TauNet: Direct Learning for Heterogeneous Treatment Effect Estimation using Deep Neural Networks

Fujin Zhu, Jie Lu, Adi Lin, Junyu Xuan, Guangquan Zhang
Centre for Artificial Intelligence, Faculty of Engineering and IT, University of Technology Sydney
NSW 2007, Australia

{Fujin.Zhu, Junyu.Xuan, Guangquan.Zhang, Jie.Lu}@uts.edu.au, Adi.Lin@student.uts.edu.au

Abstract: Causal inference from observational data lies at the heart of education, healthcare, optimal resource allocation and many other decision-making processes. Most of existing methods estimate the target treatment effect indirectly by inferring the unobserved counterfactual outcome for every individual or the underlying treatment response functions. These indirect learning methods are subject to issues of model misspecification and high variability. As a complement of existing indirect learning methods, in this paper, we propose a direct learning framework, called TauNet, for causal inference using deep multi-task learning. It is based on a novel empirical τ-risk for learning the causal effect model of direct interest in a supervised learning scheme. In our proposed framework, the target treatment effect model is parametrized as a neural network and learned jointly with other auxiliary models in an end-to-end manner. Moreover, we extend the naïve TauNet into other two variants, TauNet-Simple and Taunet-Reg, by further incorporating shared representation learning layers and a propensity prediction regularizer. Experiments on simulated and real data demonstrate that the performances of the proposed methods match or are better than that of existing state-of-arts. Moreover, by learning the target treatment effect function directly, the proposed methods tend to obtain more stable estimates than existing indirect methods.

Keywords: Causal inference; treatment effect estimation; multi-task learning; neural networks

1 Introduction

Causal inference (aka treatment effect estimation) (Imbens & Rubin 2015) is the problem of estimating the treatment effect of an intervention on a target outcome variable and lies at the heart of many domains, including precise medicine (Atan, Jordon & Schaar 2018), computational advertisement (Bottou et al. 2013), algorithmic fairness and explainability (Madras et al. 2019) (Schwab & Karlen 2019), as well as social program evaluation (Hill 2011) . Taking precise medicine as an example, when doctors want to test some medicine on a disease, the gold standard method is to conduct a double-blind randomized control trial (RCT) where patients are randomly assigned to either the treated (taking medicine) or the control group (not taking medicine) and the treatment effect of the medicine is measured as the difference of the recovery outcomes between the two groups. However, in many real-world applications, RCTs may be expensive, unethical, or even impossible (Hernán & Robins 2018). As a result, researchers have mainly focused on observational studies that conduct causal inference using purely observational data (Rosenbaum & Rubin 1983) (Imbens & Rubin 2015).

Given the fundamental problem of observational causal inference (Imbens & Rubin 2015), i.e., for any individual, only the outcome corresponding to the received treatment can be observed, while outcomes under alternative treatment options are unobservable, the task of causal inference from observational data is fundamentally different from traditional supervised machine learning where the target labels of interest are available in the training data. This leads to causal inference from observational data a "missing data" problem. Moreover, in the observational data, both the treatment assignment and the observed outcome of an individual are influenced by some of his or her covariates which are called *confounders* in the literature. As a result, the underlying treatment assignment

mechanism that determines which treatment will be assigned as well as which potential outcome is missing is unknown and not random.

To tackle these challenges, traditional methods for causal inference have mainly focused on inferring the unobserved counterfactual outcomes of each individual via matching (Stuart 2010) (Zhu, Savage & Ghosh 2018) or weighting (Rosenbaum & Rubin 1983) (Yiu & Su 2018) the observed factual outcomes of other individuals, or inferring the potential outcome models by adjusted regression (Künzel et al. 2019). Recently, more advanced machine learning techniques, including Bayesian inference (Hill 2011) (Alaa & van der Schaar 2017), ensemble models (McCaffrey, Ridgeway & Morral 2004) (Grimmer, Messing & Westwood 2017), representation learning (Johansson, Shalit & Sontag 2016) (Shalit, Johansson & Sontag 2017) (Yao et al. 2018), deep generative modelling (Louizos et al. 2017) (Zhu et al. 2018), etc. have been adopted to build more flexible adjusted regression models for inferring the treatment response functions. In general, all these methods run in a multiple steps manner by first inferring either the unobserved counterfactual outcomes or the underlying treatment response functions, and then plugin these inferred quantities to estimate the target treatment effect.

In this paper, we propose a direct learning method to tackle both the missing data and selection bias issues existed in observational causal inference. Our method parametrizes the target causal effect function with deep neural networks (DNNs) and learns it via gradient-based optimization. Once learned, the fitted causal effect function can be directly used for estimating different causal effect quantities without a detour of estimating the unknown potential outcomes or the treatment assignment mechanism. This idea is quite intuitive and motivated by policy gradient methods for policy optimization from the reinforcement learning literature (Sutton & Barto 2012). Unlike other value-based algorithms, e.g., Q-learning, which learn an optimal policy indirectly by estimating the state-action function (i.e., the Q function) first, policy gradient methods parametrize the target policy directly and learn it using gradient-based optimization. We note that Wager et al. (Wager & Athey 2018) have also proposed to directly estimate the individual treatment effect (ITE) non-parametrically using random forests with an ad-hoc leaf splitting criterion. Using this causal forest method, individuals in each leaf can be regarded as being randomly assigned as if in an RCT. However, tree-based methods need manual feature engineering which are not as automatic as the neural network based methods proposed in this paper. We also compare it with our proposed methods in the experiment section empirically.

The main contributions of this paper are: Firstly, we categorize existing methods for causal inference into those based on learning the unobserved counterfactual outcome non-parametrically and those based on fitting the underlying treatment response functions. As a complementary of these indirect methods, we propose a novel empirical τ-risk and a direct learning framework for treatment effect estimation using deep multitask learning. Secondly, in the direct learning framework, we further extend the proposed neural network by adding shared representation layers and a new propensity prediction regularizer. As a result, we proposed all together three neural network architectures for treatment effect estimation in this paper. Lastly, we validate the proposed methods with comprehensive experiments on synthetic, semi-simulated and real-world datasets. Experiment results suggest that our proposed methods generally have a better or competitive performance compared to existing state-of-art methods. Moreover, estimations of our direct learning methods are generally more stable than their competitors since they estimate directly in an end-to-end manner rather than indirectly by a two-stage process.

The remainder of the paper is organized as follows: We introduce definitions, notations and formalize the causal inference problem in Section \ref{sec:II-definition}. In Section \ref{sec:III-related}, we give a brief review of related work. As the core section of this paper, Section \ref{sec:IV-direct} introduces our proposed direct learning framework for treatment effect estimation and three practical realizations using deep neural networks. Experiments on synthetic, semi-simulated and real-world datasets are described in Section \ref{sec:V-experiment}. Section \ref{sec:VI-conclusion} concludes the paper and discusses future work.

2 Problem Formulation

Consider an observational study consisting of n observations $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), i = 1, \dots, n\}$ of the variables (X, T, Y) drawn i.i.d. from some underlying distribution such that for each i, $\mathrm{x}_i \in \mathcal{X}$ denotes the pre-treatment covariates, $t_i \in \mathcal{T}$ the assigned treatment, and $y_i \in \mathcal{Y}$ the observed outcome. Take the data from a job training as an example, for an employee with covariate $\mathrm{x} \in \mathcal{X}$, the set of treatments $\mathcal{T}$ might be whether she joined a specific job training program, and the set of outcomes might be $\mathcal{Y} = [0, 10K]$ indicating her monthly salary in dollars. In this paper, we only consider the case of binary treatment, i.e., $\mathcal{T} = \{0,1\}$. Denote the treated group as $\mathcal{T}_1 = \{i: t_i = 1\}$ and the control group as $\mathcal{T}_0 = \{i: t_i = 0\}$. For an individual i, let $Y_i(t) \in \mathcal{Y}$ be his or her *potential outcome* under the treatment option t. The fundamental problem of causal inference is that only one of the two potential outcomes, $Y_i(0)$ and $Y_i(1)$, can be observed for a given individual, i.e., $y_i = t_i Y_i(1) + (1 - t_i) Y_i(0)$. In the machine learning literature, this kind of partial feedback is called "bandit feedback" (Swaminathan & Joachims 2015a, 2015b).

For an individual with covariate value x, denote the underlying *treatment response functions* if he or she is assigned into the treated group and the control group as $\mu_1(\mathrm{x})$ and $\mu_0(\mathrm{x})$ respectively. In the language of Pearl's do-calculus (Pearl 2000), they are defined as

$$\mu_t(\mathrm{x}) = \mathbb{E}[Y_i | X_i = \mathrm{x}, do(T_i = t)] = \mathbb{E}[Y_i(t) | X_i = \mathrm{x}], \qquad t = 0,1$$

where $do(T_i = t)$ is the do-operator meaning to "set" the treatment as t rather than "seeing" the treatment t. In many real –world applications, we are interested in the covariate-specific treatment effect of the treatment t on the outcome, which is defined as the expected difference between the two potential treatment responses, i.e.,

$$\tau(\mathrm{x}) = \mu_1(\mathrm{x}) - \mu_0(\mathrm{x}) = \mathbb{E}[Y_i(1) | X_i = \mathrm{x}] - \mathbb{E}[Y_i(0) | X_i = \mathrm{x}] \qquad (1)$$

This is called the *conditional average treatment effect* (CATE), *individual treatment effect* (ITE) or *heterogeneous treatment effect* (HTE) in the causal inference literature (Imbens & Rubin 2015), and is intrinsically important in settings where we want to evaluate the efficiency of some policy and make personalized recommendations. We can use it to estimate the average treatment effect (ATE) via $ATE = \mathbb{E}[\tau(\mathrm{x}_i)]$ and the average treatment effect on the treated (ATT) via $ATT = \mathbb{E}[\tau(\mathrm{x}_i) | t_i = 1]$.

Despite of its importance, treatment effect estimation from purely observational data is fundamental impossible without causal assumptions since we can never observe both treatment responses for any individual. In order to make the individual treatment effect identifiable, we make the following assumptions as usually done in the causal inference literature (Imbens & Rubin 2015).

Assumption 1 (*Consistency*). For each individual X_i, the potential outcome under treatment $t \in \mathcal{T}$, $Y_i(t)$ is equal to the observed outcome if the actual treatment is t. That is, $Y_i(t) = Y_i$ if $T_i = t$.

Assumption 2 (*Ignorability*). For each individual X_i, the potential outcome variables $Y_i(t), t \in \mathcal{T}$ are statistically independent of the treatment actually taken. That is, $Y_i(t) \perp\!\!\!\perp T_i | X_i$ for all $i = 1,2, \dots, n$.

The *Consistency* assumption is by principle, and the *Ignorability* assumption means that there exist no unobserved confounders. Ignorability is generally uncheckable from data only and must be determined by domain knowledge. In practice, we also need the following positivity assumption to guarantee enough randomness in the data-generating process so that unobserved counterfactuals can be estimated from the observed data.

Assumption 3 (*Common support / Positivity*). The treatment propensity is positive for any covariate $\mathrm{x} \in \mathcal{X}$, i.e., $0 < P(t = 1 | \mathrm{x}) < 1$.

3 Related Work

For the task of causal inference from observational data, classic methods have focused on estimating the ATE through variants of propensity score matching or weighting (Imbens & Rubin 2015). More recent works tackle the problem of HTE estimation using standard supervised learning techniques (Künzel et al. 2019) (Nie & Wager 2017) (Wager & Athey 2018), Bayesian inference (Hill 2011) (Alaa & van der Schaar 2017) (Lin et al. 2020), representation learning (Johansson, Shalit & Sontag 2016) (Shalit, Johansson & Sontag 2017) (Yao et al. 2018) and deep generative models (Louizos et al. 2017) (Zhu et al. 2018). For a comprehensive overview of this topic, we refer the readers to (Guo et al. 2018) and (Yao et al. 2020). In this section, we present a brief review of existing methods that is closely related with our methodology.

3.1 Non-parametric Methods for Causal Inference

To tackle the fundamental problem of causal inference from observational data, many non-parametric methods attempt to transform the collected observational data to mimic a balanced one as from a randomized experiment (Imbens & Rubin 2015). Non-parametric methods do not model the relation between the pre-treatment covariates, treatment, and outcome. Instead, they realize treatment effect estimation by inferring the unobserved counterfactual outcomes using statistical techniques such as matching or weighting. On one hand, matching methods assume that similar individuals should have similar treatment outcomes, and estimate the unobserved counterfactual outcomes by matching every individual with individuals in the counterpart group. Examples of matching estimators include nearest neighborhood matching (Crump et al. 2008), propensity score matching (Stuart 2010), kernel matching (Zhu, Savage & Ghosh 2018) and optimal matching (Kallus 2017). On the other hand, inverse propensity weighting (Rosenbaum & Rubin 1983) (Zhu et al. 2020) calculates the expectation of a potential outcome using the weighted mean of observed factual outcomes in the corresponding group. The identifiability of potential outcomes is realized by inverse probability weights derived from the estimated treatment propensities.

While matching methods rely on an appropriate neighborhood metric to find a set of neighbors, weighting methods are generally designed for estimating average treatment effects and are not straightforward for heterogeneous treatment effect estimation. Recently, tree and forest based models (Athey & Imbens 2016) (Wager & Athey 2018) have been regarded as adaptive neighborhood metrics and used for non-parametric causal inference. In these estimators, ad-hoc node splitting rules targeting at treatment effect estimation are designed, trees are trained to predict propensity scores and leaves are used to predict treatment effects.

3.2 Causal Inference Based on Treatment Response Modelling

Besides inferring the unobserved counterfactual outcomes, another group of methods solve the problem of causal inference as a supervised learning problem. They fit the two treatment response functions $\mu_0(\mathrm{x})$ and $\mu_1(\mathrm{x})$ by supervised learning models (e.g., linear regression, random forest and neural networks). Then for an individual with covariates x, the treatment effect is estimated transductively via $\hat{\tau}(\mathrm{x}) = \mu_1(\mathrm{x}) - \mu_0(\mathrm{x})$. This is called simulated twins, G-computation, outcome regression or counterfactual inference in the literature (Hernán & Robins 2018) (Zhu et al. 2018) (Alejandro Schuler 2018). While classical outcome regression methods have mainly assumed generalized linear model, in recent years, advanced machine learning models such as Bayesian Additive Regression Trees (BART) (Hill 2011), multi-task Gaussian process (Alaa & van der Schaar 2017) and neural networks (Zhu et al. 2018) (Alejandro Schuler 2018) have also been adopted for treatment response modelling.

In practice, we can treat the treatment indicator $t \in \{0,1\}$ as a function indicator and learn separate treatment response models for each treatment. This is called *T-learning* (T for "two models" or "twins")

(Künzel et al. 2019). Alternatively, we can regard t as just another covariate and define treatment responses under different treatments as a single function, $\mu(\mathrm{x}, t)$. This is called *S-learning* (S for "single model") (Künzel et al. 2019). Besides T-learning and S-learning, Künzel et al (Künzel et al. 2019) also proposed *X-learning* that estimates the two treatment response functions $\mu_0(\mathrm{x})$ and $\mu_1(\mathrm{x})$ as in T-learning first, and then impute the estimated ITEs for the treated group $\mathcal{T}_1$ using $\hat{\tau}(\mathrm{x}_i) = y_i - \mu_0(\mathrm{x}_i)$ and for the control group $\mathcal{T}_0$ using $\hat{\tau}(\mathrm{x}_i) = \mu_1(\mathrm{x}_i) - y_i$. Lastly, X-learning learns the target treatment effect function $\tau(\mathrm{x})$ with the imputed ITEs in a supervised manner.

In treatment response modelling, the key is to obtain a good estimate of the underlying treatment response functions $\mu_0(\mathrm{x})$ and $\mu_1(\mathrm{x})$. The target HTE $\tau(\mathrm{x})$ is estimated indirectly by a two-stage procedure. As a result, treatment response functions fitted to minimize the prediction error for the observed outcomes are not guaranteed to produce accurate treatment effect estimation.

3.3 Causal Inference Based on Representation Learning

In recent years, deep learning techniques have been adopted for developing treatment effect estimation methods (Johansson, Shalit & Sontag 2016) (Louizos et al. 2017) (Zhu et al. 2018). By formulating counterfactual inference as a domain adaptation problem, Johansson et al (Johansson, Shalit & Sontag 2016) proposed a balancing counterfactual regression framework for treatment effect estimation. By virtue of the automatic representation learning ability of neural networks, Johansson et al. (Johansson, Shalit & Sontag 2016) proposed the Balancing Linear Regression (BLR) and Balancing Neural Network (BNN). Both of them learn a single treatment response function $\mu(\mathrm{x}, t) = h(\phi(\mathrm{x}), t)$ on top of a shared representation $\phi(\mathrm{x})$ of the pre-treatment covariates that attempts to minimize the linear discrepancy between the two groups of individuals.

Within the same framework, Shalit et al. (Shalit, Johansson & Sontag 2017) argued that the single treatment response function used in BLR and BNN may lose the influence of the scaler treatment indicator t on the shared high-dimensional representation during training. To avoid this issue, they proposed two neural networks, the Target-Agnostic Representation Network (TARNet) and the Counterfactual Regression Network (CFR), to learn two separate outcome regression models $h_0\big(\phi(\mathrm{x})\big)$ and $h_1\big(\phi(\mathrm{x})\big)$ on top of the shared representation layers $\phi(\mathrm{x})$. Moreover, to realize the goal of covariate balance in the representation space, CFR constraint the learning of the shared representation using integral probability metrics (IPMs) (Sriperumbudur et al. 2012). Later, Based on the CFR method, Yao et al. (Yao et al. 2018) proposed the local similarity-preserved individual treatment effect method that is able to learn local similarity preserved representation.

In the context of deep multi-task learning (Ruder 2017), Alaa et al. (Alaa, Weisz & Schaar 2017) proposed a deep counterfactual network for treatment effect estimation. The deep counterfactual network consists of a potential outcomes network and a propensity network. While the potential outcomes network is a deep multitask network with a set of shared representation learning layers and two outcome prediction heads, the propensity network is a feed-forward network and trained separately to estimate the treatment propensities $p(t_i|\mathrm{x}_i)$. Recently, the DragonNet proposed in (Shi, Blei & Veitch 2019) uses a three-headed multi-task neural network architecture for predicting propensity scores and potential outcomes simultaneously. The network is trained in an end-to-end manner and the obtained potential outcomes and propensity score functions can be used for downstream treatment effect estimation.

In general, existing causal inference methods based on representation learning also fall into the treatment response modelling framework introduced in the last section. They learn a single or two treatment response models for downstream treatment effect estimation and share information between the treated and control groups via the shared representation learning layers.

4 Methodology

As we can see from the last section, in treatment response modelling, the target treatment effect quantities are estimated in an indirect manner by first fitting the underlying treatment response functions. Take *T-learning* for example, the observational data $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), i = 1, \dots, n\}$ is divided into the treated subset $\mathcal{D}_1 = \{(\mathrm{x}_i, y_i), i \in \mathcal{T}_1\}$ and the control subset $\mathcal{D}_0 = \{(\mathrm{x}_i, y_i), i \in \mathcal{T}_0\}$. By parameterizing the two treatment response models $\mu_0(\cdot)$ and $\mu_1(\cdot)$ as $\mu_0(\mathrm{x}; \beta_0)$ and $\mu_1(\mathrm{x}; \beta_1)$, a naive method for learning these two models is to minimize the following *empirical μ-prediction risks* over $\mathcal{D}_0$ and $\mathcal{D}_1$ respectively,

$$\mathcal{L}_{\mu_0} = \frac{1}{n_0} \sum_{i \in \mathcal{T}_0} L(\mu_0(\mathrm{x}_i), y_i), \qquad \mathcal{L}_{\mu_1} = \frac{1}{n_1} \sum_{i \in \mathcal{T}_1} L(\mu_1(\mathrm{x}_i), y_i) \qquad (2)$$

Possible loss functions are the L_2 loss $L(\mu(\mathrm{x}_i), y_i) = (\mu(\mathrm{x}_i) - y_i)^2$ for continuous outcomes and the log-loss $L(\mu(\mathrm{x}_i), y_i) = -y_i \log \mu(\mathrm{x}_i) - (1 - y_i) \log\big(1 - \mu(\mathrm{x}_i)\big)$ for binary outcomes. Obviously, since the above μ-prediction risks are not targeted at our goal of estimating causal effects, treatment effect estimators building on them may not be reliable. In this section, we propose a direct learning framework for treatment effect estimation.

4.1 The Empirical τ-Risk for Direct Learning of HTE

Denote the interested HTE function $\tau: \mathcal{X} \to \mathbb{R}$ by a neural network $\tau(\cdot; \theta)$ with parameters $\theta \in \mathbf{\Theta}$. To train the neural network, we need a loss function that is able to guide the algorithm to update the model parameters. Suppose we have an oracle of the true ITE $\tau_i^* = \tau^*(\mathrm{x}_i) \triangleq \mu_1(\mathrm{x}_i) - \mu_0(\mathrm{x}_i)$ for every individual in the observational data $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), i = 1, \dots, n\}$. Then the optimal parameters $\theta^* \in \mathbf{\Theta}$ can be obtained via supervised learning by minimising the following PEHE-risk:

$$\mathcal{L}_\tau^*(\theta) = \frac{1}{n} \sum_{i=1}^{n} [\tau^*(\mathrm{x}_i) - \tau(\mathrm{x}_i; \theta)]^2 \qquad (3)$$

This risk is known as the *precision in estimating heterogeneous effect* (-PEHE), and is commonly used to quantify the "goodness" of a as an estimate of the true HTE model $\tau^*(\mathrm{x})$. A fundamental challenge that arises when learning the "PEHE-optimal" model $\tau(\mathrm{x}; \theta^*)$ is that we cannot compute the empirical PEHE for a particular $\theta \in \mathbf{\Theta}$ since we do not have the oracle of the true ITEs. In order to overcome this problem and to optimize the neural network parameters θ using the observed data $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), i = 1, \dots, n\}$, we need to bridge the target estimation $\tau(\mathrm{x}_i; \theta)$ with the observed outcome y_i for each individual. Obviously, if we have an oracle the counterfactual outcome y_i^{cf} for each individual x_i, several lines of algebra then imply that

$$\mathbb{E}[y_i] = (2t_i - 1)\tau(\mathrm{x}_i; \theta) + \mathbb{E}\big[y_i^{cf}\big] \qquad (4)$$

This permits us to optimize the THE model $\tau(\mathrm{x}; \theta)$ in a supervised learning manner by

$$\theta^* \in \underset{\theta \in \mathbf{\Theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \Big[y_i - \Big((2t_i - 1)\tau(\mathrm{x}_i; \theta) + y_i^{cf} \Big) \Big]^2 \qquad (5)$$

where y_i^{cf} is the counterfactual outcome of individual i and is defined by

$$\mathbb{E}\big[y_i^{cf}\big] = t_i \mu_0(\mathrm{x}_i) + (1 - t_i)\mu_1(\mathrm{x}_i) \qquad (6)$$

In classical non-parametric methods, the counterfactual outcome y_i^{cf} is usually obtained by matching with individuals in the counterpart group. In this paper, we obtain it by introducing a counterfactual prediction component that consists of the two auxiliary outcome prediction functions $\mu_0(\cdot)$ and $\mu_1(\cdot)$

in the training process. Substituting the definition Eq. (6) into Eq. (5), we define the following empirical τ-risk as a proxy of the PEHE-risk and use it for learning the HTE model:

$$\mathcal{L}_\tau(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{(2t_i - 1)\tau(\mathrm{x}_i;\theta) + t_i\mu_0(\mathrm{x}_i) + (1 - t_i)\mu_1(\mathrm{x}_i) - y_i\}^2 \quad (7)$$

Figure 1 illustrates the neural network architecture of our proposed direct learning framework. Since the method parameterizes our interested HTE model $\tau(\mathrm{x};\theta)$ using neural networks and learns the model parameters via the empirical τ-risk, we name it the TauNet.

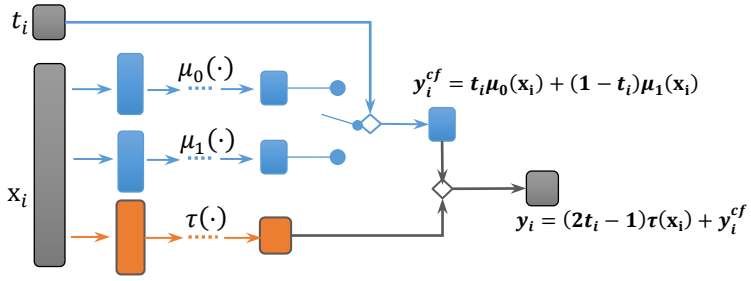

Figure 1. Neural network architecture of TauNet. The HTE prediction layers are orange. The green nodes make up the auxiliary counterfactual prediction component consisting of two treatment responses $\mu_0(x_i)$ and $\mu_1(x_i)$. The diamond toggle is switched according to the observed treatment.

Till now, we have introduced how to calculate the empirical τ-risk for a specific HTE function. This is realized by incorporating an auxiliary counterfactual outcome prediction component to bridge the target HTE estimate with the observed outcome for any individual. Note that, though relies on fitting the unknown treatment responses in the training process, the proposed TauNet differs from other treatment response modelling methods in that they fitted treatment response models are no longer needed for out-of-sample causal inference. That is, we can use the learned HTE function to estimate treatment effects for any individual directly.

4.2 Shared Representation Layers

Learning with auxiliary tasks is also known as multi-task learning (Ruder 2017) in the machine learning literature. In the context of deep multi-task learning, multi-task neural networks are usually realized with parameter sharing of hidden layers. In the TauNet architecture illustrated in Fig.1, the two auxiliary models $\mu_0(\cdot), \mu_1(\cdot)$ and our target model $\tau(\cdot;\theta)$ are parameterized independently. To improve learning efficiency by sharing information between different components, we extend the naïve TauNet by adding shared hidden layers for representation learning into the original network architecture. The extended neural network architecture is illustrated in Fig.2.

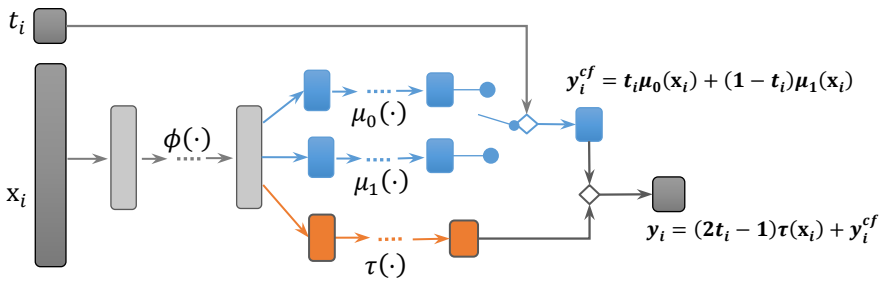

Figure 2. Neural network architecture of the extended TauNet with shared representation learning layers.

Denote the shared representation layers in the extended TauNet by $\phi: \mathcal{X} \to \Phi$ and parameterize the corresponding transformation function as $\phi(\cdot; W)$. For each individual x_i, the observed covariate will first be transformed into $\phi_W(\mathrm{x}_i) \in \Phi$. With shared representation layers, the two treatment response functions are $\mu_0(\phi_W(\mathrm{x}); \beta_0) = \mu_0(\phi_W(\mathrm{x}); \beta_0)$ and $\mu_1(\mathrm{x}; \beta_1) = \mu_1(\phi_W(\mathrm{x}); \beta_1)$. The target HTE function is parameterized as $\tau(\mathrm{x}; \theta) = \tau(\phi_W(\mathrm{x}); \theta)$. As a result, the empirical τ-risk for parameter optimization becomes

$$\mathcal{L}_\tau(W, \theta) = \frac{1}{n}\sum_{i=1}^{n}\{(2t_i - 1)\tau(\phi_W(\mathrm{x}_i); \theta) + t_i\mu_0(\phi_W(\mathrm{x}_i)) + (1 - t_i)\mu_1(\phi_W(\mathrm{x}_i)) - y_i\}^2 \quad (8)$$

In this realization, the two treatment response prediction tasks and the treatment effect estimation task can further share information in a deep multi-task learning manner through the shared representation learning layers $\phi(\cdot; W)$, whose weight matrix W is trained to minimize both the above empirical τ-risk and the *empirical μ-prediction risks* in Eq. (3).

4.3 Treatment Propensity Regularizer

While using hidden representation layers to share information among multiple tasks is easy, it is still important for us to figure out essential constraints on these shared representation layers. That is, how should we guide the parameters optimization process of these shared representation layers? In the treatment response modelling framework for causal inference, balanced representation learning methods (Shalit, Johansson & Sontag 2017) (Yao et al. 2018) propose to learn a balanced representation so that the discrepancy between the treated and control distributions induced by the learned representation is small. Based on the extended TauNet with shared representation layers, we follow the same idea used in the DragonNet by (Shi, Blei & Veitch 2019) that the shared representation should extract information that are both outcome and treatment predictive from the original pre-treatment covariates.

With such an objective in mind, we further add an auxiliary treatment prediction layer on top of the shared hidden representation layers and the extended neural network architecture is illustrated in Fig.3. In this neural network, denote the auxiliary treatment prediction layers as $g: \Phi \to [0,1]$ and parameterize it with parameters $\varphi \in \Psi$, the prediction loss for a specific $\varphi \in \Psi$ is then

$$\mathcal{L}_g(W, \varphi) = \frac{1}{n}\sum_{i=1}^{n} -t_i \log g(\phi_W(\mathrm{x}_i), \varphi) - (1 - t_i)\log\big(1 - g(\phi_W(\mathrm{x}_i), \varphi)\big) \quad (9)$$

Because this prediction loss is essentially targeted for constraining the learning of the shared representation layers such that the shared representation is both outcome and treatment predictive, we regard it as a regularizer and call it the treatment propensity regularizer.

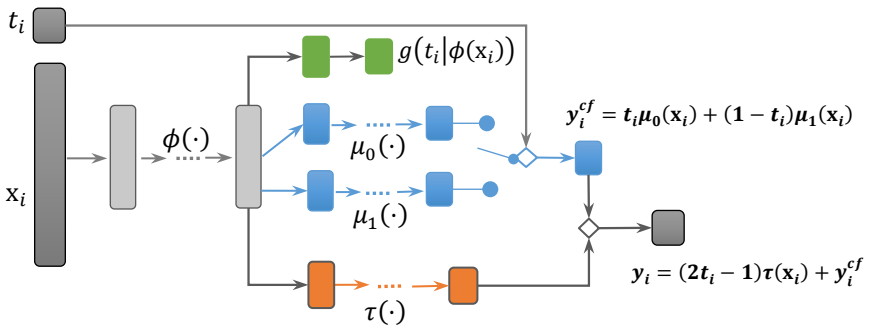

Figure 3. Neural network architecture of the extended TauNet with shared representation learning layers and treatment propensity regularization.

Table 1: Objective function specifications for different variants of TauNet

Estimators	$L_\mu^{(i)}$	$\mathcal{L}_\tau$	$\mathcal{L}_g$	Ω_{wd}
TauNet-Null	$L\left(\mu_{t_i}(\mathrm{x}_i;\beta_{t_i}),y_i\right)$	$\mathcal{L}_\tau(\theta)$	0	$\|\beta_0\|_2^2+\|\beta_1\|_2^2+\|\theta\|_2^2$
TauNet-Simple	$L\left(\mu_{t_i}(\phi_W(\mathrm{x}_i);\beta_{t_i}),y_i\right)$	$\mathcal{L}_\tau(W,\theta)$	0	$\|W\|_2^2+\|\beta_0\|_2^2+\|\beta_1\|_2^2+\|\theta\|_2^2$
TauNet-Reg	$L\left(\mu_{t_i}(\phi_W(\mathrm{x}_i);\beta_{t_i}),y_i\right)$	$\mathcal{L}_\tau(W,\theta)$	$\mathcal{L}_g(W,\varphi)$	$\|W\|_2^2+\|\beta_0\|_2^2+\|\beta_1\|_2^2+\|\varphi\|_2^2+\|\theta\|_2^2$

4.4 The Objective Functions and Algorithm

Define the outcome prediction loss for an observed sample (x_i, t_i, y_i) as

$$L_\mu^{(i)} = t_i \cdot L(\mu_0(\mathrm{x}_i), y_i) + (1-t_i)\cdot L(\mu_1(\mathrm{x}_i), y_i)$$

where $L(\mu(\mathrm{x}_i), y_i) = (\mu(\mathrm{x}_i)-y_i)^2$ for continuous outcomes and the log-loss $L(\mu(\mathrm{x}_i), y_i) = -y_i \log\mu(\mathrm{x}_i) - (1-y_i)\log\big(1-\mu(\mathrm{x}_i)\big)$ for binary outcomes. By combining the learning objectives of all components together and adding a weight decay regularization term Ω_{wd}, we obtain the following joint loss function for a general TauNet:

$$\mathcal{L}_n^{\text{TauNet}} = \mathcal{L}_\tau + \frac{\alpha}{n}\sum_{i=1}^{n} L_\mu^{(i)} + \gamma\cdot\mathcal{L}_g + \lambda\cdot\Omega_{\text{wd}} \quad (10)$$

where $\alpha, \gamma, \lambda > 0$ are hyper-parameters. Normally, the sample sizes of the two treatment groups in the training data are imbalanced. In practice, to compensate for this imbalance, we further weight the outcome prediction loss for each individual by an outcome prediction weight defined as

$$w_i = t_i + \frac{p(1-t_i)}{1-p}$$

where $p = p(t=1) = \frac{1}{n}\sum_{i=1}^{n} t_i$ is simply the treatment proportion in the training dataset. As a result, the objective function in Eq.(10) becomes

$$\mathcal{L}_n^{\text{TauNet}} = \mathcal{L}_\tau + \frac{\alpha}{n}\sum_{i=1}^{n} w_i\cdot L_\mu^{(i)} + \gamma\cdot\mathcal{L}_g + \lambda\cdot\Omega_{\text{wd}} \quad (11)$$

Since the TauNet with architecture shown in Fig.1 does not have shared hidden layers among different components, we label it the TauNet-Null. Analogously, the TauNets with architecture shown in Fig.2 and Fig.3 are labelled as TauNet-Simple and TauNet-Reg respectively since the latter use the treatment propensity prediction component to regularize the shared representation layers while the former does not. The specifications of different loss components in the general loss function Eq.(11) are listed in Table 1. We use the stochastic optimization method Adam (Kingma & Ba 2015) to train the model. The pseudocode for the joint learning process is summarized in Algorithm1.

Algorithm 1 Learning Process for TauNet

Input: Observation data $\mathcal{D} = \{(\mathrm{x}_1, t_1, y_1), \dots, (\mathrm{x}_n, t_n, y_n)\}$, hyper-parameters $\alpha, \gamma, \lambda \geq 0$, training batch size B, number of epochs K, and learning rate ϵ

Output: The learned parameters $(W, \varphi, \beta_0, \beta_1, \theta)$

Procedure:

1. Initialize parameters for $\mu_0(\cdot;\beta_0)$, $\mu_1(\cdot;\beta_1)$ and the HTE model $\tau(\cdot;\theta)$
2. Split $\mathcal{D}$ into the training set $\mathcal{D}_{train}$ and validation set $\mathcal{D}_{valid}$
3. **for** $k = 1, 2, \dots, K$, **do**
4. Sample a training batch $\mathcal{D}_{batch} = \{(\mathrm{x}_1^k, t_1^k, y_1^k), \dots, (\mathrm{x}_B^k, t_B^k, y_B^k)\}$ from $\mathcal{D}_{train}$
5. Update the parameters $(W, \varphi, \beta_0, \beta_1, \theta)$ by gradient descent to minimize Eq.(11) on $\mathcal{D}_{batch}$.
6. Test convergence using $\mathcal{D}_{valid}$, **if** converge
7. **break**
8. **end for**

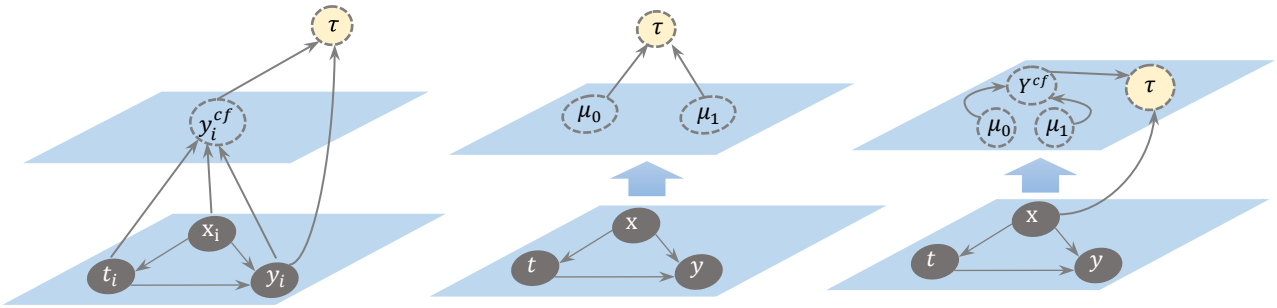

Figure 4. Graphical illustration of the causal inference process via the three groups of methods. (a) Indirect learning by inferring the counterfactual outcome via matching; (b) indirect learning methods by inferring the two underlying treatment response functions; and (c) the proposed direct learning method by modelling and inferring the target HTE function directly.

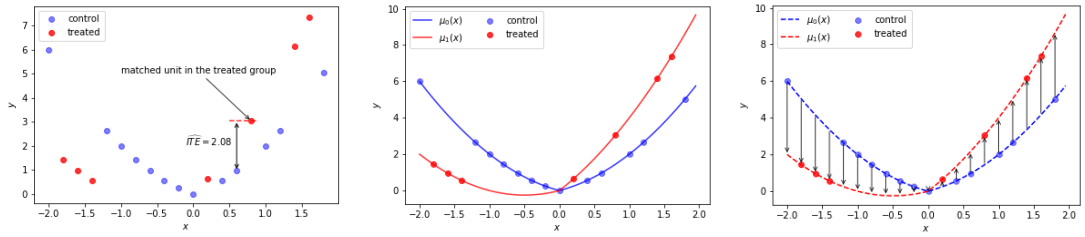

Figure 5. Illustration of the three groups of causal inference methods on the example data. (a) Indirect learning by inferring the counterfactual outcome via matching; (b) indirect learning methods by inferring the two underlying treatment response functions; and (c) the proposed direct learning methods by inferring the target HTE function directly.

4.5 Indirect *v.s.* Direct Learning Methods for Causal Inference

We have introduced a direct learning framework for causal inference and its three realizations: TauNet-Null, TauNet-Simple and TauNet-Reg. In this section, we summarize the difference between the proposed direct learning methods with existing indirect methods. Note that existing representation learning based methods intrinsically fulfill the task of causal inference by learning the treatment response functions. As a result, there are generally two groups of indirect learning methods for causal inference: (1) estimating treatment effects by inferring the counterfactual outcome for each individual; and (2) estimating treatment effects by inferring the two underlying treatment response functions. Overall, we illustrate the causal inference processes of the two groups of indirect learning methods and our proposed direct learning methods in Figure 4.

For the sake of explanation, consider a data-generating process where the observed scaler covariate $\mathrm{x} \in [-2, 2]$. The underlying treatment assignment mechanism that allocates treatments to individuals depends on the covariate value x via $t \sim \mathrm{Bern}\left(\frac{\mathrm{x}^2+0.5}{5}\right)$. With this treatment assignment mechanism, individuals with covariate value x far away from 0 are more likely to be treated while individuals with value x close to 0 are more likely be assigned into the control group. The two treatment response functions are $\mu_0(\mathrm{x}) = \mathrm{x}^2 + |\mathrm{x}|$ and $\mu_1(\mathrm{x}) = \mathrm{x}^2 + |\mathrm{x}| + 2\mathrm{x}$ respectively. With this data-generating process, it is easy to figure out that the target treatment effect function is $\tau(\mathrm{x}) = \mu_1(\mathrm{x}) - \mu_0(\mathrm{x}) = 2\mathrm{x}$. For simplicity, we generate 20 samples by this data-generating process. The observational samples $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), i = 1, \ldots, 20\}$ are illustrated in Fig.5.

Fig. 5(a) illustrates the estimation process of non-parametric causal inference via nearest neighborhood matching on the example data. Take the individual with $\mathrm{x} = 0.6$ for example, since we have observed its factual outcome under no treatment $y = \mu_0(\mathrm{x}) = 0.96$, to estimate the

ITE $\tau(\mathrm{x}=0.6)$, we firstly infer its counterfactual outcome if treated. Comparing it with other individuals in the treated group, we match the individual with its nearest neighbour whose covariate value is 0.8 and thus get its estimation as $y^{cf}=\mu_1(0.8)=3.04$. As a result, the non-parametric matching estimator obtain the estimated ITE for $\mathrm{x}=0.6$ as $\tau=y^{cf}-y=2.08$. By contrast, as shown in Fig. 5(b), methods based on treatment response modelling fit the treatment response functions $\mu_0(\mathrm{x})$ and $\mu_1(\mathrm{x})$ on the untreated and the treated samples respectively. Based on the learned treatment response functions, the ITE for any individual is calculated via $\tau(\mathrm{x})=\mu_1(\mathrm{x})-\mu_0(\mathrm{x})$.

Both non-parametric methods and outcome regression methods are very popular for causal inference in the literature. However, when the covariate dimension gets higher, it may not be possible for us to match individuals from a limit set of observations. In addition, when the underlying treatment responses are complex, mild model misspecification in treatment response modelling may lead serous bias in the final treatment effect estimation. Since the quantity we are of direct interest is the difference between the treatment response if treated versus that if untreated, why do we not learn a model of the difference directly? With such a question in mind, we propose to learn it directly using observational data rather than estimating it indirectly by firstly inferring the unobserved counterfactual outcome or the treatment response functions. The treatment effect estimation process of the direct learning framework is illustrated in Fig.5 (c).

5 Experimental Studies

In general, it is difficult to validate treatment effect estimation models on observational datasets since we have no access to all the potential outcomes or the true ITE for any individual. To evaluate the performance of our proposed direct learning method: Tau-Null, Tau-Simple and Tau-Reg, we conduct experiments on semi-simulated data, experimental data from real-world applications as well as synthetic data[1]. Details on hyper-parameter configurations are described in the appendix.

5.1 Baselines and Evaluation Metrics

We compare the proposed method empirically with the three groups of methods introduced in Section 3. Explanations of baseline methods are listed in Table 2.

Table 2: List of baseline methods

	Method	Explanation
Non-parametric methods	kNN (Crump et al. 2008)	Matching with k-nearest neighbors
	PSM (Stuart 2010)	Propensity score matching with logistic regression
	CF (Wager & Athey 2018)	Causal forest
Treatment response modelling based	OLS1/LR1	Ordinal least square / logistic regression with the treatment as a covariate
	OLS2/LR2	Separate ordinal least square / logistic regression for each treatment group
	BART (Hill 2011)	Bayesian additive regression trees
	SRF (Künzel et al. 2019)	S-Learner with random forest as meta learner
	TRF (Künzel et al. 2019)	T-Learner with random forest as meta learner
	XRF (Künzel et al. 2019)	X-Learner with random forest and logistic regression as meta learners
Representation learning based	BLR (Johansson, Shalit & Sontag 2016)	Balancing linear regression with covariate selection
	BNN (Johansson, Shalit & Sontag 2016)	Balancing neural network with linear discrepancy
	TARNet (Shalit, Johansson & Sontag 2017)	Target-agnostic representation network
	CFR-MMD (Shalit, Johansson & Sontag 2017)	Counterfactual regression network with the MMD metric
	CFR-Wass (Shalit, Johansson & Sontag 2017)	Counterfactual regression network with the Wasserstein metric
	DragonNet (Shi, Blei & Veitch 2019)	DragonNet with two outcome perdition heads and a propensity prediction head

[1] Source code will be openly accessible after revision.

It has been well known that evaluation of treatment effect estimation methods is difficult due to the fundamental problem that treatment outcomes are partially observed in the data. For synthetic and semi-simulated data where the true treatment effect for each individual is known, we use the square root of PEHE and mean absolute errors for evaluating the estimation performance of HTE, ATE and ATT respectively:

$$\epsilon_{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tau_i - \hat{\tau}_i)^2}$$

$$\epsilon_{ATE} = \left|\widehat{ATE} - ATE\right| = \frac{1}{n}\left|\sum_{i=1}^{n}(\tau_i - \hat{\tau}_i)\right|$$

$$\epsilon_{ATT} = \left|\widehat{ATT} - ATT\right| = \frac{1}{n_1}\left|\sum_{i\in\mathcal{T}_1}(\tau_i - \hat{\tau}_i)\right|$$

For all evaluation metrics, we report both the within-sample error and the out-of-sample error, where the former is computed over the training and validation sets, and the later is computed over the test set. Standard deviations for multiple replications are also reported.

5.2 Semi-simulated Data: IHDP

The first dataset we use to evaluate the proposed methods is the semi-simulated data IHDP. It was first compiled in (Hill 2011) based on the Infant Health and Development Program (IHDP), which is a randomized control trail aiming at assessing the impact of specialists' visits on children's test scores. There are 985 individuals recorded in the original dataset with each individual consisting of 6 continuous and 19 binary covariates measuring properties of a child and his/her mother. Examples of covariates include gender and birth weight of children, age and education attainment of mothers. The binary treatment t indicates whether a child was assigned into the program where both intensive high-quality childcare and home visits from a trained provider were provided. To create an observational study dataset, the records with non-white mothers in the treatment group are omitted to make the treatment and control groups unbalanced. In total there are 747 records (139 treated and 608 control) left in the new dataset.

By keeping the observed covariates and treatment variables from the original data fixed, we simulate treatment responses using both the setting "A" and "B" introduced in (Hill 2011). In particular, the setting A simulates linear treatment outcomes via $\mu_0(\mathrm{x}_i) = \mathrm{x}_i^T\beta_A$ and $\mu_1(\mathrm{x}_i) = \mathrm{x}_i{}^T\beta_A + 4$, where the coefficients in the 25-dimensional vector β_A are randomly sampled from $[0, 1,2,3,4]$ with probabilities $[0.5, 0.2,0.15, 0.1, 0.05]$. In the nonlinear outcome setting B, the two treatment response functions are $\mu_0(\mathrm{x}_i) = \exp((\mathrm{x}_i + 0.5I)^T\beta_B)$ and $\mu_1(\mathrm{x}_i) = \mathrm{x}_i{}^T\beta_B - \omega$, where the coefficients in β_B are randomly sampled from $[0, 0.1, 0.2, 0.3, 0.4]$ with probabilities $[0.6, 0.1, 0.1, 0.1, 0.1]$, and the offset ω was chosen such that the true ATE equals 4. In both simulated datasets, we observe for each individual a noisy observational outcome $y_i = t_i\mu_1(x_i) + (1 - t_i)\mu_0(x_i) + N(0,1)$.

We denote the dataset obtained via the simulation setting A and setting B as IHDP-A and IHDP-B respectively. The simulated noiseless outcomes are used to compute the true effects. With continuous outcomes, we used the L_2 loss $L(\mu(\mathrm{x}_i), y_i) = (\mu(\mathrm{x}_i) - y_i)^2$ for computing outcome prediction losses. For comparison, we follow similar neural network configurations in (Shalit, Johansson & Sontag 2017) that used 3 exponential-linear layers for the shared representation and the two auxiliary treatment response components. Layer size were 200 for representation learning and outcome prediction. Other network configurations are described in the appendix. We ran 10 replicates for selecting the hyper-parameters α, γ, λ and experimental results for 100 experiments with a 63/27/10 train/validation/test split ratio are demonstrated in Table 3.

Table 3: Within-sample and out-of-sample results on the IHDP dataset. (Lower is better)

	IHDP-A: Linear outcomes				IHDP- B: Non-linear outcomes			
	ϵ_{PEHE}^{in}	$\epsilon_{PEHE}^{\text{out}}$	ϵ_{ATE}^{in}	$\epsilon_{ATE}^{\text{out}}$	ϵ_{PEHE}^{in}	$\epsilon_{PEHE}^{\text{out}}$	ϵ_{ATE}^{in}	$\epsilon_{ATE}^{\text{out}}$
kNN	2.68±1.75	4.73±1.91	0.18±0.21	0.46±0.14	2.33±1.75	3.95±1.91	0.16±0.21	0.54±0.14
PSM	6.68±3.38	6.76±2.78	3.79±1.31	3.89±0.10	5.87±3.38	5.86±2.78	3.66±1.31	3.55±0.10
CF	4.77±2.33	5.07±2.09	0.39±0.36	0.64±0.59	3.07±2.33	3.03±2.09	0.35±0.36	0.47±0.59
OLS1	5.06±1.26	5.08±1.06	0.83±0.69	0.90±0.13	4.03±1.26	3.95±1.06	0.61±0.69	0.73±0.13
OLS2	1.89±1.58	1.96±1.33	0.15±0.48	0.25±0.13	1.61±1.58	1.67±1.33	**0.12±0.48**	0.21±0.13
BART	1.17±0.30	1.93±0.54	**0.12±0.06**	0.23±0.08	0.84±0.30	1.13±0.54	**0.12±0.06**	0.17±0.08
SRF	3.46±1.70	3.87±1.65	0.57±0.34	0.80±0.21	2.34±1.70	2.46±1.65	0.32±0.34	0.35±0.21
TRF	2.04±0.77	3.06±1.46	0.15±0.08	0.37±0.11	1.51±0.77	2.06±1.46	0.14±0.08	0.24±0.11
XRF	3.31±1.73	3.73±1.83	0.24±0.20	0.46±0.16	2.28±1.73	2.40±1.83	0.20±0.20	0.33±0.16
BLR	1.39±1.45	1.42±1.73	0.21±0.23	0.24±0.28	1.02±1.45	1.07±1.73	0.18±0.23	0.21±0.28
BNN	1.28±0.99	1.32±1.26	0.17±0.12	0.23±0.18	1.03±0.99	1.07±1.26	0.20±0.12	0.22±0.18
TARNet	1.49±1.11	1.59±1.00	0.26±0.23	0.31±0.24	1.30±1.11	1.34±1.00	0.27±0.23	0.30±0.24
CFR-MMD	1.52±1.26	1.59±1.42	0.23±0.20	0.27±0.25	1.27±1.26	1.33±1.42	0.23±0.20	0.27±0.25
CFR-Wass	1.44±0.72	1.53±1.17	0.25±0.17	0.29±0.24	1.26±0.72	1.35±1.17	0.26±0.17	0.30±0.24
DragonNet	1.56±1.35	1.61±1.47	0.33±0.38	0.37±0.44	1.32±1.35	1.46±1.47	0.24±0.38	0.28±0.44
TauNet-Null	1.27±0.90	1.39±1.13	0.17±0.13	0.24±0.21	1.07±0.90	1.19±1.13	0.15±0.13	0.19±0.21
TauNet-Simple	**0.99±0.97**	**1.17±1.42**	0.20±0.25	0.24±0.31	**0.79±0.97**	**0.97±1.42**	0.16±0.25	**0.16±0.31**
TauNet-Reg	**1.14±1.10**	**1.25±1.30**	0.17±0.15	**0.22±0.22**	1.12±1.10	1.28±1.30	0.19±0.15	0.23±0.22

As we can see from the table, while our proposed methods obtain the lowest out-sample estimation errors in both ITE and ATE estimation, representation learning and DNN based estimators perform generally better than estimators based on linear or classical statistical models in terms of individual treatment effect estimation (ϵ_{PEHE}). In addition, the BART estimator specially designed for the IHDP data obtains very similar performances in out-sample ATE estimation, with slightly higher estimation errors $\epsilon_{ATE}^{\text{out}}$ than our TauNet-Simple and TauNet-Reg on the IHDP-A and IHDP-B datasets respectively. By comparing the performances of our TauNet-Reg method with DragonNet, we find that using a learned HTE model does improve ITE and ATE estimation performances for the IHDP dataset. Moreover, comparing the three TauNet variants, it is easy to conclude that sharing information by adding shared representation layers does benefit treatment effect estimation. However, it is beyond our expectation that regularizing the shared representation to make it treatment predictive does not necessarily improve treatment effect estimation.

To further investigate the influence of each auxiliary component in the objective function Eq.(11) on the final treatment effect estimation performance. We evaluate ITE estimation errors ϵ_{PEHE} of the TauNet-Reg method with related to the hyper-parameters α, γ, λ when we change one hyper-parameter at a time ($\alpha, \gamma \in (10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}), \lambda \in (10^{-4}, 10^{-3}, 10^{-2})$) while keeping other parameters in their optimal configurations. As we can conclude from the resulting error curves in Fig.6, while a large outcome prediction parameter (α) is important for low treatment effect estimation error, changes in the treatment prediction parameter (γ) does not really influence the final estimation very much. This actually matches the performances of the results in Table 3. In additional, For the IHDP data, since the simulation treatment response functions are relatively simple, using a higher weight decay parameter ($\lambda = 0.01$) encourages simpler models and thus tends to gain lower estimation errors.

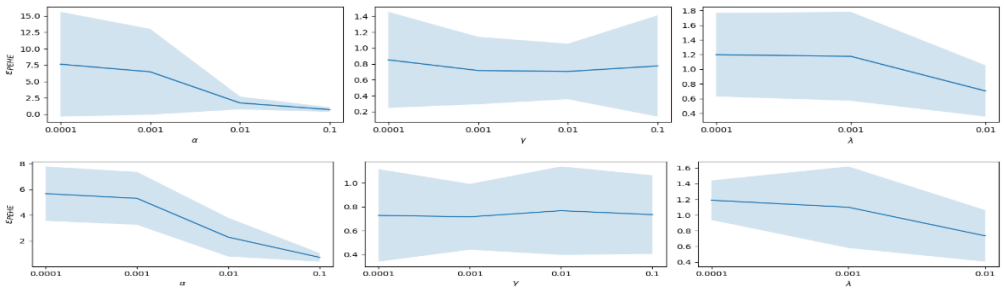

Figure 6. Out-sample ITE estimation errors ϵ_{PEHE} over different hyper-parameters for the IHDP-A dataset (top) and the IHDP-B dataset (bottom).

5.3 Real-World Data: Jobs

We also validate the proposed method using the real Jobs dataset, which combines a randomized study $\mathcal{R}$ based on the National Supported Work program with a larger observational dataset $\mathcal{O}$. This dataset was collected to evaluate the effect of job training programs on the employment status. In the original LaLonde randomized sample $\mathcal{R}$ by (LaLonde 1986), there are 722 employees (297 treated and 425 control) with 8 covariates such as age, education, and previous earnings. The binary treatment is whether an employee was enrolled in the job training program. For more details of the randomized study and data, refer[2]. To evaluate causal inference algorithms, (Shalit, Johansson & Sontag 2017) constructed the Job dataset by combining the LaLonde randomized sample $\mathcal{R}$ with the observational PSID comparison sample $\mathcal{O}$ (2490 control) to predict unemployment after job training. In the Jobs dataset, the original 8 covariates are transformed into a 17 dimension feature set. As a result, we obtain a real world binary-treatment binary-outcome dataset with 3212 examples and 17 dimensional features.

For the Jobs dataset, since the true ITEs are unknown, we are unable to calculate the RMSE ϵ_{ITE}. Following (Shalit, Johansson & Sontag 2017) and (Louizos et al. 2017), we use the policy risk estimated for the randomized subset $\mathcal{R}$ as a proxy to the ITE performance

$$R_{pol}(\pi_{\hat{\tau}}) = 1 - \big(p(\pi_{\hat{\tau}}(\mathrm{x}) = 1) \cdot \mathbb{E}[Y_1|\pi_{\hat{\tau}}(\mathrm{x}) = 1] + \big(1 - p(\pi_{\hat{\tau}}(\mathrm{x}) = 1)\big) \cdot \mathbb{E}[Y_0|\pi_{\hat{\tau}}(\mathrm{x}) = 0]\big)$$

where $\pi_{\hat{\tau}}: \mathcal{X} \to \{0,1\}$ is an policy induced from an ITE estimator $\hat{\tau}(\cdot)$ with $\pi_{\hat{\tau}}(\mathrm{x}) = 1$ if $\hat{\tau}(\mathrm{x}) > 0$, and $\pi_{\hat{\tau}}(\mathrm{x}) = 0$ otherwise. This measures the average regret when treating with the induced policy $\pi_{\hat{\tau}}$ and thus can serve as a proxy of the ITE estimation error. Instead of ATE, the NSW program aims at estimating the effect of job training on employment after training for employees enrolled in the training program, i.e., the ATT. Since all the treated individuals came from the randomized study $\mathcal{R}$, we can easily estimate ATT by

$$ATT := \frac{1}{|\mathcal{T}_1|}\sum_{i\in\mathcal{T}_1}\big(Y_1(\mathrm{x}_i) - Y_0(\mathrm{x}_i)\big) = \frac{1}{|\mathcal{T}_1|}\sum_{i\in\mathcal{T}_1} y_i - \frac{1}{|\mathcal{T}_0 \cap \mathcal{R}|}\sum_{i\in\mathcal{T}_0\cap\mathcal{R}} y_i$$

where $\mathcal{T}_1$ and $\mathcal{T}_0$ are the treated and control group in the full dataset. We replicated the experiment 50 times with a 56/24/20 train/validation/test ratio. Since we have only 297 treated samples in this dataset, we choose 500 samples for training at every training batch. The average performances and the corresponding empirical standard deviations are list in Table 4.

Table 4: Within-sample and out-of-sample results on Jobs dataset. (Lower is better)

	R_{pol}^{in}	R_{pol}^{out}	$\epsilon_{ATT}^{\text{in}}$	$\epsilon_{ATT}^{\text{out}}$
kNN	**0.08±0.01**	0.26±0.05	0.03±0.02	0.10±0.05
PSM	0.28±0.03	0.29±0.06	0.34±0.41	0.36±0.41
CF	0.17±0.02	0.24±0.05	0.02±0.01	0.07±0.05
LR1	0.22±0.01	0.23±0.05	0.01±0.01	**0.06±0.05**
LR2	0.23±0.01	0.24±0.05	**0.01±0.01**	**0.06±0.05**
BART	0.21±0.01	0.25±0.05	0.10±0.05	0.12±0.10
SRF	0.20±0.03	0.25±0.05	0.03±0.01	0.07±0.05
TRF	0.11±0.01	0.24±0.05	0.02±0.01	0.07±0.05
XRF	0.12±0.01	0.23±0.04	0.02±0.01	0.07±0.05
BLR	0.23±0.01	0.23±0.04	0.03±0.02	0.07±0.05
BNN	0.24±0.01	0.24±0.04	0.03±0.02	0.07±0.05
TARNet	0.23±0.01	0.23±0.05	0.06±0.02	0.08±0.05
CFR-MMD	0.23±0.01	0.24±0.04	0.04±0.02	0.09±0.05
CFR-Wass	0.23±0.01	0.24±0.04	0.04±0.03	0.08±0.07
DragonNet	0.17±0.02	0.22±0.04	0.04±0.04	0.10±0.08
TauNet-Null	0.18±0.02	**0.21±0.05**	0.03±0.02	0.08±0.06
TauNet-Simple	0.20±0.02	0.23±0.05	0.03±0.02	0.07±0.05
TauNet-Full	0.17±0.02	0.23±0.04	0.04±0.03	0.07±0.05

[2] Available at: http://users.nber.org/~rdehejia/data/nswdata2.html

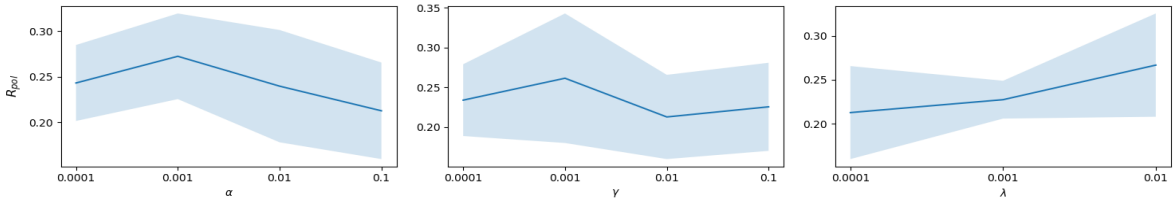

Figure 7. HTE estimation errors over different hyper-parameters for the Jobs dataset.

As we can see from the result, in general, almost all methods get decent estimation performances. In particular, straightforward modelling the treatment response models with logistic regression, either with a single outcome model or two separate outcome models, perform remarkably well in ATT estimation. However, since logistic regression is a linear model, the HTE function derived from the two logistic regression treatment response model will also be linear. As a result, we can only ascribe simple treatment policies and thus the policy risk for these methods are slightly higher than our proposed DNN-based methods. Comparing the three TauNet variants, we can conclude that adding shared representation learning layers does not improve heterogeneous treatment effect estimation but is likely to benefit ATT estimation.

We also investigated the HTE estimation performance R_{pol} of the TauNet-Reg method with related to the hyper-parameters α, γ and λ. The resulting error curves along hyper-parameter values are illustrated in Fig.7. As we can see, for this dataset, a relatively larger outcome prediction parameter (α) is important for getting a lower policy risk. Comparatively, the impact of the treatment prediction parameter (γ) fluctuates in its value range. Notably, observing the estimation errors along the weight decay parameter λ, we know that smaller λ and thus more complex models are preferred for treatment effect estimation on this dataset.

5.4 Experiment on Synthetic Data

To further check the robustness of the proposed models and their performance in different sample size and imbalance settings, we adapted the data simulation setup A in (Nie & Wager 2018) and simulated data by the following data-generating process:

$$\mathrm{x}_i \sim U(0,1)^5, \quad t_i|\mathrm{x}_i \sim \mathrm{Bern}\left(\mathrm{trim}_\eta(\sin(\pi \mathrm{x}_{i1}\mathrm{x}_{i2}))\right)$$

where $\mathrm{trim}_\eta(z) = \max\{\eta, \min(z, 1-\eta)\}$ and $\eta \in (0,0.5]$ is the imbalance parameter. The two treatment response functions and the observed factual outcome for each individual are respectively

$$\mu_0(\mathrm{x}_i) = \sin(\pi \mathrm{x}_{i1}\mathrm{x}_{i2}) + 2(\mathrm{x}_{i3} - 0.5)^2 + \mathrm{x}_{i4} + 0.5\mathrm{x}_{i5}$$

$$\mu_1(\mathrm{x}_i) = \mu_0(\mathrm{x}_i) + (\mathrm{x}_{i1} + \mathrm{x}_{i2})^2$$

$$y_i = t_i\mu_1(\mathrm{x}_i) + (1 - t_i)\mu_0(\mathrm{x}_i)$$

With this data-generating process, the underlying HTE function is $\tau(\mathrm{x}_i) = \mu_1(\mathrm{x}_i) - \mu_0(\mathrm{x}_i) = (\mathrm{x}_{i1} + \mathrm{x}_{i2})^2$. We simulated data with sample size $n = 500, 1K, 3K, 5K, 7K$ and $\eta = 0.1, 0.2, 0.3, 0.4, 0.5$. For each simulation setting, we split the data into train/validation/test sets with a ratio of 56/24/20 and replicated the experiments 50 times. We compared our direct learning methods with other baselines based on representation learning and DNNs (i.e., BLR, TARNet, BNN, CFR-MMD, CFR-WASS and DragonNet). All neural networks have similar configurations, with 2 hidden layers for each component and 50 neurons each layer. Hyper-parameters are set as $\gamma = 1$ and $\alpha = \lambda = 0.0001$. The training batch size was 200.

5.4.1 Performances in Different Imbalance Settings

According to the above data-generating process, the imbalance parameter $\eta \in (0,0.5]$ controls the data imbalance between the treated and untreated groups. Fig. 8 and Fig.9 illustrate respectively the error bar plots for out-sample ITE and ATE estimation errors with related to the imbalance parameter η in different sample sizes.

As we can see from the results, as η increases which indicates the data is getting more balanced, estimation errors of all methods generally decreased. While the two CFR methods get the largest estimation errors for both ITE and ATE estimation in almost all imbalance settings, our proposed methods obtain the lowest estimation errors in estimating ATE in all settings. Regarding the ITE estimation, our proposed TauNet-Reg method generally gets the lowest ITE estimation error except when $\eta = 0.5$ and the sample size $n = 5000$. Besides error means of different methods, it is also easy for us to see from the error bar plots that our proposed direct learning methods get generally lower empirical standard deviations than their competitors. This empirically indicates that the proposed direct learning methods are generally more stable for treatment effect estimation.

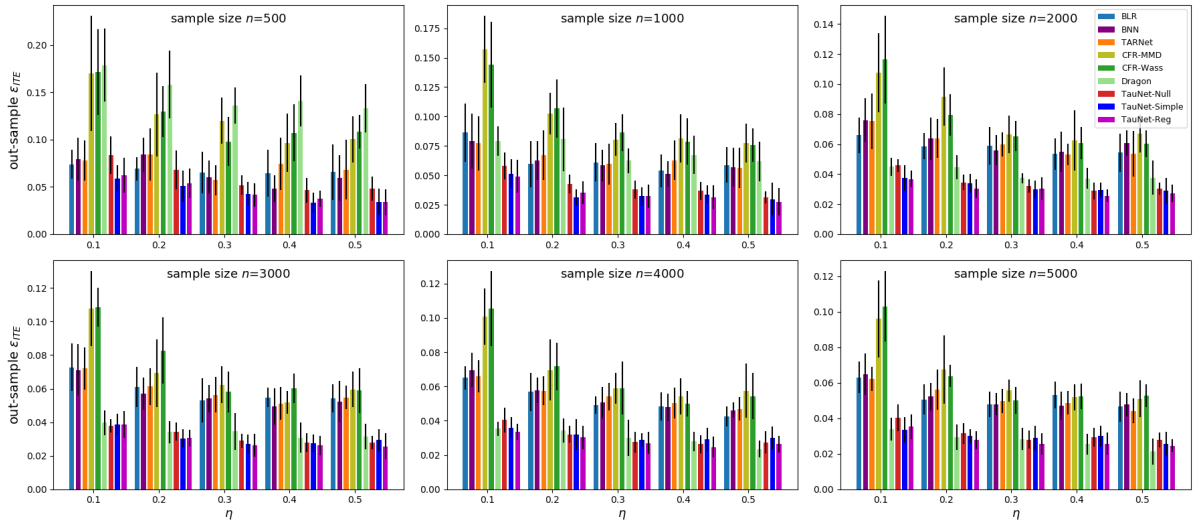

Figure 8. Comparisons of out-of-sample ITE estimation errors ϵ_{PEHE} and the corresponding empirical standard deviations with related to the imbalance parameter η in different sample size $n \in \{500, 1K, 3K, 5K, 7K\}$.

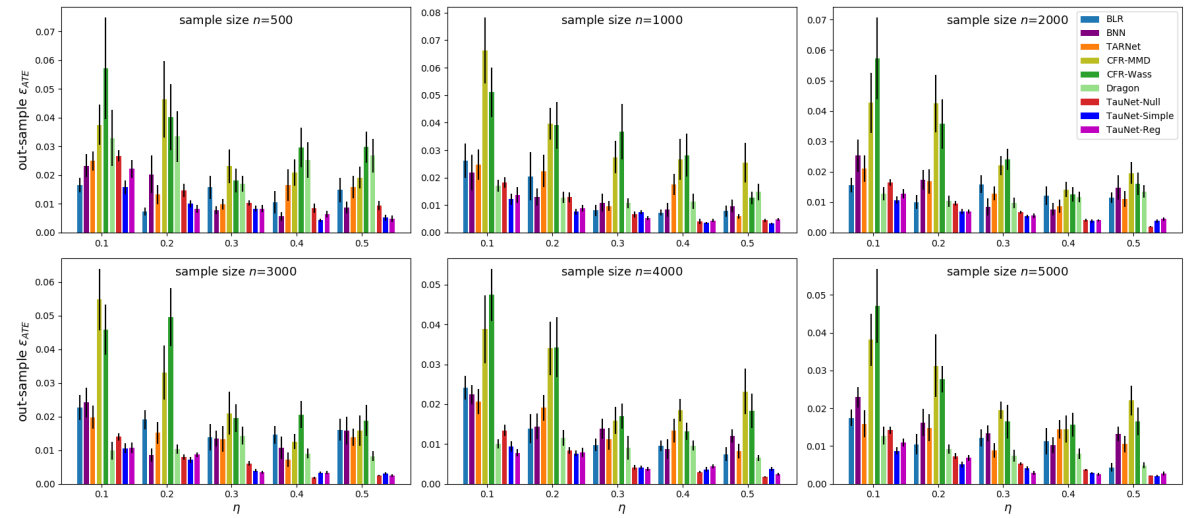

Figure 9. Comparisons of out-of-sample ATE estimation errors ϵ_{ATE} and the corresponding empirical standard deviations with related to the imbalance parameter η in different sample size $n \in \{500, 1K, 3K, 5K, 7K\}$.

5.4.2 Performances in Different Sample Sizes

To investigate the impact of sample size on estimation performances of different methods, we also illustrate the ITE and ATE estimation errors curves along sample size in different imbalance settings in Fig.10 and Fig.11 respectively. Since the performances of the three proposed methods are very similar, to avoid clutter, we only include the TauNet_Reg in Fig.10 and TauNet-Simple in Fig.11.

As we can see from Fig.10, on one hand, as the sample size gets larger, all methods generally get better estimation performances; on the other hand, our proposed TauNet-Reg generally obtains the most stable and lowest estimation errors. Although the DragonNet gets similar and even better performance as the TauNet-Reg when we have relatively large sample size, its performances when the sample size is relatively small are barely satisfactory, generally the worst among all the comparing methods. In regard to ATE estimation, it is easy to see from Fig.11 that while estimation errors of other baselines fluctuate as the sample size increase, the estimation error of our proposed TauNet-Simple is consistently lower and not sensitive to the change of sample size. In addition, empirical standard deviations of TauNet-Simple are also the smallest, indicating that our proposed methods are generally more stable.

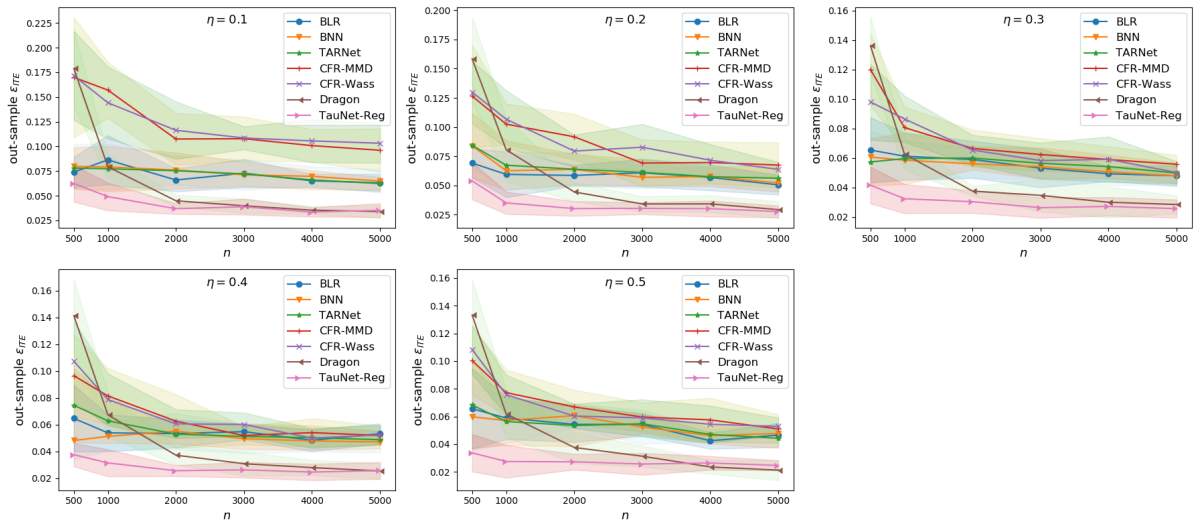

Figure 10. Comparisons of out-of-sample HTE estimation errors ϵ_{PEHE} and corresponding empirical standard deviations with related to the sample size n in different imbalance parameter settings $\eta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

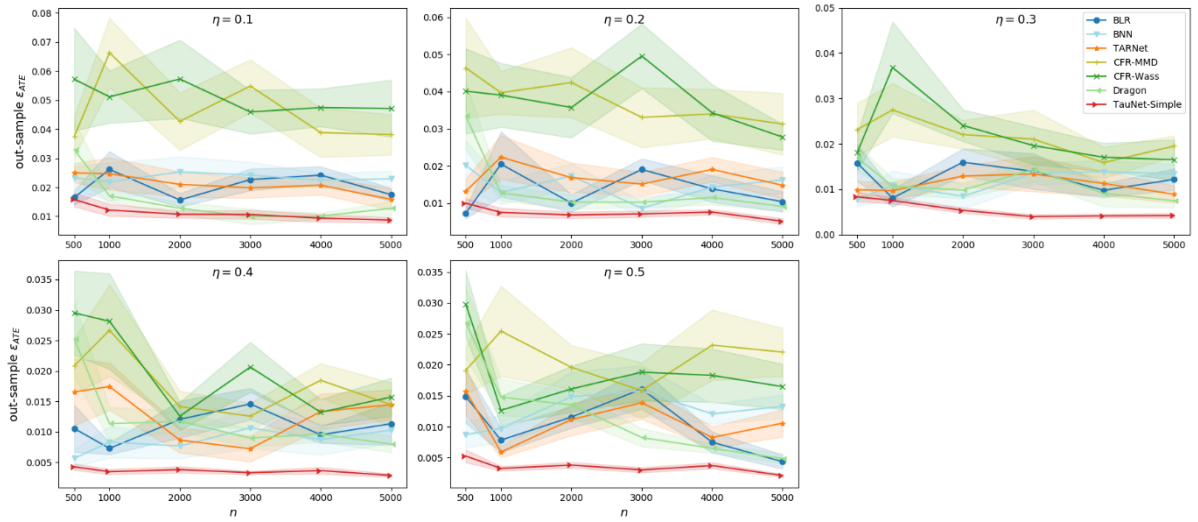

Figure 11. Comparisons of out-of-sample ATE estimation errors ϵ_{ATE} and corresponding empirical standard deviations with related to the sample size n in different imbalance parameter settings $\eta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

6 Conclusion and Discussion

In this paper, we proposed a direct learning framework called TauNet for treatment effect estimation from observational data. Compared with existing methods that complete the task in an indirect way by first inferring the unobserved counterfactual outcomes or the underlying treatment response models, the proposed TauNet parametrizes and learns the target treatment effect function directly. It builds on top of deep multi-task learning and is learned via an ad-hoc designed empirical τ-risk. As a realization of the conceptual framework, we proposed three variants of TauNet: TauNet-Null, TauNet-Simple and TauNet-Reg. To validate their effectiveness, we conducted comprehensive experiments and compared these realizations with a range of baselines. The experiment results showed that the proposed methods performed generally better than existing baselines and tended to obtain more stable estimates. Overall, we have focused on treatment effect estimation with binary treatments in this paper. An interesting future research question is how to extend the direct learning framework to settings with multivariate treatments and even continuous treatments.

Acknowledgements

This work was supported by the Australian Research Council (ARC) under Discovery Grant DP170101632.

Appendix. Experiment Configurations

Neural networks are implemented using the Tensorflow platform (Abadi et al. 2016). In all the experiments, we applied Xavier initialization (Glorot & Bengio 2010) for weight matrices, bias vectors are initialized by zeros, and scalar biases are initialized by 0.1. Search ranges of hyper-parameters are $\alpha, \gamma \in \left(10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\right), \lambda \in \left(10^{-4}, 10^{-3}, 10^{-2}\right)$. The experimental configurations for the two IHDP datasets and the Jobs dataset are listed in Table A1.

Table A1: Experiment configurations on each benchmark dataset

	IHDP-A	IHDP-B	Jobs
Outcome prediction parameter, α	0.1	0.1	0.1
Propensity regularization parameter, γ	0.01	0.1	0.01
Weight-decay parameter, λ	0.01	0.01	0.0001
Num. of representation layers	3	3	3
Num. of outcome prediction layers	2	2	3
Num. of HTE layers	2	2	2
Dim. of representation layers	200	200	200
Dim. of outcome prediction layers	200	200	200
Dim. of treatment prediction layers	100	100	100
Dim. of HTE layers	100	50	50
Batch size	100	100	500

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G. & Isard, M. 2016, 'Tensorflow: A system for large-scale machine learning', pp. 265-83.

Alaa, A.M. & van der Schaar, M. 2017, 'Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes', paper presented to the *Advances in Neural Information Processing Systems*.

Alaa, A.M., Weisz, M. & Schaar, M.v.d. 2017, 'Deep Counterfactual Networks with Propensity-Dropout', paper presented to the *Proceedings of the 34 th International Conference on Machine Learning*.

Alejandro Schuler, M.B., Robert Tibshirani, Nigam Shah 2018, 'A comparison of methods for model selection when estimating individual treatment effects', *arXiv:1804.05146*.

Atan, O., Jordon, J. & Schaar, M.v.d. 2018, 'Deep-Treat: Learning Optimal Personalized Treatments from Observational Data using Neural Networks', paper presented to the *AAAI*.

Athey, S. & Imbens, G. 2016, 'Recursive partitioning for heterogeneous causal effects', *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7353-60.

Bottou, L., Peters, J., Candela, J.Q., Charles, D.X., Chickering, M., Portugaly, E., Ray, D., Simard, P.Y. & Snelson, E. 2013, 'Counterfactual reasoning and learning systems: the example of computational advertising', *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3207-60.

Crump, R.K., Hotz, V.J., Imbens, G.W. & Mitnik, O.A. 2008, 'Nonparametric tests for treatment effect heterogeneity', *The Review of Economics and Statistics*, vol. 90, no. 3, pp. 389-405.

Glorot, X. & Bengio, Y. 2010, 'Understanding the difficulty of training deep feedforward neural networks', *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 249-56.

Grimmer, J., Messing, S. & Westwood, S.J. 2017, 'Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods', *Political Analysis*, vol. 25, no. 4, pp. 413-34.

Guo, R., Cheng, L., Li, J., Hahn, P.R. & Liu, H. 2018, 'A Survey of Learning Causality with Data: Problems and Methods', *arXiv preprint arXiv:1809.09337*.

Hernán, M.A. & Robins, J.M. 2018, *Causal Inference*, Boca Raton: Chapman & Hall/CRC, forthcoming.

Hill, J.L. 2011, 'Bayesian nonparametric modeling for causal inference', *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217-40.

Imbens, G.W. & Rubin, D.B. 2015, *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Johansson, F.D., Shalit, U. & Sontag, D. 2016, 'Learning Representations for Counterfactual Inference', paper presented to the *International conference on machine learning*.

Kallus, N. 2017, 'A framework for optimal matching for causal inference', *the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 372-81.

Kingma, D. & Ba, J. 2015, 'Adam: A method for stochastic optimization', paper presented to the *International Conference on Learning Representation*.

Künzel, S.R., Sekhon, J.S., Bickel, P.J. & Yu, B. 2019, 'Metalearners for estimating heterogeneous treatment effects using machine learning', *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4156-65.

LaLonde, R.J. 1986, 'Evaluating the econometric evaluations of training programs with experimental data', *The American economic review*, pp. 604-20.

Lin, A., Lu, J., Xuan, J., Zhu, F. & Zhang, G. 2020, 'A Causal Dirichlet Mixture Model for Causal Inference from Observational Data', *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1-29.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R. & Welling, M. 2017, 'Causal Effect Inference with Deep Latent-Variable Models', paper presented to the *Advances in Neural Information Processing Systems*.

Madras, D., Creager, E., Pitassi, T. & Zemel, R. 2019, 'Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data', *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, pp. 349-58.

McCaffrey, D.F., Ridgeway, G. & Morral, A.R. 2004, 'Propensity score estimation with boosted regression for evaluating causal effects in observational studies', *Psychological methods*, vol. 9, no. 4, p. 403.

Nie, X. & Wager, S. 2017, 'Learning Objectives for Treatment Effect Estimation', *arXiv preprint arXiv:1712.04912*.

Nie, X. & Wager, S. 2018, 'Quasi-Oracle Estimation of Heterogeneous Treatment Effects', *arXiv:1712.04912v2*.

Pearl, J. 2000, *Causality: Models, Reasoning and Inference*, Cambridge University Press.

Rosenbaum, P.R. & Rubin, D.B. 1983, 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, vol. 70, no. 1, pp. 41-55.

Ruder, S. 2017, 'An overview of multi-task learning in deep neural networks', *arXiv preprint arXiv:1706.05098*.

Schwab, P. & Karlen, W. 2019, 'CXPlain: Causal Explanations for Model Interpretation under Uncertainty', paper presented to the *Advances in Neural Information Processing Systems*.

Shalit, U., Johansson, F.D. & Sontag, D. 2017, 'Estimating individual treatment effect: generalization bounds and algorithms', paper presented to the *ICML*.

Shi, C., Blei, D. & Veitch, V. 2019, 'Adapting neural networks for the estimation of treatment effects', *Advances in Neural Information Processing Systems*, pp. 2503-13.

Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Schölkopf, B. & Lanckriet, G.R.G. 2012, 'On the empirical estimation of integral probability metrics', *Electronic Journal of Statistics*, vol. 6, pp. 1550-99.

Stuart, E.A. 2010, 'Matching methods for causal inference: A review and a look forward', *Statistical science: a review journal of the Institute of Mathematical Statistics*, vol. 25, no. 1, p. 1.

Sutton, R.S. & Barto, A.G. 2012, *Reinforcement learning: An introduction*, vol. 2, MIT press Cambridge.

Swaminathan, A. & Joachims, T. 2015a, 'Batch learning from logged bandit feedback through counterfactual risk minimization', *Journal of Machine Learning Research*, vol. 16, pp. 1731-55.

Swaminathan, A. & Joachims, T. 2015b, 'Counterfactual risk minimization: Learning from logged bandit feedback', paper presented to the *International Conference on Machine Learning*.

Wager, S. & Athey, S. 2018, 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228-42.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J. & Zhang, A. 2020, 'A Survey on Causal Inference', *arXiv preprint arXiv:2002.02770*.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J. & Zhang, A. 2018, 'Representation Learning for Treatment Effect Estimation from Observational Data', *Advances in Neural Information Processing Systems*, pp. 2634-44.

Yiu, S. & Su, L. 2018, 'Covariate association eliminating weights: a unified weighting framework for causal effect estimation', *Biometrika*.

Zhu, F., Lin, A., Zhang, G. & Lu, J. 2018, 'Counterfactual Inference with Hidden Confounders using Implicit Generative Models', paper presented to the *31th Australasian Joint Conference on Artificial Intelligence*, Wellington, New Zealand, 11th-14th December.

Zhu, F., Lu, J., Lin, A. & Zhang, G. 2020, 'A Pareto-smoothing method for causal inference using generalized Pareto distribution', *Neurocomputing*, vol. 378, pp. 142-52.

Zhu, Y., Savage, J.S. & Ghosh, D. 2018, 'A Kernel-Based Metric for Balance Assessment', *Journal of causal inference*, vol. 6, no. 2.