_\Phi^{T=1}, p_\Phi^{T=0}) + 2A_y. \tag{10}$$

Theorem 2 reveals that the ITE error is closely connected with the factual error ϵ_F and the $\mathcal{H}$-divergence between treated and controlled samples in the representation space. This new theoretical result provides a theoretical foundation for representation balancing models based on individual propensity confusion (Section 4.1.2). The proof of Theorem 2 is deferred to Section A.3.

4 Method

In the preceding section, we have stated the theoretical foundations for representation balancing methods, which are the Wasserstein distance guided error bounds (results in Shalit et al. (2017)) and $\mathcal{H}$-divergence guided error bounds (Our results). Moving on to Section 4.1, we will begin by introducing representation balancing methods without decomposed patterns. Specifically, Section 4.1.1 revisits a Wasserstein distance based representation balancing network GNet, and Section 4.1.2 demonstrates how Theorem 2 can be connected with individual propensity confusion, helping us to build a $\mathcal{H}$-divergence based representation balancing network INet. Subsequently, in Section 4.2, we will introduce how to design a representation balancing method within the scheme of decomposed patterns, based on the PDIG and PPBR methods (Section 4.2.1). The final proposed model DIGNet is presented in Section 4.2.2.

4.1 Representation Balancing without Decomposed Patterns

In representation balancing models, given the input data tuples $(\mathbf{x}, \mathbf{t}, \mathbf{y}) = \{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^{N}$, the original covariates $\mathbf{x}$ are extracted by some representation function $\Phi(\cdot)$, and representations $\Phi(\mathbf{x})$ are then fed into the outcome functions $h^1(\cdot) := h(\cdot, 1)$ and $h^0(\cdot) := h(\cdot, 0)$ that estimate the potential outcome y^1 and y^0, respectively. Finally, the factual outcome can be predicted by $h^t(\cdot) = th^1(\cdot) + (1-t)h^0(\cdot)$, and the corresponding outcome loss is

$$\mathcal{L}_y(\mathbf{x}, \mathbf{t}, \mathbf{y}; \Phi, h^t) = \frac{1}{N}\sum_{i=1}^{N} L(h^t(\Phi(\mathbf{x}_i)), y_i). \tag{11}$$

The loss function $\mathcal{L}_y$ approximates the factual error ϵ_F appeared in Theorems 1 and 2. Minimizing $\mathcal{L}_y$ also corresponds to the Principle I as mentioned in the Introduction.

4.1.1 GNet: Group Distance Minimization Guided Network

The ***group distance minimization*** focuses on learning representations that minimize the distance between the treated and controlled groups, and the corresponding theoretical foundation is supported by Wasserstein distance guided counterfactual and ITE error bounds (Theorem 1). Previous causal inference methods (e.g., Shalit et al. (2017); Yao et al. (2018); Zhang et al. (2020); Huang et al. (2022a)) commonly adopt Wasserstein distance to achieve group distance minimization. Specifically, these methods aim to minimize the empirical approximation of $\mathcal{L}_G(\mathbf{x}, \mathbf{t}; \Phi) = Wass\left(\{\Phi(\mathbf{x}_i)\}_{i:t_i=0}, \{\Phi(\mathbf{x}_i)\}_{i:t_i=1}\right)$ to learn balancing patterns. If we denote $\Phi_E(\cdot)$ by the feature extractor that extracts the original covariates $\mathbf{x}$, then the objective function designed on Theorem 1 is

$$\min_{\Phi_E, h^t} \quad \mathcal{L}_y(\mathbf{x}, \mathbf{t}, \mathbf{y}; \Phi_E, h^t) + \alpha_1 \mathcal{L}_G(\mathbf{x}, \mathbf{t}; \Phi_E). \tag{12}$$

Since the objective is to learn balancing patterns by minimizing the distributional distance between treated and controlled groups, i.e., group distance minimization, we refer to a model with the objective in equation (12) as **GNet**. For the reader's convenience, we illustrate the structure of GNet in Figure 2(a). Note that CFRNet (Shalit et al., 2017) is also the category of GNet.

4.1.2 INet: Individual Propensity Confusion Guided Network

In the field of causal inference, the propensity score plays a central role because it characterizes the probability that one receives treatment (Rosenbaum & Rubin, 1983). For example, the propensity score has been widely employed in prior literature for matching (Caliendo & Kopeinig, 2008) or weighting (Austin & Stuart, 2015) purposes. In this paper, we emphasize that the propensity score also plays an important role in representation balancing, where it serves as a natural indicator of the adequacy of leanred balancing patterns. Specifically, we propose the concept of individual propensity confusion, which aims to learn representations that are difficult to utilize for characterizing the propensity of each individual being treated or controlled. The corresponding theoretical foundation is based on our derived $\mathcal{H}$-divergence guided counterfactual and ITE error bounds (Theorem 2). In the upcoming content, we will detail how Theorem 2 establishes the connection between representation balancing and individual propensity confusion.

Let $\mathbb{I}(a)$ be an indicator function that gives 1 if a is true, and $\mathcal{H}$ be the family of binary functions as defined in Theorem 2. To achieve representation balancing, the objective function designed on Theorem 2 should aim to minimize the empirical $\mathcal{H}$-divergence $\hat{d}_{\mathcal{H}}(p_{\Phi}^{T=1}, p_{\Phi}^{T=0})$ such that

$$\hat{d}_{\mathcal{H}}(p_{\Phi}^{T=1}, p_{\Phi}^{T=0}) = 2\left(1 - \min_{\eta\in\mathcal{H}}\left[\frac{1}{N}\sum_{i:\eta(\Phi(\mathbf{x}_i))=0}\mathbb{I}[t_i = 1] + \frac{1}{N}\sum_{i:\eta(\Phi(\mathbf{x}_i))=1}\mathbb{I}[t_i = 0]\right]\right). \tag{13}$$

The "min" part in equation (13) indicates that the optimal classifier $\eta^* \in \mathcal{H}$ minimizes the classification error between the estimated treatment $\eta^*(\Phi(\mathbf{x}_i))$ and the observed treatment t_i, i.e., discriminating whether $\Phi(\mathbf{x}_i)$ is a control ($T = 0$) or treatment ($T = 1$). Equation (13) suggests that $\hat{d}_{\mathcal{H}}(p_{\Phi}^{T=1}, p_{\Phi}^{T=0})$ will be large if η^* can easily distinguish whether $\Phi(\mathbf{x}_i)$ is treated or controlled. In contrast, $\hat{d}_{\mathcal{H}}(p_{\Phi}^{T=1}, p_{\Phi}^{T=0})$ will be small if it is hard for η^* to determine whether $\Phi(\mathbf{x}_i)$ is treated or controlled. Therefore, the prerequisite of a small $\mathcal{H}$-divergence is to find a map Φ such that any classifier $\eta \in \mathcal{H}$ will get confused about the probability of $\Phi(\mathbf{x}_i)$ being treated or controlled. To achieve this goal, similar to the strategy of empirical approximation of $\mathcal{H}$-divergence (Ganin et al., 2016), we define a discriminator $\pi(\mathbf{r}) : \mathcal{R} \to [0, 1]$ that estimates the propensity score of $\mathbf{r}$, which can be regarded as a surrogate for $\eta(\mathbf{r})$. The classification error for the i^{th} individual can be empirically approximated by the cross-entropy loss between $\pi(\Phi(\mathbf{x}_i))$ and t_i:

$$\mathcal{L}_t(t_i, \pi(\Phi(\mathbf{x}_i))) = -\left[t_i \log \pi(\Phi(\mathbf{x}_i)) + (1 - t_i)\log(1 - \pi(\Phi(\mathbf{x}_i)))\right]. \tag{14}$$

As a consequence, we aim to find an optimal discriminator π^* for equation (13) such that π^* maximizes the probability that treatment is correctly classified:

$$\max_{\pi\in\mathcal{H}} \mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi, \pi) = \max_{\pi\in\mathcal{H}}\left[-\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_t(t_i, \pi(\Phi(\mathbf{x}_i)))\right]. \tag{15}$$

Given the feature extractor $\Phi_E(\cdot)$, the objective of INet can be formulated as a min-max game:

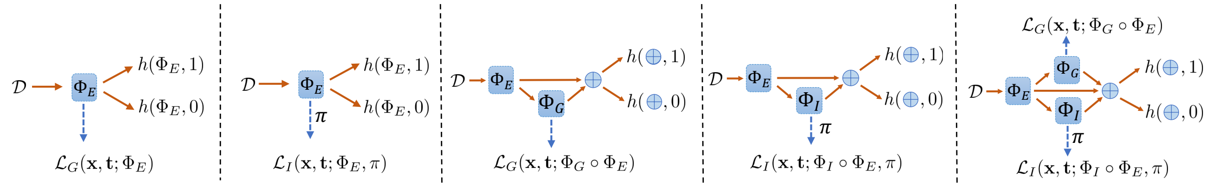
$$\min_{\Phi_E, h^t}\max_{\pi} \quad \mathcal{L}_y(\mathbf{x}, \mathbf{t}, \mathbf{y}; \Phi_E, h^t) + \alpha_2\mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi_E, \pi). \tag{16}$$

In the maximization, the discriminator π is trained to maximize the probability that treatment is correctly classified. This forces $\pi(\Phi_E(\mathbf{x}_i))$ closer to the true propensity score $e(\mathbf{x}_i)$. In the minimization, the feature extractor Φ_E is trained to fool the discriminator π. This confuses π such that $\pi(\Phi_E(\mathbf{x}_i))$ cannot correctly specify the true propensity score $e(\mathbf{x}_i)$. Eventually, the representations are balanced as the adversarial process makes it difficult for π to determine the propensity of each individual being treated or controlled. We refer to this process as ***individual propensity confusion***. For the reader's convenience, we illustrate the structure of INet in Figure 2(b).

4.2 Representation Balancing with Decomposed Patterns

4.2.1 The Proposed PDIG and PPBR Methods

PDIG. Although Theorems 1 and 2 provide solid theoretical foundation for GNet (model proposed by Shalit et al. (2017)) and INet (model proposed by us), both of these model types still encounter the inherent trade-off between representation balancing and outcome modeling. To this end, we expect to capture more effective balancing patterns by learning **P**atterns **D**ecomposed with **I**ndividual propensity confusion and


$\mathcal{D}$
Φ_E
$h(\Phi_E, 1)$
$h(\Phi_E, 0)$
$\mathcal{L}_G(\mathbf{x}, \mathbf{t}; \Phi_E)$
π
$\mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi_E, \pi)$
Φ_G
$h(\oplus, 1)$
$h(\oplus, 0)$
$\mathcal{L}_G(\mathbf{x}, \mathbf{t}; \Phi_G \circ \Phi_E)$
Φ_I
$\mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi_I \circ \Phi_E, \pi)$

(a) GNet (CFR-Wass) (b) INet (c) DGNet (d) DINet (e) DIGNet

Figure 2: Illustrations of the network architecture of the five models studied in Section 5.

Group distance minimization **(PDIG)**. More specifically, the covariates $\mathbf{x}$ are extracted by the feature extractor $\Phi_E(\cdot)$, and then $\Phi_E(\mathbf{x})$ are fed into two distinct balancing networks $\Phi_G(\cdot)$ and $\Phi_I(\cdot)$ for group distance minimization and individual propensity confusion, respectively. In summary, PDIG decomposes the balancing patterns into two distinct parts, group distance minimization and individual propensity confusion, which are respectively achieved by the following loss functions:

$$\min_{\Phi_G} \quad \mathcal{L}_G(\mathbf{x}, \mathbf{t}; \Phi_G \circ \Phi_E) \tag{17}$$

$$\min_{\Phi_I} \max_{\pi} \quad \mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi_I \circ \Phi_E, \pi). \tag{18}$$

Here, $\circ$ denotes the composition of two functions, indicating that $\Phi(\cdot)$ in $\mathcal{L}_G(\mathbf{x}, \mathbf{t}; \Phi)$ and $\mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi, \pi)$ are replaced by $\Phi_G(\Phi_E(\cdot))$ and $\Phi_I(\Phi_E(\cdot))$, respectively.

PPBR. Motivated by the discussion in Section 1, we design a framework that is capable of capturing **P**atterns of **P**re-balancing and **B**alancing **R**epresentations **(PPBR)** to mitigate potential over-balancing issue mentioned in the Introduction, aiming to preserve information that is useful for outcome predictions. In the PPBR method, the balancing patterns $\Phi_G(\Phi_E(\mathbf{x}))$ and $\Phi_I(\Phi_E(\mathbf{x}))$ are first learned over Φ_G and Φ_I, while Φ_E is remained fixed as pre-balancing patterns. Furthermore, we concatenate the balancing patterns $\Phi_G(\Phi_E(\mathbf{x}))$ and $\Phi_I(\Phi_E(\mathbf{x}))$ with the pre-balancing representations $\Phi_E(\mathbf{x})$ as attributes for outcome prediction. As a result, the proxy features used for outcome predictions are $\Phi_E(\mathbf{x}) \oplus \Phi_G(\Phi_E(\mathbf{x})) \oplus \Phi_I(\Phi_E(\mathbf{x}))$, where $\oplus$ indicates the concatenation by column. For example, if $\mathbf{a} = [1, 2]$ and $\mathbf{b} = [3, 4]$, then $\mathbf{a} \oplus \mathbf{b} = [1, 2, 3, 4]$. Consequently, representation balancing is accomplished over Φ_G and Φ_I, rather than Φ_E. Even if there may be a loss of information relevant to outcome prediction in Φ_G and Φ_I, the pre-balancing patterns Φ_E can still effectively preserve such information. Finally, the objective function with regard to outcome modeling under PPBR method becomes

$$\min_{\Phi_E, \Phi_I, \Phi_G, h^t} \quad \mathcal{L}_y(\mathbf{x}, \mathbf{t}, \mathbf{y}; \Phi_E \oplus (\Phi_I \circ \Phi_E) \oplus (\Phi_G \circ \Phi_E), h^t). \tag{19}$$

4.2.2 The Proposed DIGNet

Combining with PDIG and PPBR, we propose a new neural **Net**work model that incorporates **D**ecomposed patterns with **I**ndividual propensity confusion and **G**roup distance minimization, which we call **DIGNet**. The objective of DIGNet is separated into four stages:

$$\min_{\Phi_G} \quad \alpha_1 \mathcal{L}_G(\mathbf{x}, \mathbf{t}; \Phi_G \circ \Phi_E), \tag{20}$$

$$\max_{\pi} \quad \alpha_2 \mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi_I \circ \Phi_E, \pi), \tag{21}$$

$$\min_{\Phi_I} \quad \alpha_2 \mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi_I \circ \Phi_E, \pi), \tag{22}$$

$$\min_{\Phi_E, \Phi_I, \Phi_G, h^t} \quad \mathcal{L}_y(\mathbf{x}, \mathbf{t}, \mathbf{y}; \Phi_E \oplus (\Phi_I \circ \Phi_E) \oplus (\Phi_G \circ \Phi_E), h^t). \tag{23}$$

Within each iteration, DIGNet minimizes the group distance through equation 20, and plays an adversarial game to achieve propensity confusion through equation 21 and equation 22. In equation 23, DIGNet updates

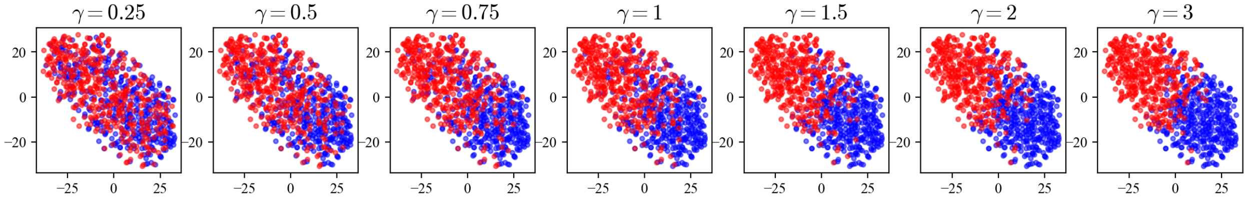


Figure 3: T-SNE visualizations of the covariates as γ varies. Red represents the treatment group and blue represents the control group. A larger γ indicates a greater imbalance between the two groups.

both the pre-balancing patterns Φ_E and balancing patterns Φ_I, Φ_G, along with the outcome function h^t to minimize the outcome prediction loss. For the reader's convenience, we illustrate the structure of DIGNet in Figure 2(e).

5 Experiments

In non-randomized observational data, the ground truth regarding treatment effects remains inaccessible due to the absence of counterfactual information. Therefore, we use simulated data and semi-synthetic benchmark data to test the performance of our methods and other baseline models. In this section, we primarily investigate the three following questions:

Q1. Is PDIG helpful in ITE estimation through Path I in the Introduction, i.e., learning more effective balancing patterns without affecting factual outcome prediction?

Q2. Is PPBR helpful in ITE estimation through Path II in the Introduction, i.e., improving factual outcome prediction without affecting learning balancing patterns?

Q3. Can the proposed DIGNet model outperform other baseline models on benchmark dataset?

Ablation models. To investigate Q1 and Q2, we conducted ablation studies and designed two ablation models, **DGNet** and **DINet**, where DGNet (or DINet) can be considered as DIGNet without PDIG, and GNet (or INet) can be considered as DGNet (or DINet) without PPBR. The structures of DGNet and DINet are shown in Figure 2(c) and Figure 2(d), and the objectives of DGNet and DINet are deferred to Section A.5.

5.1 Experimental Settings

Simulation data. Previous causal inference works assess the model effectiveness by varying the distribution imbalance of covariates in treated and controlled groups at different levels (Yao et al., 2018; Yoon et al., 2018; Du et al., 2021). As suggested by Assaad et al. (2021), we draw 1000 observational data points from the following data generating strategy:

$$\begin{aligned}
&\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot [\rho \mathbf{1}_p \mathbf{1}_p^{'} + (1-\rho)\mathbf{I}_p]),\\
&T_i \mid \mathbf{X}_i \sim \text{Bernoulli}(1/(1+\exp(-\gamma \mathbf{X}_i))),\\
&Y_i^0 = \boldsymbol{\beta}_\mathbf{0}' \mathbf{X}_i + \xi_i, \quad Y_i^1 = \boldsymbol{\beta}_\mathbf{1}' \mathbf{X}_i + \xi_i, \quad \xi_i \sim \mathcal{N}(0,1).
\end{aligned}$$

Here, $\mathbf{1}_p$ denotes the p-dimensional all-ones vector and $\mathbf{I}_p$ denotes the identity matrix of size p. We fix $p = 10, \rho = 0.3, \sigma^2 = 2, \boldsymbol{\beta}_\mathbf{0}' = [0.3, ..., 0.3], \boldsymbol{\beta}_\mathbf{1}' = [1.3, ..., 1.3]$ and vary $\gamma \in \{0.25, 0.5, 0.75, 1, 1.5, 2, 3\}$ to yield different levels of selection bias. As seen in Figure 3, selection bias becomes more severe with γ increasing. For each γ, we repeat the above data generating process to generate 30 different datasets, with each dataset split by the ratio of 56%/24%/20% as training/validation/test sets.

Semi-synthetic data. The IHDP dataset is introduced by Hill (2011). This dataset consists of 747 samples with 25-dimensional covariates collected from real-world randomized experiments. Selection bias is

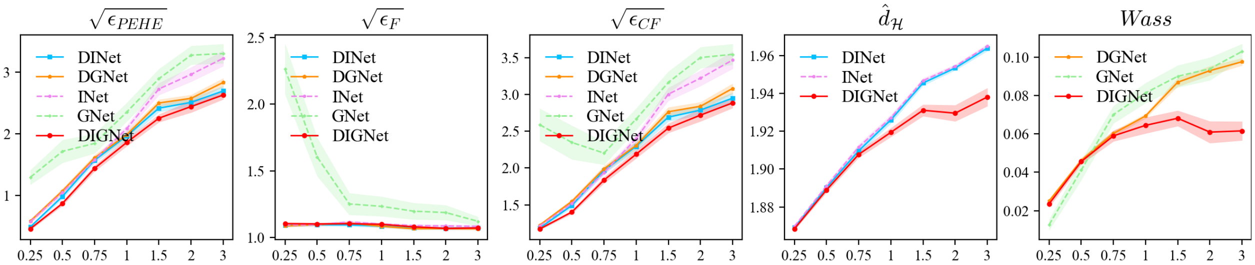


Figure 4: Plots of model performances on test set for different metrics as γ varies in $\{0.25, 0.5, 0.75, 1, 1.5, 2, 3\}$. Each graph shows the average of 30 runs with standard errors shaded. Lower lines indicate lower values of the metric.

created by removing some of treated samples. The goal is to estimate the effect of special visits (treatment) on cognitive scores (outcome). The potential outcomes are generated using the NPCI package Dorie (2021). We use the same 1000 datasets as used in Shalit et al. (2017), with each dataset split by the ratio of 63%/27%/10% as training/validation/test sets.

Models and metrics. In simulation experiments, we perform comprehensive comparisons between INet, GNet, DINet, DGNet, and DIGNet in terms of the mean and standard error for the following metrics: $\sqrt{\epsilon_{PEHE}}$, $\sqrt{\epsilon_{CF}}$, and $\sqrt{\epsilon_{F}}$ with L defined in Definition 1 being the squared loss, as well as the empirical approximations of $Wass(p_{\Phi}^{T=1}, p_{\Phi}^{T=0})$ and $d_{\mathcal{H}}(p_{\Phi}^{T=1}, p_{\Phi}^{T=0})$ (denoted by $Wass$ and $\hat{d}_{\mathcal{H}}$, respectively). Note that as shown in Figure 2, $Wass$ is over Φ_E for GNet while over Φ_G for DGNet and DIGNet; $\hat{d}_{\mathcal{H}}$ is over Φ_E for INet while over Φ_I for DINet and DIGNet. To analyze the source of gain and ensure fair comparison in simulation studies, we fix hyperparameters across all models. This way is consistent with Curth & van der Schaar (2021). We apply an early stopping rule to all models as Shalit et al. (2017) do. In IHDP experiment, we use $\sqrt{\epsilon_{PEHE}}$, as well as an additional metric $\epsilon_{ATE} = |\hat{\tau}_{ATE} - \tau_{ATE}|$ to evaluate performances of various causal models (see them in Table 4). More descriptions of the implementation details, as well as the analysis of training time and training stability, are deferred to Section A.4.

Device. All the experiments are run on Dell 7920 with one 16-core Intel Xeon Gold 6250 3.90GHz CPU and three NVIDIA Quadro RTX 6000 GPUs.

5.2 Results and Analysis

5.2.1 Preliminary Experimental Results

In this part, we first make a general comparison between different models with the degree of covariate imbalance increasing, and the relevant results are shown in Figure 4. There are four main observations:

1. DIGNet attains the lowest $\sqrt{\epsilon_{PEHE}}$ across all datasets, while GNet have inferior performances than other models;
2. DINet and DGNet outperform INet and GNet regarding $\sqrt{\epsilon_{CF}}$ and $\sqrt{\epsilon_{PEHE}}$;
3. INet, DINet, and DGNet have comparable performance to DIGNet in terms of factual outcome estimations ($\sqrt{\epsilon_{F}}$), but cannot compete with DIGNet in terms of counterfactual estimations ($\sqrt{\epsilon_{CF}}$) or ITE estimations ($\sqrt{\epsilon_{PEHE}}$);
4. DIGNet achieves smaller $\hat{d}_{\mathcal{H}}$ (or $Wass$) than DINet and INet (or DGNet and GNet), especially when the covariate shift problem is severe (e.g., when $\gamma > 1$).

In conclusion, the above study has produced several noteworthy findings. Firstly, finding (1) reveals that our proposed DIGNet model consistently performs well in ITE estimation. Secondly, as indicated by finding

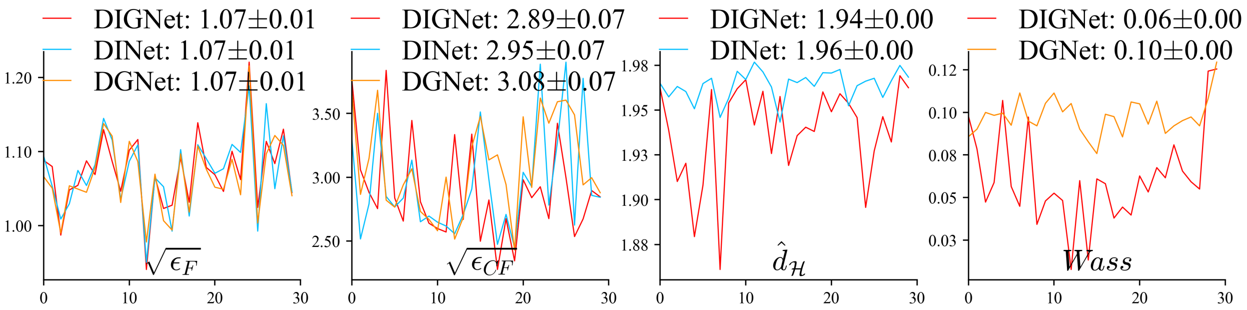


Figure 5: Plots of model performances on test set for $\sqrt{\epsilon_F}$, $\sqrt{\epsilon_{CF}}$, $\hat{d}_{\mathcal{H}}$, and $Wass$ when $\gamma = 3$. Each graph plots the metric for 30 runs. Mean ± std of each metric averaged across 30 runs are reported on the top. Lower lines indicate lower values of the metric.

(2), implementing the PPBR approach can enhance the predictive accuracy of factual and counterfactual outcomes. Lastly, findings (3) and (4) highlight the role of PDIG structure in enhancing the simultaneous reinforcement and complementarity of group distance minimization and individual propensity confusion, resulting in more balanced representations. Our subsequent analysis will step beyond these preliminary conclusions to gain a deeper understanding of the effectiveness of the proposed methods.

5.2.2 Further Ablation Studies

So far our preliminary observations have show that the relationship between the ITE errors of each model is: DIGNet<DINet<INet and DIGNet<DGNet<GNet. To further explore how PDIG and PPBR contribute to the improvement of ITE estimations, we choose the case with high selection bias ($\gamma = 3$) to analyze the source of gain for PDIG and PPBR. We plot specific metrics of 30 runs on test set in Figure 5 and Figure 6. We also report model performances (mean ± std) averaged over 30 training and test sets in Table 1. Below we discuss the source of gain in detail.

Ablation study for PDIG. ***The PDIG structure is manifest to be effective in capturing more effective balancing patterns, without affecting factual outcome predictions.*** As depicted in Figure 4, DIGNet exhibits more balanced representations, irrespective of whether the discrepancy is measured by $\hat{d}_{\mathcal{H}}$ or $Wass$, while DIGNet, DINet, and DGNet demonstrate comparable estimates of factual outcomes ($\sqrt{\epsilon_F}$). Two additional pieces of specific evidence can be observed from Figure 5: (1) Despite the absence of PDIG in DINet and DGNet when compared to DIGNet, these three models exhibit very similar performance regarding $\sqrt{\epsilon_F}$, with the performance being 1.07 ± 0.01. This indicates that PDIG does not impact the factual estimation. (2) DIGNet achieves smaller $\hat{d}_{\mathcal{H}}$ with a $|1.94/1.96 - 1| = 1.0\%$ reduction (or $Wass$ with a $|0.06/0.10 - 1| = 40\%$ reduction) compared with DINet (or DGNet). This indicates that PDIG enables the model to learn more effective balancing patterns. The above two points indicate that PDIG can capture more effective balancing patterns, without affecting factual outcome predictions. This advantage translates into superior counterfactual estimation, with DIGNet reduceing $\sqrt{\epsilon_{CF}}$ by $|2.89/2.95 - 1| = 2.0\%$ and $|2.89/3.08 - 1| = 6.2\%$ compared to DINet and DGNet, respectively. Correspondingly, DIGNet also shows superiority in treatment effect estimation ($\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE}) compared to DINet (or DGNet), as demonstrated in Table 1.

Ablation study for PPBR. ***The PPBR approach contributes to enhancing factual outcome predictions, without affecting learning balancing patterns.*** From Figure 6, we gain two important insights: (1) The difference in learned balancing patterns, measured by $\hat{d}_{\mathcal{H}}$ (or $Wass$), between DINet and INet (or DGNet and GNet), is negligible. This implies that PPBR does not affect learning balancing patterns. (2) Compared with INet, DINet achieves smaller $\sqrt{\epsilon_F}$, with $|1.07/1.08 - 1| = 0.9\%$ error reduction. Similarly, compared with GNet, DGNet achieves smaller $\sqrt{\epsilon_F}$, with $|1.07/1.12 - 1| = 4.5\%$ error reduction. These two observations reveal that PPBR can improve factual outcome predictions, without affecting learning balancing patterns. Benefiting from the advantage of PPBR, the improvement is particularly pronounced in

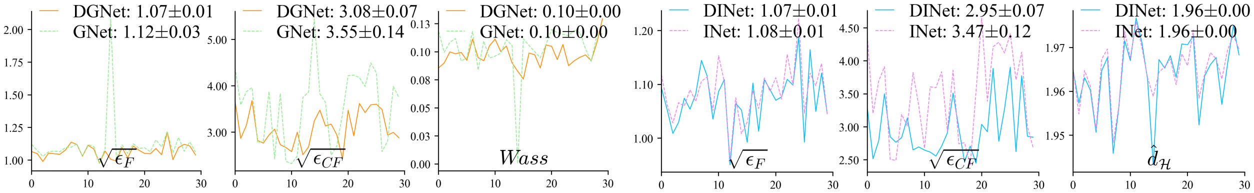


Figure 6: Plots of model performances on test set for different metrics when $\gamma = 3$. Each graph plots the metric for 30 runs, with mean ± std averaged across 30 runs reported on the top. Lower lines indicate lower values of the metric.

counterfactual estimation. Comparing DINet with INet, the reduction in $\sqrt{\epsilon_{CF}}$ amounts to $|2.95/3.47 - 1| = 15.0\%$. Similarly, comparing DGNet with GNet, the reduction is $|3.08/3.55 - 1| = 13.2\%$. Correspondingly, DINet (or DGNet) attains smaller treatment effect errors ($\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE}) compared with INet (or GNet), as shown in Table 1.

Significance analysis for the improvements. To assess the significance of the improvements observed in the above ablation studies, we conducted an additional significance analysis by recording the values of $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE} for 30 runs of each of the 5 models (GNet, INet, DGNet, DINet, and DIGNet). Subsequently, we performed a t-test for GNet vs. DGNet, INet vs. DINet, DGNet vs. DIGNet, and DINet vs. DIGNet, to investigate the statistical significance of their differences. The relevant results are reported in Table 3. The results reveal a statistically significant difference between GNet and DGNet, INet and DINet, as well as DGNet and DIGNet. Note that the difference between DINet and DIGNet is not statistically significant, despite DIGNet exhibiting smaller treatment effect estimation errors on average compared to DINet.

Table 1: Training- & test- set $\sqrt{\epsilon_{PEHE}}$ & ϵ_{ATE} when $\gamma = 3$. Mean ± standard error of 30 runs.

	Training set		Test set	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
GNet	3.30±0.15	2.58±0.14	3.30±0.16	2.59±0.14
INet	3.24±0.11	2.46±0.09	3.22±0.12	2.47±0.10
DGNet	2.86±0.06	2.15±0.03	2.83±0.07	2.15±0.04
DINet	2.70±0.06	2.12±0.04	2.69±0.08	2.13±0.05
DIGNet	**2.66±0.07**	**2.04±0.05**	**2.63±0.07**	**2.03±0.04**

Table 2: Training- & test- set $\sqrt{\epsilon_{PEHE}}$ & ϵ_{ATE} on IHDP. Mean ± standard error of 100 runs.

	Training set		Test set	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
GNet	0.71±0.15	0.12±0.01	0.77±0.18	0.15±0.02
INet	0.66±0.09	0.13±0.01	0.72±0.11	0.15±0.02
DGNet	0.53±0.07	**0.11±0.01**	0.60±0.09	0.13±0.01
DINet	0.57±0.12	0.13±0.01	0.60±0.11	0.14±0.01
DIGNet	**0.42±0.02**	**0.11±0.01**	**0.45±0.04**	**0.12±0.01**

Table 3: Significance analysis regarding the achieved improvements by comparing GNet and DGNet, INet and DINet, DGNet and DIGNet, DINet and DIGNet. The p-value ≤ 0.05 indicates difference is statistically significant.

	Training set				Test set			
	$\sqrt{\epsilon_{PEHE}}$		ϵ_{ATE}		$\sqrt{\epsilon_{PEHE}}$		ϵ_{ATE}	
	t-value	p-value	t-value	p-value	t-value	p-value	t-value	p-value
GNet vs. DGNet	2.7435	**0.0081**	2.9844	**0.0042**	2.7073	**0.0089**	2.9269	**0.0049**
INet vs. DINet	4.0812	**0.0001**	3.5222	**0.0008**	3.5665	**0.0007**	3.0824	**0.0031**
DGNet vs. DIGNet	2.0240	**0.0476**	1.8888	0.0639	2.0650	**0.0434**	2.0935	**0.0407**
DINet vs. DIGNet	0.4513	0.6535	1.3525	0.1815	0.6079	0.5456	1.5473	0.1272

5.2.3 Comparisons on IHDP benchmark.

Table 4: Training- & test- set $\sqrt{\epsilon_{PEHE}}$ & ϵ_{ATE} on IHDP. Mean ± standard error of 1000 runs.

	Training set		Test set	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
OLS/LR$_1$ (Johansson et al., 2016)	$5.8 \pm .3$	$.73 \pm .04$	$5.8 \pm .3$	$.94 \pm .06$
OLS/LR$_2$ (Johansson et al., 2016)	$2.4 \pm .1$	$.14 \pm .01$	$2.5 \pm .1$	$.31 \pm .02$
k-NN (Crump et al., 2008)	$2.1 \pm .1$	$.14 \pm .01$	$4.1 \pm .2$	$.79 \pm .05$
BART (Chipman et al., 2010)	$2.1 \pm .1$	$.23 \pm .01$	$2.3 \pm .1$	$.34 \pm .02$
CF (Wager & Athey, 2018)	$3.8 \pm .2$	$.18 \pm .01$	$3.8 \pm .2$	$.40 \pm .03$
CEVAE (Louizos et al., 2017)	$2.7 \pm .1$	$.34 \pm .01$	$2.6 \pm .1$	$.46 \pm .02$
SITE (Yao et al., 2018)	$.69 \pm .0$	$.22 \pm .01$	$.75 \pm .0$	$.24 \pm .01$
GANITE (Yoon et al., 2018)	$1.9 \pm .4$	$.43 \pm .05$	$2.4 \pm .4$	$.49 \pm .05$
BLR (Johansson et al., 2016)	$5.8 \pm .3$	$.72 \pm .04$	$5.8 \pm .3$	$.93 \pm .05$
BNN (Johansson et al., 2016)	$2.2 \pm .1$	$.37 \pm .03$	$2.1 \pm .1$	$.42 \pm .03$
TARNet (Shalit et al., 2017)	$.88 \pm .0$	$.26 \pm .01$	$.95 \pm .0$	$.28 \pm .01$
CFR-Wass (GNet) (Shalit et al., 2017)	$.73 \pm .0$	$.12 \pm .01$	$.81 \pm .0$	$.15 \pm .01$
Dragonnet (Shi et al., 2019)	$1.3 \pm .4$	$.14 \pm .01$	$1.3 \pm .5$	$.20 \pm .05$
MBRL (Huang et al., 2022a)	$.52 \pm .0$	$.12 \pm .01$	$.57 \pm .0$	$.13 \pm .01$
DIGNet (Ours)	$\mathbf{.42 \pm .0}$	$\mathbf{.11 \pm .01}$	$\mathbf{.45 \pm .0}$	$\mathbf{.12 \pm .01}$

In this part, we perform experiments on the IHDP benchmark dataset to compare the performances of different models. The corresponding results are reported in Table 2 and 4.

First, we report the ablation results on 1-100 IHDP datasets in Table 2, aiming to examine the consistent effectiveness of PDIG and PPBR. Specifically, Table 2 shows that DINet and DGNet are superior to INet and GNet but inferior to DIGNet concerning treatment effect estimation, suggesting that both PDIG and PPBR are advantageous for treatment effect estimation. For example, on the test set, DINet reduces $\sqrt{\epsilon_{PEHE}}$ by $|0.60/0.72 - 1| = 16.7\%$ for INet, and DIGNet reduces $\sqrt{\epsilon_{PEHE}}$ by $|0.45/0.60 - 1| = 25\%$ for DINet. This is consistent with the findings before: PDIG and PPBR are beneficial to treatment effect estimation.

Furthermore, we undergo comparisons between DIGNet and other causal models on 1-1000 IHDP datasets and report the results in Table 4. The results highlight the superior performance of the proposed DIGNet across all the models. Specifically, in comparison to the second-best method in test-sample performance, DIGNet achieves a substantial improvement, with error reducted by $|0.45/0.57 - 1| = 21\%$ in terms of $\sqrt{\epsilon_{PEHE}}$ and $|0.12/0.13 - 1| = 7.7\%$ in terms of ϵ_{ATE}. Moreover, it is worth noting that DIGNet consistently achieves the lowest errors across various datasets and metrics, revealing its robust performance. We also conduct an additional experiments on another benchmark dataset Twins. The details and results are deferred to Section A.4

6 Conclusion

This paper establishes a theoretical foundation by deriving counterfactual and ITE error bounds based on $\mathcal{H}$-divergence. This theoretical foundation builds a connection between representation balancing and individual propensity confusion. Furthermore, based on individual propensity confusion and group distance minimization, we suggest learning decomposed patterns for representation balancing models using the PDIG and PPBR methods. Further, building upon PDIG and PPBR, we propose a novel model DIGNet, for treatment effect estimation. Comprehensive experiments verify that PDIG and PPBR follow different pathways to improve counterfactual and ITE estimation. In particular, PDIG enables the model to capture more effective balancing patterns without affecting factual outcome prediction, while PPBR contributes to improving factual outcome predictions without influencing learning balancing patterns. We hope these findings can constitute an important step to inspire more research concerning the generalization of representation balancing models for counterfactual and ITE estimation.

Limitations. Our paper verifies the effectiveness of PDIG and PPBR in improving ITE estimation, it is also important to step beyond our empirical insights into future theoretical studies aimed at addressing the trade-off challenge mentioned in the introduction, e.g., exploring the possibility of deriving tighter theoretical error bounds based on learning decomposed patterns. Furthermore, it remains challenging to analytically determine the best divergence metric for representation balancing methods. A promising avenue for future theoretical investigations would involve developing new distributional divergences or exploring a unified theory that enables models to select appropriate divergence metrics based on the distinct data. Empirical studies can focus on discouraging the redundancy of shared information within the decomposed patterns and improving the optimization efficacy of DIGNet.

References

Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. Counterfactual representation learning with balancing weights. In International Conference on Artificial Intelligence and Statistics, pp. 1972–1980. PMLR, 2021.

Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. Observational Studies, 5(2):37–51, 2019.

Peter C Austin and Elizabeth A Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. Statistics in medicine, 34(28):3661–3679, 2015.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. Advances in neural information processing systems, 19, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. Machine learning, 79(1):151–175, 2010.

Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela Van Der Schaar. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. Clinical Pharmacology & Therapeutics, 109(1):87–100, 2021a.

Ioana Bica, Daniel Jarrett, Alihan Hüyük, and Mihaela van der Schaar. Learning ”what-if” explanations for sequential decision-making. In International Conference on Learning Representations, 2021b. URL `https://openreview.net/forum?id=h0de3QWtGG`.

Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. Journal of economic surveys, 22(1):31–72, 2008.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1):C1–C68, 2018.

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. The Annals of Applied Statistics, 4(1):266–298, 2010.

Zhixuan Chu, Stephen L. Rathbun, and Sheng Li. Graph infomax adversarial learning for treatment effect estimation with networked observational data. In KDD, pp. 176–184, 2021. URL `https://doi.org/10.1145/3447548.3467302`.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Nonparametric tests for treatment effect heterogeneity. The Review of Economics and Statistics, 90(3):389–405, 2008.

Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. Advances in Neural Information Processing Systems, 34:15883–15894, 2021.

Alicia Curth, David Svensson, Jim Weatherall, and Mihaela van der Schaar. Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation. In Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2), 2021.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In International conference on machine learning, pp. 685–693. PMLR, 2014.

Vincent Dorie. Nonparametric methods for causal inference. https://github.com/vdorie/npci, 2021.

Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. Data Mining and Knowledge Discovery, 35(4):1713–1738, 2021.

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. Journal of Econometrics, 189(1):1–23, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. The journal of machine learning research, 17(1):2096–2030, 2016.

Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. ACM Computing Surveys (CSUR), 53(4):1–37, 2020a.

Ruocheng Guo, Jundong Li, Yichuan Li, K Selçuk Candan, Adrienne Raglin, and Huan Liu. Ignite: A minimax game toward learning individual treatment effects from networked observational data. In IJCAI, pp. 4534–4540, 2020b.

Ruocheng Guo, Jundong Li, and Huan Liu. Learning individual causal effects from networked observational data. In Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 232–240, 2020c.

Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In IJCAI, pp. 5880–5887, 2019a.

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In International Conference on Learning Representations, 2019b.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20(1):217–240, 2011.

Yiyan Huang, Cheuk Hang Leung, Xing Yan, Qi Wu, Nanbo Peng, Dongdong Wang, and Zhixiang Huang. The causal learning of retail delinquency. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pp. 204–212, 2021.

Yiyan Huang, Cheuk Hang Leung, Shumin Ma, Qi Wu, Dongdong Wang, and Zhixiang Huang. Moderately-balanced representation learning for treatment effects with orthogonality information. In Pacific Rim International Conference on Artificial Intelligence, pp. 3–16. Springer, 2022a.

Yiyan Huang, Cheuk Hang Leung, Qi Wu, Xing Yan, Shumin Ma, Zhiri Yuan, Dongdong Wang, and Zhixiang Huang. Robust causal learning for the estimation of average treatment effects. In 2022 International Joint Conference on Neural Networks (IJCNN 2022). IEEE, 2022b.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In International conference on machine learning, pp. 3020–3029. PMLR, 2016.

Fredrik D. Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. Journal of Machine Learning Research, 23(166):1–50, 2022. URL `http://jmlr.org/papers/v23/19-511.html`.

Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang, and Fei Wang. Treatment effect estimation with data-driven variable decomposition. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the national academy of sciences, 116(10):4156–4165, 2019.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer joint matching for unsupervised domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1410–1417, 2014.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. Advances in neural information processing systems, 30, 2017.

Maggie Makar, Fredrik Johansson, John Guttag, and David Sontag. Estimation of bounds on potential outcomes for decision making. In International Conference on Machine Learning, pp. 6661–6671. PMLR, 2020.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. Biometrika, 108 (2):299–319, 2021.

Zhaozhi Qian, Yao Zhang, Ioana Bica, Angela Wood, and Mihaela van der Schaar. Synctwin: Treatment effect estimation with longitudinal outcomes. Advances in Neural Information Processing Systems, 34: 3178–3190, 2021.

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

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association, 100(469):322–331, 2005.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In International Conference on Machine Learning, pp. 3076–3085. PMLR, 2017.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. Advances in neural information processing systems, 32, 2019.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. Electronic Journal of Statistics, 6:1550–1599, 2012.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242, 2018.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. Advances in Neural Information Processing Systems, 31, 2018.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. ACM Transactions on Knowledge Discovery from Data (TKDD), 15(5):1–46, 2021.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In International Conference on Learning Representations, 2018.

Yao Zhang, Alexis Bellot, and Mihaela Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In International Conference on Artificial Intelligence and Statistics, pp. 1005–1014. PMLR, 2020.

A Appendix

A.1 Proof of Lemma 1

Proof of L taking the squared loss, i.e., $L(y_1, y_2) = (y_1 - y_2)^2$:

Proof. We denote $\epsilon_{PEHE}(f) = \epsilon_{PEHE}(h, \Phi)$, $\epsilon_F(f) = \epsilon_F(h, \Phi)$, $\epsilon_{CF}(f) = \epsilon_{CF}(h, \Phi)$ for $f(\mathbf{x}, t) = h(\Phi(\mathbf{x}), t)$.

$$
\begin{aligned}
&\epsilon_F(f) \\
=&\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (f(\mathbf{x},t)-y^t)^2 p(y^t|\mathbf{x})p(\mathbf{x},t)dy^t d\mathbf{x}dt \\
=&\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (f(\mathbf{x},t)-\tau^t(\mathbf{x}))^2 p(y^t|\mathbf{x})p(\mathbf{x},t)dy^t d\mathbf{x}dt \\
&+\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (\tau^t(\mathbf{x})-y^t)^2 p(y^t|\mathbf{x})p(\mathbf{x},t)dy^t d\mathbf{x}dt \\
&+2\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} (f(\mathbf{x},t)-\tau^t(\mathbf{x}))(\tau^t(\mathbf{x})-y^t) p(y^t|\mathbf{x})p(\mathbf{x},t)dy^t d\mathbf{x}dt \qquad (24)\\
=&\int_{\mathcal{X}\times\{0,1\}} (f(\mathbf{x},t)-\tau^t(\mathbf{x}))^2 p(\mathbf{x},t)d\mathbf{x}dt + \sigma^2_{y^t}(p(\mathbf{x},t)) \qquad (25)
\end{aligned}
$$

Equation (25) is by the definition of $\sigma^2_{y^t}(p(\mathbf{x},t))$ in Lemma 1 and equation (24) equaling zero since $\tau^t(\mathbf{x}) = \int_{\mathcal{Y}} y^t p(y^t|\mathbf{x})dy_t$. A similar result can be obtained for ϵ_{CF}:

$$
\epsilon_{CF}(f) = \int_{\mathcal{X}\times\{0,1\}} (f(\mathbf{x},t)-\tau^t(\mathbf{x}))^2 p(\mathbf{x},1-t)d\mathbf{x}dt + \sigma^2_{y^t}(p(\mathbf{x},1-t)).
$$

$$
\begin{aligned}
&\epsilon_{PEHE}(f) \\
=&\int_{\mathcal{X}} ((f(\mathbf{x},1)-f(\mathbf{x},0))-(\tau^1(\mathbf{x})-\tau^0(\mathbf{x})))^2 p(\mathbf{x})d\mathbf{x} \\
\le&2\int_{\mathcal{X}} ((f(\mathbf{x},1)-\tau^1(\mathbf{x}))^2+(f(\mathbf{x},0)-\tau^0(\mathbf{x}))^2)p(\mathbf{x})d\mathbf{x} \qquad (26)\\
=&2\int_{\mathcal{X}} (f(\mathbf{x},1)-\tau^1(\mathbf{x}))^2 p(\mathbf{x},T=1)d\mathbf{x} + 2\int_{\mathcal{X}} (f(\mathbf{x},0)-\tau^0(\mathbf{x}))^2 p(\mathbf{x},T=0)d\mathbf{x} \\
&+2\int_{\mathcal{X}} (f(\mathbf{x},1)-\tau^1(\mathbf{x}))^2 p(\mathbf{x},T=0)d\mathbf{x} + 2\int_{\mathcal{X}} (f(\mathbf{x},0)-\tau^0(\mathbf{x}))^2 p(\mathbf{x},T=1)d\mathbf{x} \qquad (27)\\
=&2\int_{\mathcal{X}\times\{0,1\}} (f(\mathbf{x},t)-\tau^t(\mathbf{x}))^2 p(\mathbf{x},t)d\mathbf{x}dt + 2\int_{\mathcal{X}\times\{0,1\}} (f(\mathbf{x},t)-\tau^t(\mathbf{x}))^2 p(\mathbf{x},1-t)d\mathbf{x}dt \\
=&2(\epsilon_F(f)-\sigma^2_{y^t}(p(\mathbf{x},t))) + 2(\epsilon_{CF}(f)-\sigma^2_{y^t}(p(\mathbf{x},1-t))). \qquad (28)
\end{aligned}
$$

Inequality (26) is by $(x+y)^2 \le 2(x^2+y^2)$; equation (27) is by $p(\mathbf{x}) = p(\mathbf{x},T=0)+p(\mathbf{x},T=1)$. By (equation 28) and the definition of σ^2_y in Lemma 1, we have

$$
\epsilon_{PEHE}(h,\Phi) \le 2(\epsilon_{CF}(h,\Phi)+\epsilon_F(h,\Phi)-2\sigma^2_y).
$$

□

Proof of L taking the absolute loss, i.e., $L(y_1,y_2)=|y_1-y_2|$:

Proof. We denote $\epsilon_{PEHE}(f) = \epsilon_{PEHE}(h, \Phi)$, $\epsilon_F(f) = \epsilon_F(h, \Phi)$, $\epsilon_{CF}(f) = \epsilon_{CF}(h, \Phi)$ for $f(\mathbf{x}, t) = h(\Phi(\mathbf{x}), t)$.

$$\begin{aligned}
&\epsilon_F(f) \\
=&\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} |f(\mathbf{x}, t) - y^t|p(y^t|\mathbf{x})p(\mathbf{x}, t)dy^t d\mathbf{x}dt \\
\geq&\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} |f(\mathbf{x}, t) - \tau^t(\mathbf{x})|p(y^t|\mathbf{x})p(\mathbf{x}, t)dy^t d\mathbf{x}dt \\
&- \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} |\tau^t(\mathbf{x}) - y^t|p(y^t|\mathbf{x})p(\mathbf{x}, t)dy^t d\mathbf{x}dt \qquad (29) \\
=&\int_{\mathcal{X}\times\{0,1\}} |f(\mathbf{x}, t) - \tau^t(\mathbf{x})|p(\mathbf{x}, t)d\mathbf{x}dt - A_{y^t}(p(\mathbf{x}, t)). \qquad (30)
\end{aligned}$$

Inequality (29) is by $|x + y| \geq |x| - |y|$, equation (30) is by the definition of $A_{y^t}(p(\mathbf{x}, t))$ in Lemma 1. A similar result can be obtained for ϵ_{CF}:

$$\epsilon_{CF}(f) \geq \int_{\mathcal{X}\times\{0,1\}} |f(\mathbf{x}, t) - \tau^t(\mathbf{x})|p(\mathbf{x}, 1-t)d\mathbf{x}dt - A_{y^t}(p(\mathbf{x}, 1-t)).$$

$$\begin{aligned}
&\epsilon_{PEHE}(f) \\
=&\int_{\mathcal{X}} |(f(\mathbf{x}, 1) - f(\mathbf{x}, 0)) - (\tau^1(\mathbf{x}) - \tau^0(\mathbf{x}))|p(\mathbf{x})d\mathbf{x} \\
\leq&\int_{\mathcal{X}} (|f(\mathbf{x}, 1) - \tau^1(\mathbf{x})| + |f(\mathbf{x}, 0) - \tau^0(\mathbf{x})|)p(\mathbf{x})d\mathbf{x} \qquad (31) \\
=&\int_{\mathcal{X}} |f(\mathbf{x}, 1) - \tau^1(\mathbf{x})|p(\mathbf{x}, T=1)d\mathbf{x} + \int_{\mathcal{X}} |f(\mathbf{x}, 1) - \tau^1(\mathbf{x})|p(\mathbf{x}, T=0)d\mathbf{x} \qquad (32) \\
&+ \int_{\mathcal{X}} |f(\mathbf{x}, 0) - \tau^0(\mathbf{x})|p(\mathbf{x}, T=0)d\mathbf{x} + \int_{\mathcal{X}} |f(\mathbf{x}, 0) - \tau^0(\mathbf{x})|p(\mathbf{x}, T=1)d\mathbf{x} \qquad (33) \\
=&\int_{\mathcal{X}\times\{0,1\}} |f(\mathbf{x}, t) - \tau^t(\mathbf{x})|p(\mathbf{x}, t)d\mathbf{x}dt + \int_{\mathcal{X}\times\{0,1\}} |f(\mathbf{x}, t) - \tau^t(\mathbf{x})|p(\mathbf{x}, 1-t)d\mathbf{x}dt \\
\leq&\epsilon_F(f) + A_{y^t}(p(\mathbf{x}, t)) + \epsilon_{CF}(f) + A_{y^t}(p(\mathbf{x}, 1-t)). \qquad (34)
\end{aligned}$$

Inequality (31) is by $|x+y| \leq |x|+|y|$. Equation (32) and equation (33) are by $p(\mathbf{x}) = p(\mathbf{x}, T=0)+p(\mathbf{x}, T=1)$. By equation (34) and the definition of A_y in Lemma 1, we have

$$\begin{aligned}
\epsilon_{PEHE}(h, \Phi) &\leq \epsilon_F(f) + A_{y^t}(p(\mathbf{x}, t)) + \epsilon_{CF}(f) + A_{y^t}(p(\mathbf{x}, 1-t)) \\
&\leq \epsilon_{CF}(h, \Phi) + \epsilon_F(h, \Phi) + 2A_y.
\end{aligned}$$

□

A.2 Proof of Theorem 1

Proof of equation (4):

Proof.

$$\begin{aligned}
&\epsilon_{CF}(h,\Phi) - [(1-u)\cdot\epsilon_F^{T=1}(h,\Phi) + u\cdot\epsilon_F^{T=0}(h,\Phi)] \\
=&[(1-u)\cdot\epsilon_{CF}^{T=1}(h,\Phi) + u\cdot\epsilon_{CF}^{T=0}(h,\Phi)] - [(1-u)\cdot\epsilon_F^{T=1}(h,\Phi) + u\cdot\epsilon_F^{T=0}(h,\Phi)] \\
=&(1-u)\cdot[\epsilon_{CF}^{T=1}(h,\Phi) - \epsilon_F^{T=1}(h,\Phi)] + u\cdot[\epsilon_{CF}^{T=0}(h,\Phi) - \epsilon_F^{T=0}(h,\Phi)] \\
=&(1-u)\int_{\mathcal{X}} \ell_{h,\Phi}(\mathbf{x},1)(p^{T=0}(\mathbf{x}) - p^{T=1}(\mathbf{x}))d\mathbf{x} + u\int_{\mathcal{X}} \ell_{h,\Phi}(\mathbf{x},0)(p^{T=1}(\mathbf{x}) - p^{T=0}(\mathbf{x}))d\mathbf{x} \\
=&(1-u)\int_{\mathcal{R}} \ell_{h,\Phi}(\Psi(\mathbf{r}),1)(p_\Phi^{T=0}(\mathbf{r}) - p_\Phi^{T=1}(\mathbf{r}))d\mathbf{r} + u\int_{\mathcal{R}} \ell_{h,\Phi}(\Psi(\mathbf{r}),0)(p_\Phi^{T=1}(\mathbf{r}) - p_\Phi^{T=0}(\mathbf{r}))d\mathbf{r} \qquad (35) \\
=&B_\Phi\cdot(1-u)\int_{\mathcal{R}} \frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(\mathbf{r}),1)(p_\Phi^{T=0}(\mathbf{r}) - p_\Phi^{T=1}(\mathbf{r}))d\mathbf{r} \\
&+ B_\Phi\cdot u\int_{\mathcal{R}} \frac{1}{B_\Phi}\ell_{h,\Phi}(\Psi(\mathbf{r}),0)(p_\Phi^{T=1}(\mathbf{r}) - p_\Phi^{T=0}(\mathbf{r}))d\mathbf{r} \\
\le&B_\Phi\cdot(1-u)\sup_{g\in\mathcal{G}}|\int_{\mathcal{R}} g(\mathbf{r})(p_\Phi^{T=0}(\mathbf{r}) - p_\Phi^{T=1}(\mathbf{r}))d\mathbf{r}| \\
&+ B_\Phi\cdot u\cdot\sup_{g\in\mathcal{G}}|\int_{\mathcal{R}} g(\mathbf{r})(p_\Phi^{T=1}(\mathbf{r}) - p_\Phi^{T=0}(\mathbf{r}))d\mathbf{r}| \qquad (36) \\
=&B_\Phi\cdot Wass(p_\Phi^{T=1}, p_\Phi^{T=0}) \qquad (37)
\end{aligned}$$

Equation (35) is by the change of formula, $p_\Phi^{T=0}(\mathbf{r}) = p^{T=0}(\Psi(\mathbf{r}))J_\Psi(\mathbf{r})$, $p_\Phi^{T=1}(\mathbf{r}) = p^{T=1}(\Psi(\mathbf{r}))J_\Psi(\mathbf{r})$, where $J_\Psi(\mathbf{r})$ is the absolute of the determinant of the Jacobian of $\Psi(\mathbf{r})$. Equation (37) is by Definition 2. ☐

Proof of equation (5):

Proof.

$$\begin{aligned}
&\epsilon_{PEHE}(h,\Phi) \\
\le&2(\epsilon_{CF}(h,\Phi) + \epsilon_F(h,\Phi) - 2\sigma_y^2). \qquad (38) \\
\le&2(\epsilon_F^{T=1}(h,\Phi) + \epsilon_F^{T=0}(h,\Phi) + B_\Phi\cdot Wass(p_\Phi^{T=1}, p_\Phi^{T=0}) - 2\sigma_y^2). \qquad (39)
\end{aligned}$$

Inequality (38) is by equation (2) in Lemma 1. Inequality (39) is by equation 4 in Theorem 1. ☐

Proof of equation (6):

Proof.

$$\begin{aligned}
&\epsilon_{PEHE}(h,\Phi) \\
\le&\epsilon_{CF}(h,\Phi) + \epsilon_F(h,\Phi) + 2A_y \qquad (40) \\
\le&\epsilon_F^{T=1}(h,\Phi) + \epsilon_F^{T=0}(h,\Phi) + B_\Phi\cdot Wass(p_\Phi^{T=1}, p_\Phi^{T=0}) + 2A_y \qquad (41)
\end{aligned}$$

Inequality (40) is by equation (3) in Lemma 1. Inequality (41) is by equation 4 in Theorem 1. ☐

A.3 Proof of Theorem 2

We first introduce Lemma 2 that is useful for proving Theorem 2.

Lemma 2. *Let $\mathcal{G}$ that is defined in Definition 2 be the family of binary functions. Then we obtain* $\sup_{\eta\in\mathcal{H}}\left|\int_{\mathcal{S}}\eta(s)(p_1(s) - p_2(s))ds\right| = \frac{1}{2}d_{\mathcal{H}}(p_1,p_2)$.

Proof. Let $\mathbb{I}(\cdot)$ denotes an indicator function.

$$
\begin{aligned}
& d_{\mathcal{H}}(p_1, p_2) \\
=& 2 \sup_{\eta \in \mathcal{H}} \left| \int_{\eta(s)=1} (p_1(s) - p_2(s)) ds \right| \\
=& 2 \sup_{\eta \in \mathcal{H}} \left| \int_{\mathcal{S}} \mathbb{I}(\eta(s) = 1)(p_1(s) - p_2(s)) ds \right| \\
=& 2 \sup_{\eta \in \mathcal{H}} \left| \int_{\mathcal{S}} \eta(s)(p_1(s) - p_2(s)) ds \right| \qquad (42)
\end{aligned}
$$

The last equation is because an indicator function is also a binary function. □

Proof of equation (8):

Proof.

$$
\begin{aligned}
& \epsilon_{CF}(h, \Phi) - [(1-u) \cdot \epsilon_F^{T=1}(h, \Phi) + u \cdot \epsilon_F^{T=0}(h, \Phi)] \\
=& (1-u) \int_{\mathcal{R}} \ell_{h,\Phi}(\Psi(\mathbf{r}), 1)(p_\Phi^{T=0}(\mathbf{r}) - p_\Phi^{T=1}(\mathbf{r})) d\mathbf{r} + u \int_{\mathcal{R}} \ell_{h,\Phi}(\Psi(\mathbf{r}), 0)(p_\Phi^{T=1}(\mathbf{r}) - p_\Phi^{T=0}(\mathbf{r})) d\mathbf{r} \qquad (43) \\
\leq& (1-u) \int_{p_\Phi^{T=0} > p_\Phi^{T=1}} \ell_{h,\Phi}(\Psi(\mathbf{r}), 1)(p_\Phi^{T=0}(\mathbf{r}) - p_\Phi^{T=1}(\mathbf{r})) d\mathbf{r} \\
& + u \int_{p_\Phi^{T=1} > p_\Phi^{T=0}} \ell_{h,\Phi}(\Psi(\mathbf{r}), 0)(p_\Phi^{T=1}(\mathbf{r}) - p_\Phi^{T=0}(\mathbf{r})) d\mathbf{r} \qquad (44) \\
\leq& (1-u) K \int_{p_\Phi^{T=0} > p_\Phi^{T=1}} (p_\Phi^{T=0}(\mathbf{r}) - p_\Phi^{T=1}(\mathbf{r})) d\mathbf{r} + u \cdot K \int_{p_\Phi^{T=1} > p_\Phi^{T=0}} (p_\Phi^{T=1}(\mathbf{r}) - p_\Phi^{T=0}(\mathbf{r})) d\mathbf{r} \qquad (45) \\
=& (1-u) K \int_{\mathcal{R}} \mathbb{I}(p_\Phi^{t=0} > p_\Phi^{T=1})(p_\Phi^{T=0}(\mathbf{r}) - p_\Phi^{T=1}(\mathbf{r})) d\mathbf{r} \\
& + u \cdot K \int_{\mathcal{R}} \mathbb{I}(p_\Phi^{T=1} > p_\Phi^{T=0})(p_\Phi^{T=1}(\mathbf{r}) - p_\Phi^{T=0}(\mathbf{r})) d\mathbf{r} \\
\leq& (1-u) K \sup_{\eta \in \mathcal{H}} | \int_{\mathcal{R}} \eta(\mathbf{r})(p_\Phi^{T=1}(\mathbf{r}) - p_\Phi^{T=0}(\mathbf{r})) d\mathbf{r} | \\
& + u \cdot K \cdot \sup_{\eta \in \mathcal{H}} | \int_{\mathcal{R}} \eta(\mathbf{r})(p_\Phi^{T=1}(\mathbf{r}) - p_\Phi^{T=0}(\mathbf{r})) d\mathbf{r} | \qquad (46) \\
\leq& K \cdot \sup_{\eta \in \mathcal{H}} | \int_{\mathcal{R}} \eta(\mathbf{r})((p_\Phi^{T=1}(\mathbf{r}) - p_\Phi^{T=0}(\mathbf{r}))) d\mathbf{r} | \\
=& \frac{K}{2} d_{\mathcal{H}}(p_\Phi^{T=1}, p_\Phi^{T=0}) \qquad (47)
\end{aligned}
$$

Equation (43) is derived in the same way as equation (35). Equation (44) is by $\ell_{h,\Phi} \geq 0$ for all $\mathbf{r}$ and t. Inequality (45) is by the definition of K in Theorem 2. Inequality (46) is because an indicator function is also a binary function. Equation (47) is by Lemma 2. □

Proof of equation (9):

Proof.

$$
\begin{aligned}
& \epsilon_{PEHE}(h, \Phi) \\
\leq& 2(\epsilon_{CF}(h, \Phi) + \epsilon_F(h, \Phi) - 2\sigma_y^2) \qquad (48) \\
\leq& 2(\epsilon_F^{T=1}(h, \Phi) + \epsilon_F^{T=0}(h, \Phi) + \frac{K}{2} d_{\mathcal{H}}(p_\Phi^{T=1}, p_\Phi^{T=0}) - 2\sigma_y^2) \qquad (49)
\end{aligned}
$$

Inequality (48) is by equation 2 in Lemma 1. Inequality (49) is by equation 8 in Theorem 2. ☐

Proof of equation (10):

Proof.

$$\begin{aligned}
&\epsilon_{PEHE}(h,\Phi)\\
\leq&\epsilon_{CF}(h,\Phi)+\epsilon_{F}(h,\Phi)+2A_y \qquad (50)\\
\leq&\epsilon_{F}^{T=1}(h,\Phi)+\epsilon_{F}^{T=0}(h,\Phi)+\frac{K}{2}d_{\mathcal{H}}(p_{\Phi}^{T=1},p_{\Phi}^{T=0})+2A_y \qquad (51)
\end{aligned}$$

Inequality (50) is by equation 3 in Lemma 1. Inequality (51) is by equation 8 in Theorem 2. ☐

A.4 Additional Experimental details

Additional results on Twins Benchmark. To investigate the applicability of our model DIGNet to benchmark datasets beyond the commonly used IHDP benchmark, we conducted additional comparisons with several baseline models, including linear, tree, matching, and representation learning methods, on the Twins benchmark, as presented in Table 5.

The Twins dataset comprises records of twin births in the USA between 1989 and 1991. After preprocessing, each unit contains 30 covariates relevant to parents, pregnancy, and birth. The treatment $D=1$ indicates the heavier twin, while $D=0$ indicates the lighter twin. The binary outcome variable Y represents 1-year mortality. For more comprehensive details on this dataset and the limitation of IHDP, refer to Curth et al. (2021).

Notably, for ϵ_{ATE}, the simple linear or matching estimator performs best across different methods. On the other hand, when assessing ITE performance using the AUC of potential outcomes, representation learning models all demonstrate strong performance, with AUC values exceeding 0.800 on both training and test sets. The observation might stem from the fact that representation balancing models are based on ITE error bounds, rather than ATE error bounds, thereby optimizing for AUC instead of ϵ_{ATE}. Moreover, among all the models, our DIGNet achieves the second-best AUC results. The best results are achieved by MBRL, which involves the orthogonality information (similar to doubly robust estimators) in representation balancing. This, in turn, inspires us to explore ATE error bounds, or consider involving doubly robust methods in future research.

Table 5: Training- & test- set AUC & ϵ_{ATE} on Twins. Mean ± standard error of 100 runs.

	Training set		Test set	
	AUC	ϵ_{ATE}	AUC	ϵ_{ATE}
OLS/LR$_1$ Johansson et al. (2016)	$.660 \pm .005$	$.004 \pm .003$	$.500 \pm .028$	$.007 \pm .006$
OLS/LR$_2$ Johansson et al. (2016)	$.660 \pm .004$	$.004 \pm .003$	$.500 \pm .016$	$.007 \pm .006$
k-NN Crump et al. (2008)	$.609 \pm .010$	$.003 \pm .002$	$.492 \pm .012$	$.005 \pm .004$
BART Chipman et al. (2010)	$.506 \pm .014$	$.121 \pm .024$	$.500 \pm .011$	$.127 \pm .024$
CEVAE Louizos et al. (2017)	$.845 \pm .003$	$.022 \pm .002$	$.841 \pm .004$	$.032 \pm .003$
SITE Yao et al. (2018)	$.862 \pm .002$	$.016 \pm .001$	$.853 \pm .006$	$.020 \pm .002$
BLR Johansson et al. (2016)	$.611 \pm .009$	$.006 \pm .004$	$.510 \pm .018$	$.033 \pm .009$
BNN Johansson et al. (2016)	$.690 \pm .008$	$.006 \pm .003$	$.676 \pm .008$	$.020 \pm .007$
TARNet Shalit et al. (2017)	$.849 \pm .002$	$.011 \pm .002$	$.840 \pm .006$	$.015 \pm .002$
CFR-Wass (GNet) Shalit et al. (2017)	$.850 \pm .002$	$.011 \pm .002$	$.842 \pm .005$	$.028 \pm .003$
MBRL (Huang et al., 2022a)	$.879 \pm .000$	$.003 \pm .000$	$.874 \pm .001$	$.007 \pm .00q$
DIGNet (Ours)	$.874 \pm .001$	$.004 \pm .001$	$.871 \pm .001$	$.008 \pm .001$

Hyperparameters. In simulation studies, we ensure a fair comparison by fixing all the hyperparameters in all datasets across different models. The relevant details are stated in Table 6. In IHDP studies, to compare with the baseline model CFR-Wass (GNet), we remain the hyperparameters of INet, DGNet, DINet and the early stopping rule the same as those used in CFR-Wass Shalit et al. (2017). Since DIGNet is more complex than other four models, we adjust the hyperparameters of Φ_E, Φ_G, Φ_I, α_1, and α_2 for DIGNet as Shalit et al. (2017) do. The relevant details are stated in Table 7.

Table 6: Hyperparameters of different models in simulation studies.

	Φ_E	Φ_G	Φ_I	π	h^1	h^0	α_1	α_2	batchsize	iteration	learning rate	learning rate for π
Gnet	$(100, 100, 100, 100)$	–	–	–	$(100, 100)$	$(100, 100)$	0.1	–	100	300	$1e^{-3}$	–
Inet	$(100, 100, 100, 100)$	–	–	$(100, 100, 100)$	$(100, 100)$	$(100, 100)$	–	0.1	100	300	$1e^{-3}$	$1e^{-4}$
DGNet	$(100, 100, 100, 100)$	$(100, 100)$	–	–	$(100, 100)$	$(100, 100)$	0.1	–	100	300	$1e^{-3}$	–
DINet	$(100, 100, 100, 100)$	–	$(100, 100)$	$(100, 100, 100)$	$(100, 100)$	$(100, 100)$	–	0.1	100	300	$1e^{-3}$	$1e^{-4}$
DIGNet	$(100, 100, 100, 100)$	$(100, 100)$	$(100, 100)$	$(100, 100, 100)$	$(100, 100)$	$(100, 100)$	0.1	0.1	100	300	$1e^{-3}$	$1e^{-4}$

Table 7: Hyperparameters of different models in IHDP experiments.

	Φ_E	Φ_G	Φ_I	π	h^1	h^0	α_1	α_2	batchsize	iteration	learning rate	learning rate for π
Gnet	$(100, 100, 100, 100)$	–	–	–	$(100, 100, 100)$	$(100, 100, 100)$	1	–	100	600	$1e^{-3}$	–
Inet	$(100, 100, 100, 100)$	–	–	$(200, 200, 200)$	$(100, 100, 100)$	$(100, 100, 100)$	–	1	100	600	$1e^{-3}$	$1e^{-3}$
DGNet	$(100, 100, 100, 100)$	$(100, 100)$	–	–	$(100, 100, 100)$	$(100, 100, 100)$	1	–	100	600	$1e^{-3}$	–
DINet	$(100, 100, 100, 100)$	–	$(100, 100)$	$(200, 200, 200)$	$(100, 100, 100)$	$(100, 100, 100)$	–	1	100	600	$1e^{-3}$	$1e^{-3}$
DIGNet	$(100, 100, 100, 100, 100, 100)$	$(100, 100, 100)$	$(100, 100, 100)$	$(200, 200, 200)$	$(100, 100, 100)$	$(100, 100, 100)$	0.1	1	100	600	$1e^{-3}$	$1e^{-3}$

Analysis of training time and training stability. We record the time it took for different models to run through 100 IHDP datasets, and each model is trained within 600 epochs. Following Shalit et al. (2017), all models adopt the early stopping rule. We also record the average early stopping epoch on 100 runs and the actual time on 100 runs, where (actual time) = (total time) × (average early stopping epoch)/600. Not surprisingly, GNet took the least amount of time with 3096 seconds since the objective of GNet is the simplest. However, it is very interesting that the proposed methods, DGNet and DINet, are the first two to early stop. As a result, though DGNet and DINet have multi-objectives, they spent less actual training time but achieved better ITE estimation compared to GNet and INet. Since GNet and INet are actually DGNet and DINet with PPBR ablated, we find that PPBR component can help a model achieve better ITE estimates with less time. In addition, we find that DIGNet spent the longest time to optimize since it has the most complex objective. To further study the stability of the model training, we also plot the metrics $\sqrt{\epsilon_F}$, Wass, $\hat{d}_{\mathcal{H}}$, and $\sqrt{\epsilon_{PEHE}}$ for the first 100 epochs of each model on the first IHDP dataset. We find that the training process of DIGNet is stable, even steadier than GNet and INet. From this perspective, we haven't seen a difficulty of optimizing DIGNet.

Table 8: Training time records on 100 IHDP datasets.

Model	Time for 600 epochs	Avg early stopping	Actual time	$\sqrt{\epsilon_{PEHE}}$ on test set
GNet	3096s	240.61	1241s	0.77±0.18
INet	4042s	254.19	1712	0.72±0.11
DGNet	3775s	169.17	1064s	0.60±0.09
DINet	3212s	157.98	846s	0.60±0.11
DIGNet	4984s	226.76	1884s	0.45±0.04

A.5 Objectives of Different Models

Objective of GNet.

$$\min_{\Phi_E, h^t} \quad \mathcal{L}_y(\mathbf{x}, \mathbf{t}, \mathbf{y}; \Phi_E, h^t) + \alpha_1 \mathcal{L}_G(\mathbf{x}, \mathbf{t}; \Phi_E).$$

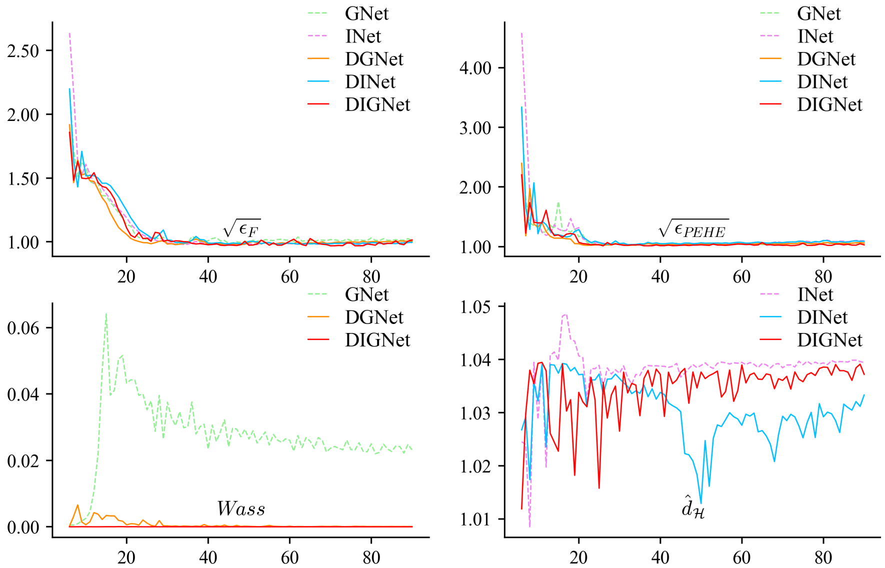


Figure 7: Training loss plots for the first 100 epochs on the first IHDP dataset.

Objective of INet.

$$\begin{aligned}
&\max_{\pi} \quad \alpha_2 \mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi_E, \pi), \\
&\min_{\Phi_E, h^t} \quad \mathcal{L}_y(\mathbf{x}, \mathbf{t}, \mathbf{y}; \Phi_E, h^t) + \alpha_2 \mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi_E, \pi).
\end{aligned}$$

Objective of DINet. Note that similar to DIGNet, the pre-balancing patterns are preserved by only updating Φ_I but fixing Φ_E in the second step.

$$\begin{aligned}
&\max_{\pi} \quad \alpha_2 \mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi_I \circ \Phi_E, \pi), \\
&\min_{\Phi_I} \quad \alpha_2 \mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi_I \circ \Phi_E, \pi), \\
&\min_{\Phi_E, \Phi_I, h^t} \quad \mathcal{L}_y(\mathbf{x}, \mathbf{t}, \mathbf{y}; \Phi_E \oplus (\Phi_I \circ \Phi_E), h^t).
\end{aligned}$$

Objective of DGNet. Note that similar to DIGNet, the pre-balancing patterns are preserved by only updating Φ_G but fixing Φ_E in the first step.

$$\begin{aligned}
&\min_{\Phi_G} \quad \alpha_1 \mathcal{L}_G(\mathbf{x}, \mathbf{t}; \Phi_G \circ \Phi_E), \\
&\min_{\Phi_E, \Phi_G, h^t} \quad \mathcal{L}_y(\mathbf{x}, \mathbf{t}, \mathbf{y}; \Phi_E \oplus (\Phi_G \circ \Phi_E), h^t).
\end{aligned}$$

Objective of DIGNet.

$$\begin{aligned}
&\min_{\Phi_G} \quad \alpha_1 \mathcal{L}_G(\mathbf{x}, \mathbf{t}; \Phi_G \circ \Phi_E), \\
&\max_{\pi} \quad \alpha_2 \mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi_I \circ \Phi_E, \pi), \\
&\min_{\Phi_I} \quad \alpha_2 \mathcal{L}_I(\mathbf{x}, \mathbf{t}; \Phi_I \circ \Phi_E, \pi), \\
&\min_{\Phi_E, \Phi_I, \Phi_G, h^t} \quad \mathcal{L}_y(\mathbf{x}, \mathbf{t}, \mathbf{y}; \Phi_E \oplus (\Phi_I \circ \Phi_E) \oplus (\Phi_G \circ \Phi_E), h^t).
\end{aligned}$$