Estimating Heterogeneous Treatment Effects: Mutual Information Bounds and Learning Algorithms

Xingzhuo Guo [*1,2] **Yuchen Zhang** [*1] **Jianmin Wang** [1] **Mingsheng Long** [1]

Abstract

Estimating heterogeneous treatment effects (HTE) from observational studies is rising in importance due to the widespread accumulation of data in many fields. Due to the selection bias behind the inaccessibility of counterfactual data, the problem differs fundamentally from supervised learning in a challenging way. However, existing works on modeling selection bias and corresponding algorithms do not naturally generalize to non-binary treatment spaces. To address this limitation, we propose to use mutual information to describe selection bias in estimating HTE and derive a novel error bound using the mutual information between the covariates and the treatments, which is the first error bound to cover general treatment schemes including multinoulli or continuous spaces. We then bring forth theoretically justified algorithms, the Mutual Information Treatment Network (MitNet), using adversarial optimization to reduce selection bias and obtain more accurate HTE estimations. Our algorithm reaches remarkable performance in both simulation study and empirical evaluation.

1. Introduction

Estimating heterogeneous treatment effects from observational data is a central problem in a broad variety of fields, including drug development in medical studies (Foster et al., 2011), policy analysis in economics (Heckman, 2000) and pollution assessment in ecology (Wang et al., 2016). A plethora of machine learning methods have been designed for inferring treatment effects with the increasing availability of observational data in these fields, which involve the use of tree models (Hill, 2011; Athey & Imbens, 2016; Wager & Athey, 2018), Gaussian processes (Alaa & Van Der Schaar, 2017; 2018) and neural networks (Johansson et al., 2016; Shalit et al., 2017).

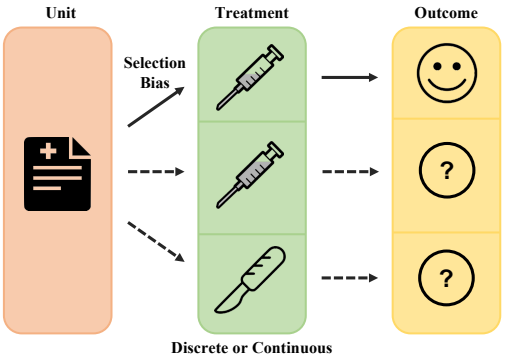

Figure 1. We study the learning bounds and algorithms for estimating heterogeneous treatment effect on *general treatment spaces*.

Following the Neyman-Rubin causal model (Rubin, 2005), the treatment effect for any individual is defined as the difference between the potential outcomes of receiving and not receiving a particular treatment (referred to as the intervention of being treated and controlled). In contrast to the standard supervised learning setups, for any individual only the factual outcome is observed while the counterfactual is not obtained. This is close to "learning from logged bandit feedback" (Strehl et al., 2010; Swaminathan & Joachims, 2015) in machine learning literature, with the distinction that we do not have access to the model generating the actions (or interventions in causal contexts). One simple way to overcome this barrier is to model the response surfaces of potential outcomes for both treated and control groups by some machine learning methods to enable causal estimation.

Different from randomized controlled trials (RCT) where interventions are randomly assigned to each individual, observational studies are usually subject to the sample *selection bias*, which is the main focus in this article. Sample selection bias is when the distributions differ as a result of an unknown sample rejection process (Quionero-Candela et al., 2009), causing *confounding* where the covariates affect both the factual treatment and the potential outcomes.

[*]Equal contribution [1]School of Software, Tsinghua University. [2]Institute for Interdisciplinary Information Sciences, Tsinghua University. Xingzhuo Guo <gxz19@mails.tsinghua.edu.cn>. Correspondence to: Mingsheng Long <mingsheng@tsinghua.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

For example, richer patients might better afford a certain medical treatment, so individuals measured with a larger value of wealth has a larger probability of receiving the treatment. Thus the data drawn from the control group would be less concentrated for the rich and vice versa. Ignoring selection bias could result in highly inaccurate models and biased average effects (Dorie et al., 2019).

For the case of binary interventions, the selection bias could be estimated using the propensity score or some distributional discrepancy in previous works. The propensity score is defined as the probability for an individual to receive the treatment. Matching (Rubin, 1973) and weighting (Rosenbaum & Rubin, 1983) methods are proposed to reduce selection bias based on the estimation and adjustment of propensity scores. Another measurement of selection bias is the distributional difference between factual and counterfactual data since estimating HTE requires predicting outcomes over the counterfactual distribution distinct from the observed one. Such a phenomenon contradicts the common *i.i.d.* assumption in standard supervised learning and is closely related to domain adaptation (Quionero-Candela et al., 2009; Cortes & Mohri, 2014; Zhang et al., 2019). Typically, representation learning methods (Johansson et al., 2016) are proposed to reduce such distribution shift.

Shalit et al. (2017) propose to use the integral probability metrics (IPM) including maximum mean discrepancy and Wasserstein distance to develop a learning error bound for this problem. To be specific, the expected precision in the estimation of heterogeneous effect (PEHE, Equation (5)) is controlled by the expected mean squared error and an IPM term. With notations defined in Section 2, their bound is interpreted as

$$\begin{aligned}\epsilon_{\text{PEHE}}[f] \leqslant 2(&\epsilon_{1\text{F}}[f] + \epsilon_{0\text{F}}[f] \\ &+ B_G \operatorname{IPM}_G(\mu_{\mathbf{X}\,|\,T=1}, \mu_{\mathbf{X}\,|\,T=0})),\end{aligned} \tag{1}$$

where IPM_G is the integral probability metric with respect to function class G and B_G is a constant related both to G and hypothesis space of the predictors. In their algorithm, Shalit et al. (2017) extract features from the original covariates and minimize the right-hand side of Equation (1) in the representation space.

Although they provide a novel view and rigorous theoretical guarantees for this problem, there are still a few drawbacks. First, their measurement could not be naturally extended to general intervention schemes as the counterfactual distribution could not be well-defined beyond binary settings. In addition, their theory relies on strong assumptions such as bijectivity and twice differentiability of the feature learner, which are not trivially satisfied in practical algorithms.

To overcome these shortcomings, we propose a novel theory and guided algorithms using mutual information to capture and control selection bias. Compared with previous works, our theory has two main advantages. First, mutual information is a versatile measurement for variables on general spaces, hence can be used beyond the tasks with binary treatments. Second, as an essential analytical tool, the use of mutual information allows us to obtain a more precise estimation of the dependencies which reflect the degree of selection bias. Inspired by our theory and the representation learning methods similar to Shalit et al. (2017), we map the original covariates to a new space and use representation from that space to predict potential outcomes in the hope that the learned features could be less dependent on the interventions. In particular, we use the mutual information as a regularization term to reduce the dependence between the learned features of covariants and the treatments. Our main contributions are summarized as follows:

- We develop a unified theory with rigorous error bounds using mutual information, which could not only deal with binary treatments, but also problems with general forms of treatments.
- We bring forward representation learning algorithms utilizing mutual information estimator with neural networks. The algorithms are theoretically justified and reach state-of-art performance in both simulation and semi-synthetic data.

2. Preliminaries

Our analysis hinges on a generalized version of the Neyman-Rubin potential outcomes model (Rubin, 2005). Let $\mathcal{X} \subset \mathbb{R}^d, \mathcal{T}, \mathcal{Y} \subset \mathbb{R}$ be a covariate space, a treatment space and an outcome space respectively. For a unit with features $\mathbf{x} \in \mathcal{X}$ and any treatment $t \in \mathcal{T}$, there is a potential outcome $Y_t \in \mathcal{Y}$. In particular, Y_0 is the potential outcome of receiving no treatment.

Definition 2.1. Let

$$m(\mathbf{x}, t) = \mathbb{E}[Y_t \,|\, \mathbf{x}]. \tag{2}$$

The *heterogeneous treatment effect (HTE)* or the *conditional average treatment effect (CATE)* of treatment t is defined as

$$\tau(\mathbf{x}, t) = m(\mathbf{x}, t) - m(\mathbf{x}, 0) = \mathbb{E}[Y_t - Y_0 \,|\, \mathbf{x}]. \tag{3}$$

Denote by $\mathcal{Y}^{\mathcal{T}}$ the space for random vectors or functions gathering all potential outcomes. Assume that the data $(\mathbf{x}, t, \mathbf{y})$ satisfies some underlying distribution μ on $\mathcal{X} \times \mathcal{T} \times \mathcal{Y}^{\mathcal{T}}$. Unfortunately, in real applications, for each unit $\mathbf{x}$ only one treatment t is assigned and the only observation y is the corresponding factual outcome $y = \mathbf{y}_t$ (also known as the *consistency assumption*). Under this setting, we are interested in learning a function $f : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ from the training data consisting of partially observed samples

$\mathcal{S} = \{(\mathbf{x}^{(i)}, t^{(i)}, y^{(i)}) : i = 1, \cdots, n\}$, such that

$$\widehat{\tau}_f(\mathbf{x}, t) = f(\mathbf{x}, t) - f(\mathbf{x}, 0) \tag{4}$$

serves as an estimation of the HTE $\tau(\mathbf{x}, t)$ on unit $\mathbf{x}$ and treatment t. Due to the missing of counterfactual outcomes and the selection bias, the learning of f differs from standard supervised learning. Next we will introduce criterons for the ideal and factual cases, respectively.

Assume all potential outcomes can be observed. In order to judge how well τ is approximated, Hill (2011) suggested a criterion called *precision in estimation of heterogeneous effect (PEHE)* for the binary treatment case:

$$\begin{aligned}\epsilon_{\text{PEHE}}[f] &= \mathbb{E}_{\mu_{\mathbf{X}}}[(\widehat{\tau}_f(\mathbf{X}, 1) - \tau(\mathbf{X}, 1))^2] \\ &= \mathbb{E}_{\mu_{\mathbf{X}}}[((f(\mathbf{X}, 1) - f(\mathbf{X}, 0)) \\ &\quad - (m(\mathbf{X}, 1) - m(\mathbf{X}, 0)))^2].\end{aligned} \tag{5}$$

In this paper, we study more general cases where treatment t is multinoulli or continuous, and extend the above notion by defining the generalized PEHE with respect to f:

Definition 2.2. (*Generalized PEHE*)

$$\begin{aligned}\epsilon_{\text{PEHE}}[f] &= \mathbb{E}_{\mu_{T|T\neq 0}}\mathbb{E}_{\mu_{\mathbf{X}}}[(\widehat{\tau}_f(\mathbf{X}, T) - \tau(\mathbf{X}, T))^2] \\ &= \mathbb{E}_{\mu_{T|T\neq 0}}\mathbb{E}_{\mu_{\mathbf{X}}}[((f(\mathbf{X}, T) - f(\mathbf{X}, 0)) \\ &\quad - (m(\mathbf{X}, T) - m(\mathbf{X}, 0)))^2],\end{aligned} \tag{6}$$

where $\mu_{\mathbf{X}}$ is the marginal distribution of $\mathbf{X}$ and $\mu_{T|T\neq 0}$ is the conditional distribution of T given that $T \neq 0$.

Remark 2.3. This extension is natural and reasonable since in real applications we care about the precision of effects not only for a particular treatment but also for various treatments chosen according to some population-level information. Thus, our goal is to learn a function f that minimizes $\epsilon_{\text{PEHE}}[f]$.

However, as we have discussed previously, only the factual outcomes are actually observed. We therefore define:

Definition 2.4. The (expected) *factual error* is the (expected) mean squared error of predicting potential outcomes for individuals with the observed treatment:

$$\epsilon_{\text{F}}[f] = \mathbb{E}_{\mu_{\mathbf{X},T}}[(f(\mathbf{X}, T) - m(\mathbf{X}, T))^2], \tag{7}$$

where the expectation is taken over the joint distribution of covariate $\mathbf{X}$ and treatment T.

Due to the selection bias mentioend previously, $\mathbf{X}$ and T might be highly correlated, and it is not enough to train f to only minimize $\epsilon_{\text{F}}[f]$ for a good HTE estimation. One general approach to handle selection bias is by creating a pseudo group which is approximately close to the interested group (Yao et al., 2021). Along this line, we will introduce a bound on $\epsilon_{\text{F}}[f]$ as a guidance for the learning of a feature representation to create pseudo groups.

In order to make the conditional causal effect identifiable, we need an extension of the *strong ignorability* condition (Rubin, 2005):

Assumption 2.5. (*Strong Ignorability*)

- (Ignorability) $\mathbf{Y} \perp\!\!\!\perp T \,|\, \mathbf{X} = \mathbf{x}$ for all $\mathbf{x}$, where $\mathbf{Y}$ denotes the random vector or function collecting all the potential outcomes.
- (Overlap) $\forall$ measurable set $\mathcal{M} \subset \mathcal{T}$, $0 < \mathbb{P}[T \in \mathcal{M}] < 1$ implies $0 < \mathbb{P}[T \in \mathcal{M} \,|\, \mathbf{X} = \mathbf{x}] < 1$ for all $\mathbf{x}$.

Remark 2.6. The first condition ensures that there would be "no-hidden confounding" and the second describes the overlap of distributions across treatment groups. Note that the validity of strong ignorability could be only determined by domain knowledge and prior understanding of the causal relationships and cannot be assessed solely from training data (Pearl, 2017).

3. Mutual Information Bounds for HTE

We provide a theoretical bound on the generalized PEHE $\epsilon_{\text{PEHE}}[f]$ in Equation (6) with the mutual information term to measure the impact of selection bias. We further extend it to a tightened bound through representation learning and discuss the correction of the selection bias. We start with a brief introduction to the notion of mutual information so as to present our main result.

Definition 3.1. In probability theory and information theory, the *mutual information (MI)* of two random variables Z_1 and Z_2 is a measure of the mutual dependence between them:

$$\begin{aligned}&\text{MI}(Z_1; Z_2) \\ &= \text{KL}(\mu_{Z_1,Z_2} \,\|\, \mu_{Z_1} \otimes \mu_{Z_2}) \\ &= \int_{\mathcal{Z}\times\mathcal{Z}} p_{Z_1,Z_2}(z_1, z_2) \log \frac{p_{Z_1,Z_2}(z_1, z_2)}{p_{Z_1}(z_1) p_{Z_2}(z_2)} \mathrm{d}z_1 \mathrm{d}z_2,\end{aligned} \tag{8}$$

where $\text{KL}(\cdot \,\|\, \cdot)$ denotes the Kullback-Leibler divergence. μ_{Z_1,Z_2} is the joint probability distribution of Z_1 and Z_2. μ_{Z_1} and μ_{Z_2} are the marginal distributions and $\mu_{Z_1} \otimes \mu_{Z_2}$ is their product distribution. $p_\star$ indicates the probability mass or density function corresponding to distribution $\mu_\star$. A larger divergence between the joint and product implies stronger dependence between Z_1 and Z_2. In particular, the mutual information vanishes for fully independent variables.

In the following, we present our main theorem on mutual information-based bounds for HTE estimation.

Theorem 3.2 (Mutual Information Bound)**.** *Let* $\pi = \mathbb{P}[T \neq 0]$. *Suppose the value of potential outcomes are bounded by a constant* B. *With the assumption of strong igorability and* $0 < \pi < 1$, *the expected PEHE could be controlled by the expected factual error together with the mutual information between* $\mathbf{X}$ *and* T*:*

$$\epsilon_{\text{PEHE}}[f] \leqslant \frac{1}{\pi(1-\pi)}\left(4B^2\sqrt{2\,\text{MI}(\mathbf{X};T)} + \epsilon_{\text{F}}[f]\right). \quad (9)$$

Proof. See Appendix A.1. □

Remark 3.3. In Theorem 3.2, the mutual information is used as a bridge to establish the binding relationship between PEHE and factual error. Formally, the PEHE could be bounded jointly by the factual error and the mutual information term between covariants $\mathbf{X}$ and treatments T. When there is no *selection bias*, the treatment T is independent on $\mathbf{X}$ and the bound degenerates into the factual error ($\mu_{\mathbf{X},T}$). Otherwise, the MI term in the bound could give a quantitive measurement of the error caused by the selection bias.

For multinoulli treatment spaces, there is another natural extension (Schwab et al., 2019; Kaddour et al., 2021), which is not the main focus of this paper. As a supplement, we briefly introduce this definition and the corresponding theory. The form is no more than a coefficient compared with Theorem 3.2, hence further analyses are all applicable.

Corollary 3.4 (Mutural Information Bound on Alternative Extension for Multinoulli Treatment Space)**.** *Given treatment space* $\mathcal{T} = \{1, 2, ..., k\}$, *define*

$$\epsilon_{\text{mPEHE}}[f] = \sum_{1\leq j<l\leq k} \mathbb{E}_{\mu_{\mathbf{x}}}\Big[\big((f(\mathbf{x},j) - f(\mathbf{x},l)) \\ (m(\mathbf{x},j) - m(\mathbf{x},l))\big)^2\Big], \quad (10)$$

then we have

$$\epsilon_{\text{mPEHE}}[f] \leq \frac{2}{k}\cdot\frac{1}{\pi^2}\left(\epsilon_F[f] + 4B^2\sqrt{2\,\text{MI}(\mathbf{X};T)}\right). \quad (11)$$

Further, we consider a corollary of this theorem for representation learning. We consider the hypothesis f with the form $f(\mathbf{x},t) = g(\phi(\mathbf{x}),t)$ where $\phi : \mathcal{X} \to \mathcal{R}$ is a feature learner from the covariate space $\mathcal{X}$ to a feature space $\mathcal{R}$ and $g : \mathcal{R}\times\mathcal{T} \to \mathcal{Y}$ is a predictor. For simplicity we denote

$$\epsilon_\star[\phi, g] = \epsilon_\star[f]. \quad (12)$$

Fixing the feature learner ϕ, we have the following collary.

Corollary 3.5 (Mutual Information Bound with Representation Learning)**.**

$$\epsilon_{\text{PEHE}}[\phi, g] \leqslant \frac{1}{\pi(1-\pi)}\left(4B^2\sqrt{2\,\text{MI}(\phi(\mathbf{X});T)} + \epsilon_{\text{F}}[\phi, g]\right). \quad (13)$$

Remark 3.6. Corollary 3.5 is a natural extension of Theorem 3.2 under a specific hypothesis space. Note that $\text{MI}(\phi(\mathbf{X});T) \leq \text{MI}(\mathbf{X};T)$ (see proof in Appendix A.3), so the bound in the corollary is tightened from the original version with arbitrary hypothesis f.

Corollary 3.5 implies that we can benefit from correcting selection bias by training the feature learner ϕ to minimize the mutual information. Now we show how far representation learning can go in correcting selection bias while preserving correlation of potential outcomes. To elaborate, consider the chain rule of mutual information:

$$\begin{aligned}&\text{MI}(\phi(\mathbf{X});T)\\ &= \text{MI}((\phi(\mathbf{X}),\mathbf{Y});T) - \text{MI}(\mathbf{Y};T\,|\,\phi(\mathbf{X}))\\ &= \text{MI}(\mathbf{Y};T) + \text{MI}(\phi(\mathbf{X});T\,|\,\mathbf{Y}) - \text{MI}(\mathbf{Y};T\,|\,\phi(\mathbf{X})).\end{aligned} \quad (14)$$

When ϕ preserves all correlation of potential outcomes, the ignorability can be proved to extend to $\text{MI}(\mathbf{Y};T\,|\,\phi(\mathbf{X})) = 0$ (see Appendix A.4.). Now that $\text{MI}(\mathbf{Y};T)$ is constant for a particular learning problem, minimizing $\text{MI}(\phi(\mathbf{X});T)$ is actually minimizing $\text{MI}(\phi(\mathbf{X});T\,|\,\mathbf{Y})$. The ideal state is $\text{MI}(\phi(\mathbf{X});T\,|\,\mathbf{Y}) = 0$ and $\text{MI}(\phi(\mathbf{X});T) = \text{MI}(\mathbf{Y};T)$, implying that the features that have impact on treatments rather than outcomes are removed.

In summary, our theory applies to *general treatment spaces* without any reliance on superfluous assumptions other than compactness. We further tighten this bound under representation learning for guiding HTE estimation algorithms.

4. Mutual Information Treatment Networks

According to our theory, we use mutual information (MI) as a regularization term in our objective function[1]:

$$\min_{\phi,g}\widehat{\epsilon}_{\text{F}}[\phi, g] + \alpha\widehat{\text{MI}}(\phi(\mathbf{X});T), \quad (15)$$

where α is a hyperparameter. $\widehat{\epsilon}_{\text{F}}$ and $\widehat{\text{MI}}$ are estimations of ϵ_{F} and MI from finite samples. We present a class of representation learning algorithms, coined *Mutual Information Treatment Network (MitNet)*, using this objective function. MitNet consists of three parts: a feature learner ϕ, a predictor g, and *a co-trained MI estimator* ψ. The design of ϕ and g basically follows Shalit et al. (2017). The feature learner is a multilayer network with ReLU activations. For discrete intervention settings with $\mathcal{T} = \{0, 1, ..., k\}$, we parameterize $g_t(\cdot) = g(\cdot, t)$ as $k+1$ separate heads. For the continuous treatment space with dosage $t \in [0, 1]$, we treat the dosage variable as an additional feature and concatenate $t\cdot\phi(\mathbf{x})$ with $\phi(\mathbf{x})$ as input to g to handle the dimension scaling issue.

[1]For computational convenience, we remove the square root function. Empirically, this does not show a significant difference.

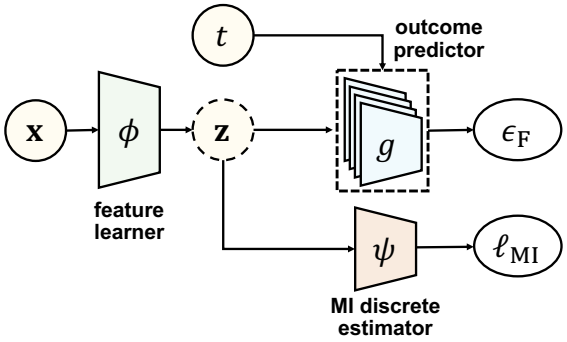

Figure 2. Architecture of MitNet for the direct esimation of MI.

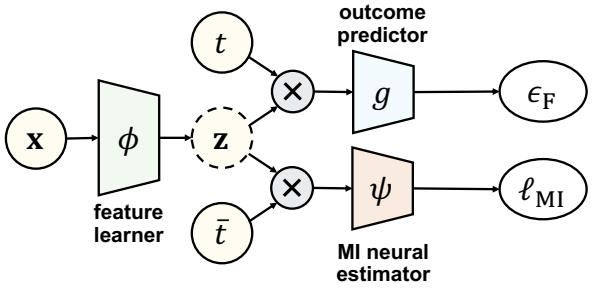

Figure 3. Architecture of MitNet for the neural estimator of MI.

On the design of the MI estimator ψ, despite the simple form of the objective function, estimating and minimizing MI have historically been difficult (Paninski, 2003). Recent works of Belghazi et al. (2018) offer a general-purpose Mutual Information Neural Estimator (MINE) that is compatible with representation learning models and could perform well if the estimator networks are chosen properly.

Therefore, we suggest two variants of MitNet with different types of MI estimators to deal with HTE estimation towards different treatment spaces. The first is a lightweight algorithm for discrete treatment spaces and computes the reduced form of MI by obtaining the equilibrium of a well-designed minimax game, akin to the generative adversarial network (GAN) (Goodfellow et al., 2014). The latter works for more general treatment spaces with better empirical performances by adopting variations of MINE in Belghazi et al. (2018) and also requires minimax optimization.

4.1. MitNet for Discrete Treatments

When the treatment space is discrete, we find that mutual information (MI) can be reduced to a generalized Jensen-Shannon divergence and be computed via adversarial optimization (Goodfellow et al., 2014). Let $\pi_j = \mathbb{P}[T = j]$, $j \in \{0, 1, \cdots, k\}$. Denoting by H the Shannon entropy and JS the generalized JS divergence, we have

$$\begin{aligned}
&\mathrm{MI}(\phi(\mathbf{X}); T) \\
&= \mathrm{H}\Big(\sum_{j=0}^{k} \pi_j \mu_{\phi(\mathbf{X})\,|\,T=j}\Big) - \sum_{j=0}^{k} \pi_j \,\mathrm{H}\left(\mu_{\phi(\mathbf{X})\,|\,T=j}\right) \\
&= \mathrm{JS}_{\pi_0, \pi_1, \cdots, \pi_k}\left(\mu_{\phi(\mathbf{X})\,|\,T=0}, \mu_{\phi(\mathbf{X})\,|\,T=1}, \cdots, \mu_{\phi(\mathbf{X})\,|\,T=k}\right).
\end{aligned} \tag{16}$$

Thereby, we estimate the mutual informtion through ψ as a discrete treatment discriminator that maximizes the classification loss of ψ by adversarial optimization. Specifically, for problems with a binary treatment space (see Figure 2), we parameterize $\psi : \mathcal{R} \to [0, 1]$ as a binary discriminator with the cross-entropy loss, which is trained to distinguish between the treated and control groups:

$$\ell_{\mathrm{MI}}[\phi, \psi] = -\sum_{i \text{ s.t. } t_i=1} \log \psi \circ \phi(\mathbf{x}_i) - \sum_{i \text{ s.t. } t_i=0} \log(1 - \psi \circ \phi(\mathbf{x}_i)). \tag{17}$$

Similarly, for problems with a multinoulli treatment space, we parameterize $\psi : \mathcal{R} \to [0, 1]^{k+1}$ as a treatment classifier with multi-class softmax output and the cross-entropy loss, which is trained to identify the treatment for features $\phi(\mathbf{x})$:

$$\ell_{\mathrm{MI}}[\phi, \psi] = -\sum_{i=1}^{n} \log \psi_{t_i}(\phi(\mathbf{x}_i)), \tag{18}$$

where $\psi_t(\phi(\mathbf{x}))$ denotes the predicted probability of $\mathbf{x}$ taking treatment t. Let

$$\widehat{\mathrm{MI}}(\phi(\mathbf{X}); T) = -\min_{\psi \in \Psi} \ell_{\mathrm{MI}}[\phi, \psi] + C, \tag{19}$$

where $C = -\sum_{j=0}^{k} \pi_j \log \pi_j$. Similar to Ganin et al. (2016), we train ψ to minimize ℓ_{MI} and ϕ to minimize our bound

$$\widehat{\epsilon}_{\mathrm{F}}[\phi, g] + \alpha \widehat{\mathrm{MI}}(\phi(\mathbf{X}); T). \tag{20}$$

It is important to note that when the equilibrium is reached,

$$\begin{aligned}
&\mathrm{MI}(\phi(\mathbf{X}); T) \\
&= \sum_{j=0}^{k} \pi_j \mathbb{E} \log \frac{p_{\phi(\mathbf{X})\,|\,T=j}(\phi(\mathbf{X}_j))}{p_{\phi(\mathbf{X})}(\phi(\mathbf{X}_j))} \\
&= \sum_{j=0}^{k} \pi_j \mathbb{E} \log \frac{p_{\phi(\mathbf{X}), T}(\phi(\mathbf{X}_j), j)}{\pi_j p_{\phi(\mathbf{X}_j)}(\phi(\mathbf{X}))} \\
&= \sum_{j=0}^{k} \pi_j \mathbb{E} \log \mathbb{P}[T = j \,|\, \phi(\mathbf{X}_j)] + C \\
&= \mathbb{E}_{\mu_{\mathbf{X}, T}} \log \mathbb{P}[T \,|\, \phi(\mathbf{X})] + C,
\end{aligned} \tag{21}$$

where $\mathbf{X}_j \sim \mu_{\mathbf{X}\,|\,T=j}$ for $j = 0, \cdots, k$. Thus when equilibrium, $\widehat{\mathrm{MI}}(\phi(\mathbf{X}); T)$ serves as an approximation of $\mathrm{MI}(\phi(\mathbf{X}); T)$.

4.2. MitNet with General Treatments

For general treatment spaces such as continuous dosage, MitNet adopts Mutual Information Neural Estimator (MINE)

(Belghazi et al., 2018) as the MI estimator. We first present the core ideas of MINE and its algorithms that have been adjusted for our settings. MINE originates from the following lemma on dual representation of KL divergence:

Lemma 4.1 (Donsker-Varadhan Representation(Donsker & Varadhan, 1983))**.** *Suppose μ and ν are two probability distributions on Ω. The KL divergence admits the following dual representation*

$$\mathrm{KL}(\mu \,\|\, \nu) = \sup_{\psi:\Omega\to\mathbb{R}} \mathbb{E}_{\mu}[\psi] - \log \mathbb{E}_{\nu}[e^{\psi}], \tag{22}$$

where the supremum is taken over all functions ψ such that the two expectations are finite.

In what follows, given a distribution μ, we denote by $\widehat{\mu}^{(n)}$ as the empirical distribution associated to n i.i.d. samples.

Definition 4.2. Denote Ψ as the family of functions $\psi : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ parametrized by neural networks with parameters θ. Then the *neural information measure* and the *mutual information neural estimator (MINE)* (Belghazi et al., 2018) are defined as

$$\mathrm{MI}_{\Psi}(X;Z) = \sup_{\psi\in\Psi} \mathbb{E}_{\mu_{X,Z}}[\psi] - \log \mathbb{E}_{\mu_X}\mathbb{E}_{\mu_Z}[e^{\psi}], \tag{23}$$

$$\widehat{\mathrm{MI}}_{\Psi}(X;Z) = \sup_{\psi\in\Psi} \mathbb{E}_{\widehat{\mu}^{(n)}_{X,Z}}[\psi] - \log \mathbb{E}_{\widehat{\mu}^{(n)}_{X}}\mathbb{E}_{\widehat{\mu}^{(n)}_{Z}}[e^{\psi}]. \tag{24}$$

Belghazi et al. (2018) show that $\widehat{\mathrm{MI}}_{\Psi}$ is strongly consistent. In other words, the estimator MI_{Ψ} can be close enough to MI if Ψ is well-chosen and the empirical $\widehat{\mathrm{MI}}_{\Psi}$ converges to the expected MI_{Ψ} with the increase of sample size n.

Following this theoretical guarantee, we adopt the MI estimation head ψ as a regressor for Equation 24 which takes $\phi(\mathbf{x})$ and t as input. Due to similar input and output spaces, we instantiate ψ as the same structure as the treatment effect predictor g (separate heads for discrete treatments and concatenating as an additional feature for continuous dosages). The special case of ψ for continuous treatments is shown in Figure 3. With this co-trained ψ, it suffices to maximize

$$\ell_{\mathrm{MI}}[\phi,\psi] = \frac{1}{n}\sum_{i=1}^{n} \psi\left(\phi(\mathbf{x}_i), t_i\right) - \log \frac{1}{n}\sum_{i=1}^{n} e^{\psi(\phi(\mathbf{x}_i),\bar{t}_i)}, \tag{25}$$

where $\bar{t}_i$ is resampled from distribution μ_T. The complete training process of MitNet for general treatments based on MINE (Belghazi et al., 2018) is shown in Algorithm 1.

Finally, noting that for discrete treatment spaces, MINE can be further improved by using the exact distribution of μ_T instead of an empirical one, reduced to a better loss function

$$\ell_{\mathrm{MI}}[\phi,\psi] = \frac{1}{n}\sum_{i=1}^{n} \psi\left(\phi(\mathbf{x}_i), t_i\right) - \sum_{j=0}^{k} \pi_j \log \frac{1}{n}\sum_{i=1}^{n} e^{\psi(\phi(\mathbf{x}_i),j)}. \tag{26}$$

Algorithm 1 MitNet with general treatments.

Initialize parameters θ of the MINE network ψ.
repeat
  Draw b minibatch samples from the joint distribution: $(\mathbf{x}_1, t_1), \cdots, (\mathbf{x}_b, t_b) \sim \mu_{\mathbf{X},T}$.
  Draw b samples from the marginal distribution of T: $\bar{t}_1, \cdots, \bar{t}_b \sim \mu_T$.
  Evaluate the lower bound of mutual information:

$$\ell_{\mathrm{MI}}[\phi,\psi_\theta] = \frac{1}{b}\sum_{i=1}^{b} \psi_\theta(\phi(\mathbf{x}_i), t_i) - \log \frac{1}{b}\sum_{i=1}^{b} e^{\psi_\theta(\phi(\mathbf{x}_i),\bar{t}_i)}.$$

  $\theta \leftarrow \theta + \nabla_\theta \ell_{\mathrm{MI}}[\phi,\psi_\theta]$.
until convergence.

5. Experiments

In this section, we propose detailed experiments to show the efficiency and generality of our methods. Code and data are available at https://github.com/thuml/MitNet.

5.1. Simulation Study

We conduct simulation studies to identify the regularization effect of mutual information (MI). We follow and extend the experiment settings introduced in (Colangelo et al., 2019). In all experiments, we generate $n = 2000$ i.i.d. samples from Gaussian distribution with covariates

$$\mathbf{X} = (X_1, \cdots, X_{1000}) \sim \mathcal{N}(0, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma} = \mathrm{diag}\{1, \cdots, 1\}$. We define two 1000-dimensional vectors θ and β: $\theta_j = \frac{1}{j}$ for $j \le 500$ and $\theta_j = 0$ for $j > 500$, while $\beta_j = 0$ for $j \le 500$ and $\beta_j = \frac{1}{j-500}$ for $j > 500$. It can be verified that θ and β are orthogonal.

The basic idea behind the data generation is to create treatment assignments that depend on the values of covariates. For this purpose, let the probability of subject $\mathbf{X}_i$ assigned to the t-th group be generated by a binomial distribution with parameter $\pi_{i,0} \in [0,1]$:

$$\pi_{i,0} = 0.5 + 0.1 * \mathrm{sign}(A_i D),$$

where A_i, D are defined as $A_i = (1, X_{i,1}, X_{i,2}, X_{i,3})$, $D = (0,1,1,1)'$. We consider a continuous treatment:

$$T_i = \Phi\left(\sum_{j=1}^{1000} X_{i,j}\theta_j\right) + 0.01\nu, \quad \Phi \text{ is the CDF of } \mathcal{N}(0,1)$$

with probability $1 - \pi_{i,0}$ and $T_i = 0$ with probability $\pi_{i,0}$. $\nu \sim \mathcal{N}(0,1)$ is Gaussian noise. The corresponding potential

outcome is given by

$$Y_i = \alpha\left(T_i + T_i^2 + T_i X_{i,1} + \sum_{j=1}^{1000} X_{i,j}\theta_j\right) + (1-\alpha)\left(\sum_{j=1}^{1000} \beta_j X_{i,j}\right) + 0.01\varepsilon.$$

Here $\alpha \in [0,1]$ is the dependence ratio. When α is as large as 1, the treatment effect Y highly depends on treatment T. When α approaches 0, Y is related to $\sum_{j=1}^{1000} \beta_j X_{i,j}$. As T_i is sampled by means $\Phi(\sum_{j=1}^{1000} X_{i,j}\theta_j)$ and θ, β are orthogonal, Y is independent to T in this case. We sample a group of datasets when α is set as $0.1, 0.3, 0.5, 0.7, 0.9$ respectively. We randomly split the generated dataset to train, validation and test parts according to the ratio of $63/27/10$. We use generalized PEHE (6) as the evaluation metric. We implement MitNet by constructing a three-layer neural network for each one of the feature learner, predictor and MI estimator.

We show the simulation results with respect to the dependence ratio α in Table 1. We provide both in-sample results evaluated on training data (In) and the out-of-sample results evaluated on test data (Out). When the factor α becomes larger, output $\mathbf{Y}$ will be highly dependent on treatment T. In the meantime $\mathrm{MI}(\mathbf{Y}; T)$ will increase the error bound and enlarge the difficulties of predicting Y under different treatments. We can see that the regularization of MI helps reduce PEHE over all variations of the dependence ratio. The convidence interval is reported in Appendix B.

Table 1. Results of Generalized PEHE on simulation datasets.

	α	0.1	0.3	0.5	0.7	0.9
w/o MI	In	0.085	0.26	0.46	0.67	0.77
	Out	0.092	0.28	0.51	0.74	0.85
w/ MI	In	**0.079**	**0.23**	**0.38**	**0.58**	**0.70**
	Out	**0.086**	**0.25**	**0.42**	**0.63**	**0.78**

5.2. Empirical Evaluation

We performed numerical experiments on three real-world semi-synthetic datasets with binary or multinoulli treatments in order to illustrate the efficacy and generality of MitNet.

IHDP The Infant Health and Development Program (IHDP) dataset (Hill, 2011) contains data from a randomized study on the impact of specialist visits on the cognitive development of infants. It consists of 747 children with 25 covariates describing properties of the children and their mothers. Children that receive home visits from specialists form the treated group while those who receive no visits form the control group. We use the treatment assignments and potential outcomes implemented as setting "A" in the NPCI package (Dorie, 2016), the same as Shalit et al. (2017).

News The News dataset was first proposed as a benchmark for counterfactual inference by Johansson et al. (2016) and was extended to the multinoulli treatment setting by Schwab et al. (2019). It consists of 5000 randomly sampled news articles from the NY Times corpus containing data on the opinion of media consumers on various news items. The units are different news items represented by word counts with dimension 2870, and the outcomes are the reader's opinion of the news item. k available treatments ($\mathcal{T} = \{1, \cdots, k\}$) represent various devices that could be used for viewing, e.g. smartphone, tablet, desktop, television or others. We assign the observed treatment and generate the potential outcomes according to the case k =4, 8 and κ =10 in Schwab et al. (2019) where κ is a treatment assignment bias coefficient.

TCGA The Cancer Genomic Atlas (TCGA) project collected gene expression data from various types of cancers in 9659 individuals with 20531 covariates (Weinstein et al., 2013). There were k =3 available clinical treatment options including medication, chemotherapy and surgery. Each medication is assigned with a continuous treatment dosage. A synthetic outcome function called the dose-response curve was used to simulate the risk of cancer recurrence after receiving either of the treatment options based on the real-world gene expression data. To model the outcomes, we follow the same approach as in Schwab et al. (2020) with the treatment assignment bias coefficient κ =10.

Baselines We examine MitNet together with several categories of benchmarks: traditional statistical methods (linear regression with the treatment as a feature (OLS/LR-1), linear regression with separate regressors for each treatment group (OLS/LR-2), inverse probability weighting (IPW), k-nearest neighbor matching (k-NN)), tree-based algorithms (BART (Chipman et al., 2010; Hill, 2011), Random Forests (Breiman, 2001), Causal Forests (Wager & Athey, 2018)), Gaussian processes (CMGP (Alaa & Van Der Schaar, 2017), NSGP (Alaa & Van Der Schaar, 2018)) and representation learning methods (GANITE (Yoon et al., 2018), BNN (Johansson et al., 2016), TARNet, CFR-Wasserstein (Shalit et al., 2017), PM (Schwab et al., 2019), DRNet (Schwab et al., 2020)). Note that some of these benchmarks are not applicable for News and TCGA (denoted as n.a. in Table 2) and DRNet is regarded as an extension of TARNet for the treatment scheme in TCGA.

Evaluation Although the generalized PEHE (6) could serve as a prior choice to measure the performance of HTE

Table 2. Results for HTE estimation on IHDP, News and TCGA. A lower metric indicates better performance.

Method	IHDP		News-4		News-8		TCGA
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{mPEHE}}}$	ϵ_{mATE}	$\sqrt{\epsilon_{\text{mPEHE}}}$	ϵ_{mATE}	$\sqrt{\epsilon_{\text{MISE}}}$
OLS/LR-1	5.1±.4	.90±.06	n.a.	n.a.	n.a.	n.a.	35.7±.1
OLS/LR-2	2.4±.1	.30±.02	43.2±2.4	23.1±2.5	40.3±2.2	19.6±2.4	n.a.
IPW	5.8±.3	.35±.03	39.5±3.3	15.6±2.7	38.4±3.7	12.9±1.6	26.3±.1
k-NN	2.7±.2	.79±.06	27.9±2.4	19.4±3.1	26.2±2.2	15.1±2.3	n.a.
BART (2011)	2.3±.2	.34±.02	26.4±3.1	17.1±3.5	25.8±2.7	14.8±2.6	n.a.
Random Forest (2001)	2.2±.2	.94±.06	26.6±3.0	18.0±3.2	23.8±2.1	12.4±2.3	16.3±.3
Causal Forest (2018)	2.8±.2	.48±.03	23.9±2.5	13.5±2.5	22.6±2.3	9.7±1.9	15.2±.1
CMGP (2017)	.76±.01	.29±.01	n.a.	n.a.	n.a.	n.a.	n.a.
NSGP (2018)	.64±.03	.23±.01	n.a.	n.a.	n.a.	n.a.	n.a.
GANITE (2018)	3.8±.8	.58±.07	24.5±2.3	13.8±2.7	23.6±2.5	11.2±2.8	15.4±.2
BNN (2016)	2.2±.2	.46±.03	n.a.	n.a.	n.a.	n.a.	n.a.
TARNet (2017) & DRNet (2020)	.95±.03	.28±.01	23.4±2.2	13.6±2.2	22.4±2.3	9.4±2.0	9.6±.0
CFR-Wasserstein (2017)	.76±.02	.27±.01	22.7±2.0	13.0±1.7	21.6±1.8	8.8±1.7	n.a.
PM (2019)	.80±.06	.31±.02	21.6±2.6	10.0±2.7	20.8±1.9	6.5±1.7	9.7±.2
MitNet (discrete)	.66±.04	.30±.02	19.3±1.7	13.3±2.1	19.1±1.9	9.8±2.2	n.a.
MitNet (general)	**.60±.03**	.25±.01	**19.2±2.2**	11.2±2.0	**18.9±2.0**	7.9±2.1	**9.3±.2**

estimation, it may be difficult to compute and does not directly apply to the News and TCGA datasets as there does not exist an additional control group. The treatment effect is directly computed as the difference of potential outcomes between a pair of treatment groups. Therefore, we introduce several variations of such metric with respect to each of the three datasets according to Schwab et al. (2019; 2020). For News, we consider the average of PEHE between every possible pair of treatments following Schwab et al. (2019):

$$\begin{aligned}\widehat{\epsilon}_{\text{mPEHE}}[f] &= \frac{2}{k(k-1)}\sum_{j=1}^{k}\sum_{l=1}^{j-1}\widehat{\epsilon}_{\text{PEHE},j,l}[f] \\ &= \frac{2}{k(k-1)n}\sum_{j=1}^{k}\sum_{l=1}^{j-1}\sum_{i=1}^{n}((f(\mathbf{x}_i,j)-f(\mathbf{x}_i,l)) \\ &\quad -(y_{i,j}-y_{i,l}))^2,\end{aligned}$$

where $y_{i,j}$ denotes the potential outcome for the i-th individual in the sample to receive treatment j.

For TCGA, we aim to measure a predictive model's ability to recover the dose-response curve across the range of dosage values and use the mean integrated square error (MISE) as defined in Schwab et al. (2020):

$$\widehat{\epsilon}_{\text{MISE}}[f] = \frac{1}{kn}\sum_{t=1}^{k}\sum_{i=1}^{n}\int_0^1 (f(x_i,(t,s))-v_i(t,s))^2\mathrm{d}s.$$

Note that the treatment in this problem is given by (t,s) and $v_i(t,s)$ denotes the potential outcome for the i-th individual given the t-th treatment with dosage s.

Implementation To enable a fair comparison, we chose the same hyperparameters including the batch size and number of hidden layers for TARNet, CFR-Wasserstein, PM, and MitNet. We adopt the model selection method mentioned in Schwab et al. (2019) to choose optimal α and learning rate η. We use the implementation from Dorie (2016) and Shalit et al. (2017) for previous methods on IHDP and the reported results of Schwab et al. (2019) and Schwab et al. (2020) on News-4/8 and TCGA if available. For IHDP we report the mean value and 95% confidence interval of results averaged over 1000 realizations of the potential outcomes. For News-4/8 we report the mean value and standard deviation of results in 50 repeated experiments. All experiments are performed with 63/27/10 train/validation/test splits. Early stopping is performed for network-based methods according to the value of the objective function on the validation set and all results are reported on the test set.

Analysis We show our methods can work well for general forms of treatments through three kinds of datasets when the types of treatments cover binary values (IHDP), multinoulli values (News) and continuous dosage (TCGA). From Table 2, we could see that MitNet for general cases ranked the first on all three datasets, while lightweight MitNet for discrete spaces reached the third on IHDP and second place on News-4/8. They consistently outperform all other network models on a large margin in terms of HTE estimation. Among all baselines, TARNet is a clean model that assigns each treatment a separate network without any selection bias reduction. CRF, PM and MitNet are developed to collect

selection bias based on the architecture of TARNet. Our methods based on mutual information control are the most flexible and effective compared with predecessors that could only get advantages on certain kinds of treatments.

6. Related Work

Machine learning for estimating HTE. Machine learning has longly been applied to estimating Heterogeneous Treatment Effect (HTE). Traditional non-parametric models (Hill, 2011; Hahn et al., 2020; Athey & Imbens, 2016; Wager & Athey, 2018; Alaa & Van Der Schaar, 2017; 2018) and learning schemes (Nie & Wager, 2021; Künzel et al., 2019) are designed to tackle this problem. Recently, neural networks have become a prevalent choice due to their flexibility and convenience (Louizos et al., 2017; Atan et al., 2018; Yoon et al., 2018; Zhang et al., 2020; Jesson et al., 2020; van Amersfoort et al., 2021; Jesson et al., 2021; Kaddour et al., 2021).

Handling selection bias. To cope with selection bias in observational data for HTE estimation, a series of methods from the perspective of data include Perfect Match (PM) (Schwab et al., 2019), context-aware importance weighting (Hassanpour & Greiner, 2019), and targeted regularization (Shi et al., 2019). For neural network-based approaches, the impact of selection bias can be mitigated by reducing the distribution shift in the representation space. Johansson et al. (2016) propose to add a regularization term called the discrepancy distance (Mansour et al., 2009). Shalit et al. (2017) propose counterfactual regression networks (CFR) with IPM regularization and Yao et al. (2018) add a local similarity preserving component. Zhang et al. (2020) propose to keep the distribution overlap. Jesson et al. (2020); van Amersfoort et al. (2021); Jesson et al. (2021) quantify the uncertainty. Our work goes along with representation learning methods and gives a novel theoretical analysis to quantify selection bias at the representation level.

Non-binary intervensions. While the vast majority of works deal with a single treatment, fewer account for non-binary interventions. For Average Treatment Effect (ATE) and Average Potential Outcome (APO), existing methods could be adapted to both multinoulli (Lopez et al., 2017; Scotina & Gutman, 2019; Tu et al., 2013) and continuous cases (Wu et al., 2022; Fong et al., 2018; Kallus & Santacatterina, 2019; Kennedy et al., 2017; Colangelo et al., 2019; Nie et al., 2021; Jesson et al., 2022). As in the case of HTE, although several machine learning methods designed for binary interventions could also be naturally applied (Gu et al., 2020; Schwab et al., 2019; 2020), few of them could deal with the problem of selection bias in the meantime. However, our proposed algorithms could reduce the selection bias in all settings of treatments.

HTE estimation and domain adaptation. As pointed out by Johansson et al. (2016), HTE estimation with the selection bias has technical connections with domain adaption. Both our error bound and the ones of IPM (Shalit et al., 2017) have similarities with generalization bounds in domain adaptation given by Mansour et al. (2009); Ben-David et al. (2010); Cortes & Mohri (2014); Zhang et al. (2019). Our algorithm is related to domain adversarial neural network (Ganin et al., 2016) in the adversarial paradigm.

7. Discussion

Additional Parameters. MitNet introduces a mutual information estimator in the model architecture, which takes additional time and memory over the fundamental regression model in the training process. Thanks to our extended theory to representation learning scenarios, we can append the mutual information estimator upon the learned representation. Hence the estimator is lightweight for direct estimation and has the same architecture as the prediction head for neural estimation. While there still exists extra time and memory consumption at the training stage, at the inference stage we can completely remove the mutual information estimator and thus eliminate such extra consumption.

Optimization Issues. MitNet adopts a minimax objective function for mutual information estimation, which is harder to optimize than plain models. Despite the remarkable performance gains in Table 2, we have not improved MitNet from the perspective of optimization. Noting that minimax optimization has been the foundation of many well-established algorithms such as generative adversarial networks (GANs), we believe our algorithms for estimating heterogeneous treatment effects are acceptable. However, a future research direction is to design optimization-friendly models or algorithms from our mutual information bounds.

8. Conclusion

We propose to use mutual information to describe selection bias in estimating HTE and derive a rigorous error bound using the mutual information between the covariates and treatments. We then bring forth theoretically justified algorithms called the Mutual Information Treatment Network (MitNet), which applies adversarial optimization to reduce selection bias and obtain a more accurate estimation of HTE.

Acknowledgements

This work was supported by the National Key Research and Development Plan (2021YFB1715200), National Natural Science Foundation of China (62022050 and 62021002), and Beijing Nova Program (Z201100006820041).

References

Alaa, A. M. and Van Der Schaar, M. Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes. *NeurIPS*, pp. 3425–3433, 2017.

Alaa, A. M. and Van Der Schaar, M. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *ICML*, pp. 209–222, January 2018.

Atan, O., Jordon, J., and Van Der Schaar, M. Deep-Treat - Learning Optimal Personalized Treatments From Observational Data Using Neural Networks. *AAAI*, pp. 2071–2078, 2018.

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, July 2016.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *ICML*, pp. 531–540, 2018.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001.

Chipman, H. A., George, E. I., and McCulloch, R. E. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, March 2010.

Colangelo, K., Lee, Y.-Y., et al. Double debiased machine learning nonparametric inference with continuous treatments. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, 2019.

Cortes, C. and Mohri, M. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

Donsker, M. D. and Varadhan, S. S. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36 (2):183–212, 1983.

Dorie, V. Npci: Non-parametrics for causal inference, 2016. URL https://github.com/vdorie/npci.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34:43–68, 2019.

Fong, C., Hazlett, C., Imai, K., et al. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NeurIPS*, pp. 2672–2680, 2014.

Gu, C., Lopez, M. J., and Hu, L. The estimation of causal effects of multiple treatments in observational studies using bayesian additive regression trees. *Statistical Methods in Medical Research*, 2020.

Hahn, P. R., Murray, J. S., Carvalho, C. M., et al. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 2020.

Hassanpour, N. and Greiner, R. CounterFactual Regression with Importance Sampling Weights. *IJCAI*, August 2019.

Heckman, J. J. Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics*, 115(1):45–97, 2000.

Hill, J. L. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, January 2011.

Jesson, A., Mindermann, S., Shalit, U., and Gal, Y. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33:11637–11649, 2020.

Jesson, A., Mindermann, S., Gal, Y., and Shalit, U. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, pp. 4829–4838. PMLR, 2021.

Jesson, A., Douglas, A., Manshausen, P., Meinshausen, N., Stier, P., Gal, Y., and Shalit, U. Scalable sensitivity and uncertainty analysis for causal-effect estimates of continuous-valued interventions. *arXiv preprint arXiv:2204.10022*, 2022.

Johansson, F. D., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *ICML*, pp. 4407–4418, January 2016.

Kaddour, J., Zhu, Y., Liu, Q., Kusner, M. J., and Silva, R. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems*, 34: 24841–24854, 2021.

Kallus, N. and Santacatterina, M. Kernel optimal orthogonality weighting: A balancing approach to estimating effects of continuous treatments. *arXiv preprint arXiv:1910.11972*, 2019.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79 (4):1229–1245, 2017.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Lopez, M. J., Gutman, R., et al. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, 32(3):432–454, 2017.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R. S., and Welling, M. Causal Effect Inference with Deep Latent-Variable Models. *NeurIPS*, 2017.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

Nie, L., Ye, M., Liu, Q., and Nicolae, D. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*, 2021.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 06 2021.

Paninski, L. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Pearl, J. Detecting latent heterogeneity. *Sociological Methods & Research*, 46(3):370–389, 2017.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. The MIT Press, 2009.

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

Rubin, D. B. Matching to remove bias in observational studies. *Biometrics*, pp. 159–183, 1973.

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

Schwab, P., Linhardt, L., and Karlen, W. Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks. *arXiv preprint arXiv:1810.00656*, 2019.

Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M., and Karlen, W. Learning Counterfactual Representations for Estimating Individual Dose-Response Curves. In *AAAI*, 2020.

Scotina, A. D. and Gutman, R. Matching algorithms for causal inference with multiple treatments. *Statistics in medicine*, 38(17):3139–3167, 2019.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*, pp. 4709–4718, January 2017.

Shi, C., Blei, D. M., and Veitch, V. Adapting Neural Networks for the Estimation of Treatment Effects. In *NeurIPS*, June 2019.

Strehl, A., Langford, J., Li, L., and Kakade, S. M. Learning from logged implicit exploration data. In *NeurIPS*, pp. 2217–2225, 2010.

Swaminathan, A. and Joachims, T. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

Tu, C., Koh, W. Y., and Jiao, S. Using generalized doubly robust estimator to estimate average treatment effects of multiple treatments in observational studies. *Journal of Statistical Computation and Simulation*, 83(8):1518–1526, 2013.

van Amersfoort, J., Smith, L., Jesson, A., Key, O., and Gal, Y. On feature collapse and deep kernel learning for single forward pass uncertainty. *arXiv preprint arXiv:2102.11409*, 2021.

Wager, S. and Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113:1228–1242, 2018.

Wang, Y., Kloog, I., Coull, B. A., Kosheleva, A., Zanobetti, A., and Schwartz, J. D. Estimating causal effects of long-term pm2. 5 exposure on mortality in new jersey. *Environmental health perspectives*, 124(8):1182–1188, 2016.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

Wu, X., Mealli, F., Kioumourtzoglou, M.-A., Dominici, F., and Braun, D. Matching on generalized propensity scores with continuous exposures. *Journal of the American Statistical Association*, 12 2022.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation Learning for Treatment Effect Estimation from Observational Data. *NeurIPS*, pp. 2633–2643, 2018.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data*, 15(5):1–46, 2021.

Yoon, J., Jordon, J., and Van Der Schaar, M. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. *ICLR*, February 2018.

Zhang, Y., Liu, T., Long, M., and Jordan, M. Bridging theory and algorithm for domain adaptation. In *ICML*, 2019.

Zhang, Y., Bellot, A., and Schaar, M. Learning overlapping representations for the estimation of individualized treatment effects. In *AISTATS*, pp. 1005–1014, 2020.

A. Mathemetical Proof

A.1. Proof of Theorem 3.2

In this subsection we prove our main theorem

$$\epsilon_{\text{PEHE}}[f] \leqslant \frac{1}{\pi(1-\pi)}\big(4B^2\sqrt{2\,\text{MI}(\mathbf{X};T)} + \epsilon_{\text{F}}[f]\big).$$

Proof. By Cauchy inequality and Pinsker's inequality

$$\begin{aligned}
&\epsilon_{\text{PEHE}}[f] && (27)\\
=&\mathbb{E}_{\mu_{T|T\neq 0}}\mathbb{E}_{\mu_{\mathbf{X}}}\left[((f(\mathbf{X},T)-f(\mathbf{X},0))-(m(\mathbf{X},T)-m(\mathbf{X},0)))^2\right] && (28)\\
=&\mathbb{E}_{\mu_{T|T\neq 0}}\mathbb{E}_{\mu_{\mathbf{X}}}\left[((f(\mathbf{X},T)-m(\mathbf{X},T))-(f(\mathbf{X},0)-m(\mathbf{X},0)))^2\right] && (29)\\
\overset{\text{Cauchy}}{\leqslant}&\Big(\frac{1}{\pi}+\frac{1}{1-\pi}\Big)\,\mathbb{E}_{\mu_{T\,|\,T\neq 0}}\mathbb{E}_{\mu_{\mathbf{X}}}\left[\pi(f(\mathbf{X},T)-m(\mathbf{X},T))^2+(1-\pi)(f(\mathbf{X},0)-m(\mathbf{X},0))^2\right] && (30)\\
=&\frac{1}{\pi(1-\pi)}\mathbb{E}_{\mu_T}\mathbb{E}_{\mu_{\mathbf{X}}}\left[(f(\mathbf{X},T)-m(\mathbf{X},T))^2\right] && (31)\\
\leqslant&\frac{1}{\pi(1-\pi)}\left(\mathbb{E}_{\mu_{\mathbf{X},T}}\left[(f(\mathbf{X},T)-m(\mathbf{X},T))^2\right]+4B^2\int_{\mathcal{X}\times\mathcal{T}}|p_{\mathbf{X}}(\mathbf{x})p_T(t)-p_{\mathbf{X},T}(\mathbf{x},t)|\,\mathrm{d}\mathbf{x}\mathrm{d}t\right) && (32)\\
\overset{\text{Pinsker}}{\leqslant}&\frac{1}{\pi(1-\pi)}\left(4B^2\sqrt{2\,\text{KL}(\mu_{\mathbf{X},T}\,\|\,\mu_{\mathbf{X}}\otimes\mu_T)}+\epsilon_{\text{F}}[f]\right) && (33)\\
\leqslant&\frac{1}{\pi(1-\pi)}\left(4B^2\sqrt{2\,\text{MI}(\mathbf{X};T)}+\epsilon_{\text{F}}[f]\right). && (34)
\end{aligned}$$

□

A.2. Proof of Corollary 3.4

In this subsection we prove the extended theory

$$\epsilon_{\text{mPEHE}}[f] \leq \frac{2}{k}\cdot\frac{1}{\pi^2}\left(\epsilon_F[f]+4B^2\sqrt{2\,\text{MI}(\mathbf{X};T)}\right)$$

Proof. We have

$$\begin{aligned}
&\epsilon_{\text{mPEHE}}[f]\\
=&\frac{2}{k(k-1)}\sum_{1\leq j<l\leq k}\mathbb{E}_{\mu_{\mathbf{X}}}\left[(f(\mathbf{X},j)-f(\mathbf{X},l))-(m(\mathbf{X},j)-m(\mathbf{X},l))^2\right]\\
=&\frac{2}{k(k-1)}\sum_{1\leq j<l\leq k}\mathbb{E}_{\mu_{\mathbf{X}}}\left[(f(\mathbf{X},j)-m(\mathbf{X},j))-(f(\mathbf{X},l)-m(\mathbf{X},l))^2\right]\\
\overset{\text{Cauchy}}{\leqslant}&\frac{2}{k(k-1)}\sum_{1\leq j<l\leq k}\frac{1}{\pi_j\pi_l}\mathbb{E}_{\mu_{\mathbf{X}}}\left[\pi_j(f(\mathbf{X},j)-m(\mathbf{X},j))^2+\pi_l(f(\mathbf{X},l)-m(\mathbf{X},l))^2\right]\\
\leq&\frac{2}{k(k-1)}\sum_{1\leq j<l\leq k}\frac{1}{\pi^2}\mathbb{E}_{\mu_{\mathbf{X}}}\left[\pi_j(f(\mathbf{X},j)-m(\mathbf{X},j))^2+\pi_l(f(\mathbf{X},l)-m(\mathbf{X},l))^2\right]\\
=&\frac{2}{k}\cdot\frac{1}{\pi^2}\mathbb{E}_{\mu_T}\mathbb{E}_{\mu_{\mathbf{X}}}\left[(f(\mathbf{X},t)-m(\mathbf{X},t))^2\right].
\end{aligned}$$

Until now, we obtain a similar term as (31), and following the same procedure we complete the proof. □

A.3. Proof of Mutual Information Decay

In this subsection we prove $\text{MI}(\phi(\mathbf{X});T)\leq\text{MI}(\mathbf{X};T)$.

Proof. According to the chain rule of mutual information, we have

$$\begin{aligned}&\mathrm{MI}(\phi(\mathbf{X});T) + \mathrm{MI}(\mathbf{X};T\,|\,\phi(\mathbf{X}))\\ =&\,\mathrm{MI}((\mathbf{X},\phi(\mathbf{X});T)\\ =&\,\mathrm{MI}(\mathbf{X};T) + \mathrm{MI}(\phi(\mathbf{X});T\,|\,\mathbf{X}).\end{aligned}$$

Since $\phi(\mathbf{X}) \perp\!\!\!\perp T\,|\,\mathbf{X}$, it follows $\mathrm{MI}(\phi(\mathbf{X});T|\mathbf{X}) = 0$, which completes the proof. □

A.4. Proof of Extended Ignorability

In this subsection we aim to discuss the extended ignorability

$$\mathbf{Y} \perp\!\!\!\perp T\,|\,\phi(\mathbf{X}),$$

when ϕ simply preserves correlation of potential outcomes, i.e., $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}\,|\,\phi(\mathbf{X})$.

Proof. From the ignorability, we have

$$\begin{aligned}&P(\mathbf{Y},T\,|\,\mathbf{X}) = P(\mathbf{Y}\,|\,\mathbf{X})P(T\,|\,\mathbf{X})\\ \Rightarrow&P(\mathbf{Y},T,\mathbf{X})P(\mathbf{X}) = P(\mathbf{Y},\mathbf{X})P(T,\mathbf{X})\\ \Rightarrow&P(\mathbf{Y},T,\mathbf{X},\phi(\mathbf{X}))P(\mathbf{X},\phi(\mathbf{X})) = P(\mathbf{Y},\mathbf{X},\phi(\mathbf{X}))P(T,\mathbf{X},\phi(\mathbf{X}))\\ \Rightarrow&P(\mathbf{Y},T,\mathbf{X}\,|\,\phi(\mathbf{X}))P(\mathbf{X}\,|\,\phi(\mathbf{X})) = P(\mathbf{Y},\mathbf{X}\,|\,\phi(\mathbf{X}))P(T,\mathbf{X}\,|\,\phi(\mathbf{X})).\end{aligned}$$

Since $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}\,|\,\phi(\mathbf{X})$, it follows

$$P(\mathbf{Y},\mathbf{X}\,|\,\phi(\mathbf{X})) = P(\mathbf{Y}\,|\,\phi(\mathbf{X}))P(\mathbf{X}\,|\,\phi(\mathbf{X})).$$

By combining the above two equations, we finally have

$$\begin{aligned}&P(\mathbf{Y},T,\mathbf{X}\,|\,\phi(\mathbf{X})) = P(\mathbf{Y}\,|\,\phi(\mathbf{X}))P(T,\mathbf{X}\,|\,\phi(\mathbf{X}))\\ \Rightarrow&P(\mathbf{Y},T\,|\,\phi(\mathbf{X})) = P(\mathbf{Y}\,|\,\phi(\mathbf{X}))P(T\,|\,\phi(\mathbf{X}))\\ \Rightarrow&\mathbf{Y} \perp\!\!\!\perp T\,|\,\phi(\mathbf{X}).\end{aligned}$$

□

Note that the preservation of confounders is a weaker condition than the bijectivity of ϕ, with which the extended ignorability holds trivially. For example, let $\mathbf{x} = (a,b)'$, $t = b$ and $\mathbf{y}_{t_0} = a + t_0$ for any t_0, then a possible ϕ is $\phi(\mathbf{x}) = a$, which is not a bijection.

B. Simulation Study with Confidence Interval

In this section we report the 95% confidence interval for the simulation study in Table 1. Results are shown in Table 3. Although introducing MI gives a slightly larger variance, MitNet still outperforms the baseline in a significant way.

Table 3. Results of Generalized PEHE on simulation datasets with 95% confidence intervals.

	α	0.1	0.3	0.5	0.7	0.9
w/o MI	In	0.085 ± 0.001	0.26 ± 0.003	0.46 ± 0.009	0.67 ± 0.012	0.77 ± 0.018
	Out	0.092 ± 0.003	0.28 ± 0.004	0.51 ± 0.010	0.74 ± 0.017	0.85 ± 0.021
w/ MI	In	**0.079 ± 0.003**	**0.23 ± 0.005**	**0.38 ± 0.014**	**0.58 ± 0.019**	**0.70 ± 0.029**
	Out	**0.086 ± 0.005**	**0.25 ± 0.007**	**0.42 ± 0.019**	**0.63 ± 0.028**	**0.78 ± 0.036**

C. License of Assets

Consider the datasets we used throughout experiments: IHDP and News are under MIT License. TCGA is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.