Counterfactual Prediction for Outcome-oriented Treatments

Hao Zou [1] **Bo Li** [2] **Jiangang Han** [3] **Shuiping Chen** [3] **Xuetao Ding** [3] **Peng Cui** [1]

Abstract

Large amounts of efforts have been devoted into learning counterfactual treatment outcome under various settings, including binary/continuous/multiple treatments. Most of these literature aims to minimize the estimation error of counterfactual outcome for the whole treatment space. However, in most scenarios when the counterfactual prediction model is utilized to assist decision-making, people are only concerned with the small fraction of treatments that can potentially induce superior outcome (i.e. outcome-oriented treatments). This gap of objective is even more severe when the number of possible treatments is large, for example under the continuous treatment setting. To overcome it, we establish a new objective of optimizing counterfactual prediction on outcome-oriented treatments, propose a novel Outcome-oriented Sample Re-weighting (OOSR) method to make the predictive model concentrate more on outcome-oriented treatments, and theoretically analyze that our method can improve treatment selection towards the optimal one. Extensive experimental results on both synthetic datasets and semi-synthetic datasets demonstrate the effectiveness of our method.

1. Introduction

In many fields, such as healthcare (Bica et al., 2020b;a) and marketing (Charles et al., 2013; Xu et al., 2022), it is beneficial for decision makers to accurately forecast individual outcome given different treatments. The randomized control trials (RCT) (Booth & Tannock, 2014), which is the golden standard to answer this question in causal inference, is expensive in time/resource (Kohavi & Longbotham, 2011) and even can be impossible (Charles et al., 2013; Kuang et al., 2017). Fortunately, the accumulation of observational data offer an opportunity to learn individual outcome of counterfactual treatments from the observational study.

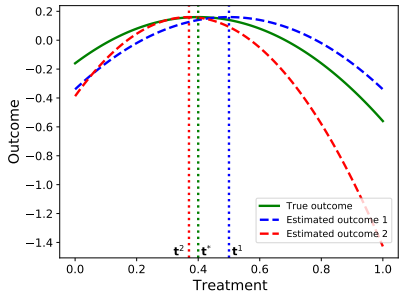


Figure 1. An example plot of two estimated outcome curve(red and blue dashed lines) and the ground truth(green solid line) for one fixed sample. The treatment value $\mathbf{t}^1$, $\mathbf{t}^2$ represent the pseudo-optimal treatment of the estimated outcome curve 1 and 2 respectively, and $\mathbf{t}^*$ represents the true-optimal treatment.

An important challenge in counterfactual prediction is the selection bias from confounding (Hassanpour & Greiner, 2019; Assaad et al., 2021) which indicates that the treatments are assigned not randomly, but with some explicit/implicit assignment policy manifested as correlations with some other covariates called confounders. Therefore, vanilla machine learning methods may induce systematic bias when predicting the outcome for the treatments assigned with a different assignment policy from the one in the training dataset. A large amount of literature in causal inference field have attempted to resolve this problem. Some literature (Johansson et al., 2016; Shalit et al., 2017; Bica et al., 2020a; Yao et al., 2018) introduce the idea of treatment invariant representation learning borrowed from domain adaptation field (Ganin & Lempitsky, 2015; Bousmalis et al., 2016). As an alternative method, sample re-weighting method (Hassanpour & Greiner, 2020; Assaad et al., 2021; Qian et al., 2021; Lim et al., 2018) adjusts the joint distribution of treatments and confounders to make them independent. In addition, some approaches (Yoon et al., 2018; Bica et al., 2020b; Qian et al., 2021) model the data distri-

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China; email:zouh18@mails.tsinghua.edu.cn, cuip@tsinghua.edu.cn [2]School of Economics and Management, Tsinghua University, Beijing, China; email: libo@sem.tsinghua.edu.cn [3]Meituan, Beijing, China; {hanjiangang,chenshuiping,dingxuetao}@meituan.com. Correspondence to: Peng Cui <cuip@tsinghua.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

bution to impute the counterfactual outcome and augment the biased dataset. Although targeting on various settings (e.g. different type of treatments, static or longitudinal data), the main target of these works is to minimize the estimation error of counterfactual outcome over the whole treatment space (Yoon et al., 2018; Bica et al., 2020b; Schwab et al., 2020).

In many application scenarios, however, when the researchers utilize the counterfactual prediction model to assist decision-making, they are only concerned with the treatments that can potentially induce better outcome (i.e. outcome-oriented treatments). It has also been acknowledged for long in the management science literature that better overall outcome predictions on all treatments may not result in better decisions (den Boer & Sierag, 2021; Besbes & Zeevi, 2015; Fernández-Loría & Provost, 2022). In these circumstances, the focus of counterfactual prediction models can be put more on minimizing the estimation error of counterfactual outcome over the outcome-oriented treatments (rather than solely the whole treatment space as in previous methods). But in real cases, we can hardly attain the true outcome of a counterfactual treatment to judge whether it is an outcome-oriented treatment. To remedy this, here we focus on the setting of continuous treatments, and assume a smooth outcome curve over treatments. In this way, we can reasonably use the optimal treatment derived from the counterfactual prediction model as a pseudo-optimal treatment, and regard the treatments around the pseudo-optimal treatment as outcome-oriented treatments.

We present a motivating example in Figure 1(den Boer & Sierag, 2021). The figure presents the true outcome curve and two estimated curves for a continuous treatment. We can observe that curve 1 achieves smaller predictive error on the entire treatment space, while curve 2 reports smaller errors on the outcome-oriented treatments. It is obvious that the pseudo-optimal treatment $\mathbf{t}^2$ of curve 2 is closer to the true optimal treatment $\mathbf{t}^*$ than that of curve 1 $\mathbf{t}^1$, and the true outcome of $\mathbf{t}^2$ is better than $\mathbf{t}^1$.

Theoretical analysis can also reveal that the treatment selection performance of model, which is characterized as the average outcome gap between true-optimal treatment and pseudo-optimal treatment over the population, is connected to the predictive error on small fraction of treatments instead of average error over the whole treatment space.

Inspired by the motivation and theoretical analysis, for making counterfactual learning more favorable to treatment selection, we propose Outcome-oriented Sample Reweighting (OOSR) algorithm. Specifically, it iteratively identifies the outcome-oriented treatments based on the current model and strengthen the outcome prediction on them, while ensuring the prediction over the whole treatment space.

Contribution Starting from improving treatment selection of counterfactual prediction model, we briefly define treatment selection regret (Fernández-Loría & Provost, 2022) as the performance metric and theoretically analyze that this optimization target is highly related to the outcome prediction error on the true/pseudo-optimal treatments instead of the average prediction error over the whole treatment space. To enhance the treatment selection performance, we derive a computationally tractable approximation of the regret bound as the objective function and give an easy-to-implement algorithm to minimize it. Specifically, we borrow the idea of sample re-weighting to simultaneously remove the original correlation between treatments and confounders in dataset and make the counterfactual learning concentrated on the outcome-oriented treatments. Extensive experimental results on both synthetic datasets and semi-synthetic datasets report that our method outperforms the existing methods in achieving smaller treatment selection regret.

2. Related Works

The related works consists of three parts, which are respectively counterfactual prediction, offline policy learning and targeted maximmum likelihood learning.

2.1. Counterfactual Prediction

Most of the previous literature about counterfactual prediction focus on the setting without unobserved confounders. The main idea of them is to remove the correlation between treatments and confounders in the observational dataset and achieve precise counterfactual outcome prediction on arbitrary treatments.

To realize this target, some works (Johansson et al., 2016; Shalit et al., 2017; Tanimoto et al., 2021; Bica et al., 2020a) characterize the undesired correlation as the distribution imbalance of confounder between the different treatment populations. Then it introduces the idea from domain invariant learning (Tzeng et al., 2014; Ganin & Lempitsky, 2015; Zhang et al., 2021) to learn the treatment invariant transformation of confounders, which is of the balanced distribution across treatment populations, and predict the outcome based on the transformed representation and treatment variable.

Besides the treatment invariant representation learning, sample re-weighting is an alternative method. Assaad et al. (2021) claims that over-enforcing the balancing property of representation may harm the predictive power while sample re-weighting schema can avoid it. Hence, some works (Hassanpour & Greiner, 2019; 2020; Johansson et al., 2018) calculate the sample weights based on trained propensity score model or directly learn the weights by distribution balance.

Yoon et al. (2018) propose to train an auxiliary model to

model the data distribution and generate the counterfactual data points. By augmenting the observational dataset with these counterfactual data points, the selection bias can be removed. There are also some literature (Alaa & van der Schaar, 2018; 2017; Zhang et al., 2020) which introduces gaussian process to model the data distribution and minimize the variance of counterfactual outcome prediction.

To deal with more complex treatments (e.g. continuous treatments, multiple treatments), some literature extends the strategies above. Arbour et al. (2021); Zou et al. (2020) applies density ratio estimation (Qin, 1998; Bickel et al., 2007; Sugiyama et al., 2012) between the original data distribution and designed target distribution to calculate sample weights. Tanimoto et al. (2021) propose to learn representation of both treatments and confounders which are independent. Bica et al. (2020b); Qian et al. (2021) propose the data augmentation methods for continuous treatment and multiple treatment.

As presented above, a large amount of works achieve accurate counterfactual prediction over the whole treatment space. However, lower outcome prediction error over the whole treatment space does not exactly means better decision making (Fernández-Loría & Provost, 2022). Tanimoto et al. (2021) also realize this gap and propose a regularizer to resolve the problem. However, the model is mainly designed for multi-dimensional binary treatment setting and the proposed regularizer aims for reducing the classification loss of whether the treatment outcome is larger than the average outcome. It can only help identify relatively good treatments (i.e. treatments with outcome superior than a baseline value). In this work, we focus on the continuous treatment setting where the number of treatments is infinite and target at the best treatment selection.

2.2. Offline Policy Learning

The paradigm of offline policy learning methods is typically defining a class of policy functions, which take confounders as input and output the treatment (distribution), and selecting the policy with the optimal estimated utility.

The class of policy function is usually defined as parameterized model, for example linear models (Swaminathan & Joachims, 2015a) and deep neural networks (Joachims et al., 2018). The objective function for optimizing the parameters is the utility estimated by offline policy evaluation methods. There have been many estimators proposed in the previous literature for accurate policy evaluation. The direct methods (Wang et al., 2019) learn an outcome predictive model from datasets and use the predicted result to estimate the utility. The sample re-weighting based methods, such as inverse propensity score (IPS) estimator (Swaminathan & Joachims, 2015a; Zhao et al., 2012), self-normalized estimator (Swaminathan & Joachims, 2015b; Joachims et al., 2018), attempt to balance the joint distribution of confounders and treatments between behavior policy and target policy to calculate the re-weighted outcome as utility. When the propensity score is unknown, the weights can be estimated by density ratio estimation (Sondhi et al., 2020) or directly balancing the distribution (Kallus, 2018). Some literature (Wang et al., 2017; Thomas & Brunskill, 2016; Dudík et al., 2011; Su et al., 2019) further combines the above two paradigms to take both the advantages of them to obtain more accurate evaluation. To the best of our knowledge, this branch of literature mainly focuses on the finite sample property of estimators, such as trade-off of bias and variance for precise evaluation. However, less attention has been paid to building counterfactual prediction model based on the rapidly developing machine learning for directly selecting treatments to minimize regret bypass offline policy evaluation. Although Wang et al. (2019) also involves learning predictive model with sample re-weighting for policy learning, it only targets to minimize the evaluation error of utility of the policy while neglecting the relationship between regret (i.e. the final goal of treatment selection) and the learned model.

2.3. Targeted Maximum Likelihood Learning

Our idea is potentially related to targeted maximum likelihood learning (TMLE) (Van Der Laan & Rubin, 2006), which is a framework in semi-parametric inference family. It proposes to learn an empirical estimator which minimize the estimation error of a target estimand. Although its application on some specific estimand (e.g. average treatment effect (ATE) estimation) is well-developed, making TMLE framework applicable to the problem in this paper is still worthy of studying to the best of our knowledge.

3. Notations and Problem Formulation

We define $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^d$ as the observed confounder variables, $\mathbf{t} \in \mathcal{T} = [a, b] \subseteq \mathbb{R}$ as continuous treatment assigned to each individual and $\mathbf{y} \in \mathbb{R}$ as outcome determined by the confounders and treatment. Hence, the observational dataset can be denoted as $\{(\mathbf{x}_i, t_i, y_i)\}_{1 \leq i \leq n}$, which is independently sampled from the joint distribution $p(\mathbf{X}, \mathbf{t}, \mathbf{y})$. The number n is the sample size.

We follow the potential outcome framework (Rosenbaum & Rubin, 1983; Rubin, 1984) in causal inference and assume there exist potential outcome function $Y_{\mathbf{X}}(\mathbf{t})$ denoting the potential outcome when assigned treatment $\mathbf{t}$ for sample with confounder $\mathbf{X}$. With the observational dataset, we aim to learn a model $f : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$ that predict potential outcome for the individual based on confounders and treatment. Since we can only observe the factual outcome y_i corresponding to the treatment t_i, we assume the following standard assumption to make the model f identifiable:

Assumption 3.1. Stable Unit Treatment Value The poten-

tial outcome of one sample is independent of the treatment assignments on the other samples.

Assumption 3.2. Unconfoundedness The assigned treatments and potential outcomes are independent conditional on observed covariates. Formally, $\mathbf{t} \perp \{Y_{\mathbf{X}}(\mathbf{t}')|\mathbf{t}' \in \mathcal{T}\}|\mathbf{X}$.

Assumption 3.3. Overlap For arbitrary $\mathbf{X} \in \mathcal{X}$ that satisfies $p(\mathbf{X}) > 0$, we have $p(\mathbf{t}|\mathbf{X}) > 0$ for each $\mathbf{t} \in \mathcal{T}$.

To characterize the treatment selection performance of a model f, we briefly define the treatment selection regret metric borrowed from decision theory (Fernández-Loría & Provost, 2022) as following:

$$Regret(f) = \mathbb{E}_{\mathbf{X}}\left[Y_{\mathbf{X}}(\rho^*(\mathbf{X})) - Y_{\mathbf{X}}(\rho^f(\mathbf{X}))\right], \quad (1)$$

where $\rho^*(\cdot)$ and $\rho^f(\cdot)$ are respectively the true-optimal treatment function and the pseudo-optimal treatment function derived from predictive model $f(\mathbf{X}, \mathbf{t})$. Formally,

$$\rho^*(\mathbf{X}) = \arg\max_{\mathbf{t}} Y_{\mathbf{X}}(\mathbf{t}), \quad (2)$$

$$\rho^f(\mathbf{X}) = \arg\max_{\mathbf{t}} f(\mathbf{X}, \mathbf{t}). \quad (3)$$

In this work, we assume larger outcome is preferred, for example, gross merchandise volume (GMV) in marketing. When smaller outcome is preferred in some applications, the regret metric and the algorithm/analysis below can be obtained in a similar manner.

4. OOSR: The Proposed Method

In this section, we firstly analyze the treatment selection regret, that is the optimization target of this problem and give an upper bound of it. Then inspired by the analysis, we give a computationally tractable approximation of the bound as the loss function and our Outcome-oriented Sample Reweighting (OOSR) method for training model. Finally, the detailed implementation of our algorithm is presented.

4.1. Theoretical Analysis on the Regret

We present that the treatment selection regret can be controlled by the outcome prediction error on two treatment points rather than the whole treatment space of each sample. The detailed relationship between the regret and predictive error is as following:

Proposition 4.1. *With the confounders* $\mathbf{X}$*, treatments* $\mathbf{t}$*, potential outcome function* $Y_{\mathbf{X}}(t)$ *defined as above, the treatment selection regret (i.e. Equation 1) of counterfactual prediction model* f *satisfies the following inequality:*

$$Regret(f) \leq \sqrt{\mathbb{E}_{\mathbf{X}}[(Y_{\mathbf{X}}(\rho^f(\mathbf{X})) - f(\mathbf{X}, \rho^f(\mathbf{X})))^2]} + \sqrt{\mathbb{E}_{\mathbf{X}}[(Y_{\mathbf{X}}(\rho^*(\mathbf{X})) - f(\mathbf{X}, \rho^*(\mathbf{X})))^2]} \quad (4)$$

The proof can be found in the appendix. From the result in Propostion 4.1, it can be concluded that if the outcome prediction error on the true optimal treatment $\rho^*(\mathbf{X})$ and pseudo-optimal treatment $\rho^f(\mathbf{X})$ are optimized to zero, we can achieve perfect treatment selection.

The first term of the r.h.s in Equation 4 can be calculated by the inverse propensity weighting (IPW) (Swaminathan & Joachims, 2015a) estimator as following:

$$\begin{aligned} & \mathbb{E}_{\mathbf{X}}[(Y_{\mathbf{X}}(\rho^f(\mathbf{X})) - f(\mathbf{X}, \rho^f(\mathbf{X})))^2] \\ = \quad & \mathbb{E}_{\mathbf{X},\mathbf{t}\sim p(\mathbf{X},\mathbf{t})}\left[\frac{\delta_{\rho^f(\mathbf{X})}(\mathbf{t})}{p(\mathbf{t}|\mathbf{X})}(Y_{\mathbf{X}}(\mathbf{t}) - f(\mathbf{X}, \mathbf{t}))^2\right] \end{aligned} \quad (5)$$

where $\delta_{\rho^f(\mathbf{X})(\mathbf{t})}(\mathbf{t})$ is Dirac delta function. The empirical version of the estimator in Equation 5 can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_{\rho^f(\mathbf{x}_i)}(t_i)}{p(t_i|\mathbf{x}_i)}(y_i - f(\mathbf{x}_i, t_i))^2. \quad (6)$$

It is an unbiased estimator if the denominator $p(t_i|\mathbf{x}_i)$ is true value (Strehl et al., 2010). However, under the continuous treatment setting, the empirical result of Equation 6 can easily be 0 since $p(t_i = \rho^f(\mathbf{x}_i)) = 0$ for each unit. To make the estimator more practical, we utilize the result in Kallus & Zhou (2018) and approximate Equation 6 as following:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{K\left((\rho^f(\mathbf{x}_i) - t_i)/\tau\right)}{\tau p(t_i|\mathbf{x}_i)}(y_i - f(\mathbf{x}_i, t_i))^2, \quad (7)$$

where $K(\cdot)$ is the kernel function that smooth $\delta_{\rho^f(\mathbf{x}_i)}(t_i)$. There are several candidate function for $K(\cdot)$, for example Epanechnikov kernel and Gaussian kernel. In this work, we choose $K(u) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}}$. For the sake of conciseness, we denote the estimator in Equation 7 as $\mathcal{A}(f)$.

Remark 4.2. We can perceive the estimator $\mathcal{A}(f)$ concentrate on the outcome-oriented treatment region around $\rho^f(\mathbf{x}_i)$ rather than the single point. The hyper-parameter τ control the strength of concentration. When τ approaches 0, Equation 7 degenerates to Equation 6.

Remark 4.3. The approximation in Equation 7 relies on that the error curve $(Y_{\mathbf{X}}(\mathbf{t}) - f(\mathbf{X}, \mathbf{t}))^2$ is smooth, which is also implied by previous literature (Kallus & Zhou, 2018; Hansen, 2009). When the outcome curve is non-smooth, the error curve will also be non-smooth and the estimation result of Equation 7 will suffer from large approximation error. It may bring damage to performance of our method. More detailed discuss can be found in appendix.

The second term of the r.h.s in Equation 4 represents the outcome prediction error on the true-optimal treatment. Unlike $\rho^f(\mathbf{X})$, $\rho^*(\mathbf{X})$ is intractable and can be arbitrary value in $\mathcal{T}$ in general. Hence, in order to reduce the predictive error

of $Y_{\mathbf{X}}(\rho^*(\mathbf{X}))$, we resort to minimize the mean predictive error over the treatment space $\mathcal{T}$, which is also the original optimization target of counterfactual prediction. Under some mild condition, the predictive error of $Y_{\mathbf{X}}(\rho^*(\mathbf{X}))$ can be upper bounded by the mean predictive error over $\mathcal{T}$ plus a constant.

Proposition 4.4. *Given the treatment space $\mathcal{T} = [a, b]$, if we assume the predictive loss function $\mathcal{G}(\mathbf{X}, \mathbf{t}) = (Y_{\mathbf{X}}(\mathbf{t}) - f(\mathbf{X}, \mathbf{t}))^2$ is L-Lipschitz on $\mathbf{t}$, then we have:*

$$\begin{aligned} & \mathbb{E}_{\mathbf{X}}\left[(Y_{\mathbf{X}}(\rho^*(\mathbf{X})) - f(\mathbf{X}, \rho^*(\mathbf{X})))^2\right] \\ \leq\ & \mathbb{E}_{\mathbf{X}}\left[\frac{1}{b-a}\int_{\mathbf{t}=a}^{b} \mathcal{G}(\mathbf{X}, \mathbf{t}) d\mathbf{t}\right] + L \cdot \frac{b-a}{2} \\ =\ & \mathbb{E}_{\mathbf{X}, \mathbf{t}\sim p(\mathbf{X}, \mathbf{t})}\left[\frac{\mathcal{G}(\mathbf{X}, \mathbf{t})}{(b-a)p(\mathbf{t}|\mathbf{X})}\right] + L \cdot \frac{b-a}{2} \\ \approx\ & \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - f(\mathbf{x}_i, t_i))^2}{(b-a)p(t_i|\mathbf{x}_i)} + L \cdot \frac{b-a}{2} \end{aligned} \tag{8}$$

We also denote the last estimator in Equation 8 as $\mathcal{B}(f)$ for conciseness. There is much research space in dealing with the second term of the r.h.s in Equation 4, for example reducing the scope of $\rho^*(\mathbf{X})$ by some extra assumption and domain knowledge, or building the relationship between $\rho^*(\mathbf{X})$ and $\rho^f(\mathbf{X})$. We will leave it to future work.

4.2. Objective Function

With the assumptions and conclusions above, we can approximate the upper bound of treatment selection regret as $\sqrt{\mathcal{A}(f)} + \sqrt{\mathcal{B}(f)}$. We attempt to reduce the regret by training counterfactual prediction model f to minimize the approximated upper bound.

For the stability of training process, we adopt to minimize the weighted combination $\gamma\mathcal{A}(f) + \mathcal{B}(f)$.

Proposition 4.5. *Assuming the function is parameterized by θ, that is f_θ, and the functions $\mathcal{A}(f_\theta)$ and $\mathcal{B}(f_\theta)$ are differentiable and strictly convex on θ, θ^* is the global minimum point of $\sqrt{\mathcal{A}(f_\theta)} + \sqrt{\mathcal{B}(f_\theta)}$, then there exists $\gamma \in \mathbb{R}^+$ such that*

$$\theta^* = \arg\min_{\theta} \gamma\mathcal{A}(f_\theta) + \mathcal{B}(f_\theta) \tag{9}$$

Remark 4.6. The minimization objective $\gamma\mathcal{A}(f_\theta) + \mathcal{B}(f_\theta)$ demonstrate that our algorithm strengthen the outcome prediction on the outcome-oriented treatment region (i.e. the treatment region around $\rho^f(\mathbf{X})$), while simultaneously ensure the global predictive performance over the whole treatment space to some extent.

Therefore, the final loss function to optimize is as following:

$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \frac{1+\lambda K\left((\rho^{f_\theta}(\mathbf{x}_i) - t_i)/\tau\right)}{(b-a)p(t_i|\mathbf{x}_i)} \cdot (y_i - f_\theta(\mathbf{x}_i, t_i))^2, \tag{10}$$

where $\lambda = (b-a)\gamma/\tau$. We set λ and τ to be the hyperparameters of our loss function for training model f_θ.

4.3. Implementation

We successively introduce the components in our methods.

Inverse Propensity Score Estimation The direct estimation of $p(t_i|\mathbf{x}_i)$ usually requires to assume the type of conditional distribution. For example, the conditional distribution $p(t_i|\mathbf{x}_i)$ can be assumed to be a gaussian distribution, formally $p(t_i|\mathbf{x}_i) = \mathcal{N}(\mu(\mathbf{x}_i), \sigma(\mathbf{x}_i))$. However, the assumption may be incorrect in many applications.

To reduce dependency on the assumption of $p(t_i|\mathbf{x}_i)$, we resort to density ratio estimation by solving a binary classification problem. Specifically, we define a uniform target distribution $p^u(\mathbf{X}, \mathbf{t}) = p(\mathbf{X})p^u(\mathbf{t}|\mathbf{X})$, where $p^u(\mathbf{t}|\mathbf{X})$ is a uniform distribution on $\mathcal{T}$ and equals $\frac{1}{b-a}$ in this problem. Therefore, the inverse of propensity score becomes

$$\frac{1}{p(\mathbf{t}|\mathbf{X})} = \frac{(b-a)p^u(\mathbf{t}|\mathbf{X})}{p(\mathbf{t}|\mathbf{X})} = \frac{(b-a)p^u(\mathbf{X}, \mathbf{t})}{p(\mathbf{X}, \mathbf{t})}. \tag{11}$$

To estimate the density ratio between $p(\mathbf{X}, \mathbf{t})$ and $p^u(\mathbf{X}, \mathbf{t})$, we label the samples in observational dataset $\{(\mathbf{x}_i, t_i)\}_{1\leq i\leq n}$ as positive samples ($L = 1$) and uniformly sample treatments $t_i' \sim \mathrm{Unif}(a, b)$ to generate samples $\{(\mathbf{x}_i, t_i')\}_{1\leq i\leq n}$ with negative label ($L = 0$). Then we can get the density ratio as following:

$$\frac{p^u(\mathbf{X}, \mathbf{t})}{p(\mathbf{X}, \mathbf{t})} = \frac{p(\mathbf{X}, \mathbf{t}|L=0)}{p(\mathbf{X}, \mathbf{t}|L=1)} = \frac{p(L=1)}{p(L=0)} \cdot \frac{p(L=0|\mathbf{X}, \mathbf{t})}{p(L=1|\mathbf{X}, \mathbf{t})}$$

After fitting these data points into a deep neural network based classifer, the term $p(L|\mathbf{X}, \mathbf{t})$ can be estimated by the output of classifer $\hat{p}(L|\mathbf{X}, \mathbf{t})$. Considering the ratio $\frac{p(L=1)}{p(L=0)} = 1$ for all the samples, the inverse propensity score can be estimated by:

$$\frac{1}{\hat{p}(\mathbf{t}|\mathbf{X})} = \frac{(b-a)\hat{p}(L=0|\mathbf{X}, \mathbf{t})}{\hat{p}(L=1|\mathbf{X}, \mathbf{t})}$$

Outcome-oriented Re-weighting Since we do not have knowledge about $\rho^{f_\theta}(\mathbf{X})$ initially, we can divide the the training process of model f_θ into two stages. In the first stage, we train the model with sample weights $w_i^{(0)} = \frac{1}{(b-a)\hat{p}(t_i|\mathbf{x}_i)}$ which removes the undesired correlation between treatments and confounders in data, and get the model $f_\theta^{(0)}$.

In the second stage, we alternately update the sample weights and model parameters θ for m rounds. For the j^{th} round, we calculate the sample weights $\mathbf{w}^{(j)}$ based on the predictive model of the previous round $f_\theta^{(j-1)}$

$$w_i^{(j)} = \frac{1 + \lambda K\left((\rho^{f_\theta^{(j-1)}}(\mathbf{x}_j) - t_j)/\tau\right)}{(b-a)\hat{p}(t_j|\mathbf{x}_j)}$$

Table 1. The experimental results on synthetic datasets with the sample size n varying. The metrics are Mean±STD over 10 repeated experiments. The best performance is marked bold.

Linear setting: Fix the degree of selection bias $\alpha = 6.0$, varying the sample size n								
n	$n = 4000$		$n = 6000$		$n = 8000$		$n = 10000$	
Methods	Within-S.	Out-of-S.	Within-S.	Out-of-S.	Within-S.	Out-of-S.	Within-S.	Out-of-S.
MLP	0.914±0.133	0.929±0.131	0.887±0.160	0.895±0.160	0.804±0.236	0.811±0.239	0.833±0.207	0.849±0.208
SCIGAN	0.156±0.002	0.166±0.002	0.140±0.002	0.146±0.003	0.126±0.002	0.132±0.002	0.130±0.003	0.136±0.002
RMNet	0.343±0.285	0.347±0.290	0.286±0.241	0.287±0.244	0.181±0.098	0.178±0.096	0.192±0.136	0.193±0.137
IPS-BanditNet	0.125±0.021	0.130±0.022	0.105±0.018	0.109±0.019	0.104±0.014	0.108±0.015	0.103±0.019	0.107±0.020
BCRI	0.199±0.046	0.204±0.047	0.172±0.035	0.175±0.035	0.150±0.026	0.154±0.027	0.137±0.015	0.139±0.014
MLP-Debias	0.100±0.048	0.107±0.051	0.081±0.057	0.083±0.058	0.074±0.047	0.073±0.047	0.053±0.029	0.055±0.030
OOSR	**0.040**±0.018	**0.043**±0.020	**0.034**±0.023	**0.046**±0.024	**0.020**±0.011	**0.037**±0.011	**0.015**±0.010	**0.016**±0.010
Exponential setting: Fix the degree of selection bias $\alpha = 5.0$, varying the sample size n								
n	$n = 4000$		$n = 6000$		$n = 8000$		$n = 10000$	
Methods	Within-S.	Out-of-S.	Within-S.	Out-of-S.	Within-S.	Out-of-S.	Within-S.	Out-of-S.
MLP	0.699±0.190	0.716±0.196	0.829±0.109	0.847±0.114	0.769±0.156	0.787±0.161	0.754±0.157	0.772±0.161
SCIGAN	0.210±0.132	0.220±0.135	0.219±0.137	0.229±0.142	0.279±0.101	0.295±0.103	0.114±0.063	0.119±0.067
RMNet	0.320±0.307	0.326±0.313	0.325±0.231	0.331±0.228	0.233±0.170	0.234±0.173	**0.092**±0.054	**0.094**±0.053
IPS-BanditNet	0.066±0.011	0.069±0.013	**0.047**±0.010	**0.050**±0.011	**0.060**±0.026	**0.063**±0.026	0.103±0.102	0.122±0.105
BCRI	0.145±0.039	0.147±0.040	0.144±0.041	0.147±0.043	0.115±0.015	0.116±0.014	0.129±0.018	0.129±0.017
MLP-Debias	0.221±0.092	0.224±0.093	0.171±0.060	0.174±0.059	0.250±0.101	0.253±0.100	0.264±0.099	0.266±0.099
OOSR	**0.053**±0.035	**0.053**±0.034	0.068±0.020	0.069±0.020	0.085±0.061	0.085±0.062	0.099±0.077	0.098±0.078
Logit setting: Fix the degree of selection bias $\alpha = 5.0$, varying the sample size n								
n	$n = 4000$		$n = 6000$		$n = 8000$		$n = 10000$	
Methods	Within-S.	Out-of-S.	Within-S.	Out-of-S.	Within-S.	Out-of-S.	Within-S.	Out-of-S.
MLP	0.639±0.020	0.648±0.021	0.626±0.035	0.634±0.036	0.628±0.034	0.636±0.036	0.609±0.047	0.616±0.048
SCIGAN	0.279±0.119	0.277±0.124	0.297±0.130	0.300±0.132	0.236±0.069	0.237±0.072	0.212±0.038	0.210±0.037
RMNet	0.211±0.057	0.212±0.057	0.151±0.125	0.152±0.125	0.162±0.144	0.163±0.145	0.110±0.066	0.109±0.065
IPS-BanditNet	0.075±0.002	0.071±0.002	0.060±0.003	0.059±0.003	0.060±0.010	0.061±0.011	0.054±0.008	0.055±0.008
BCRI	0.114±0.015	0.115±0.015	0.106±0.012	0.107±0.012	0.104±0.011	0.104±0.011	0.088±0.010	0.089±0.010
MLP-Debias	0.148±0.059	0.146±0.058	0.114±0.051	0.113±0.050	0.120±0.040	0.120±0.039	0.112±0.045	0.111±0.044
OOSR	**0.041**±0.021	**0.042**±0.022	**0.025**±0.004	**0.023**±0.004	**0.022**±0.006	**0.022**±0.007	**0.050**±0.034	**0.049**±0.033

and then update the parameters of model θ based on the sample weights $\mathbf{w}^{(j)}$ for a number of iterations to obtain model $f_\theta^{(j)}$.

Outcome Predictive Model With the sample weights $\mathbf{w}^{(j)}$, the loss function at the j^{th} round is as following:

$$\mathcal{L}^{(j)} = \frac{1}{n}\sum_{i=1}^{n} w_i^{(j)} \cdot (f_\theta^{(j)}(\mathbf{x}_i, t_i) - y_i)^2$$

Specifically, because of the potential nonlinear relationship between outcomes $\mathbf{y}$ and combination of confounders $\mathbf{X}$ and treatment $\mathbf{t}$, we apply deep neural networks with parameters θ as the predictive model f_θ. The pseudo-code of the whole algorithm can be found in appendix.

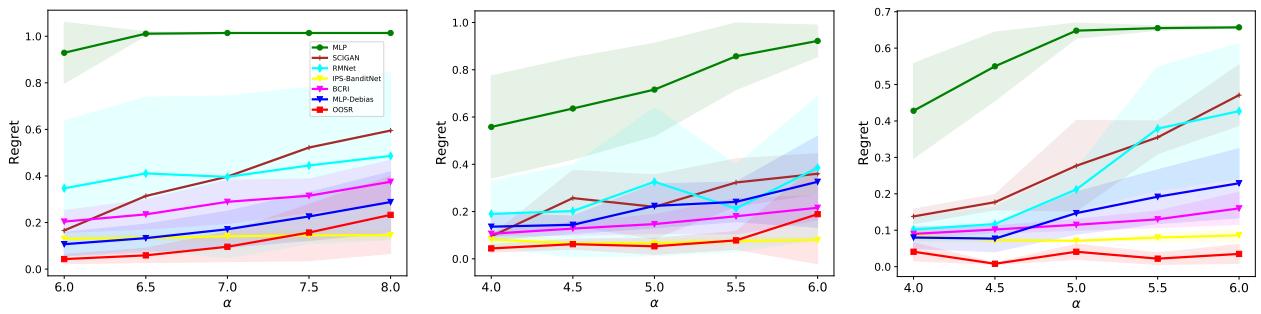
5. Empirical Results

Due to the lack of the observation of counterfactual outcomes in real-world datasets, we evaluate our proposed methods on both synthetic datasets and semi-synthetic datasets.

5.1. Experimental Setup

Baselines To demonstrate the advantage of our method, we compare it with the following approaches:

- Standard multilayer perceptron (MLP): It directly trains a predictive model on the observational dataset. The model takes the concatenation of confounder vector $\mathbf{X}$ and treatment variables $\mathbf{t}$ as input and output the corresponding outcome.
- MLP trained on debiased dataset (MLP-Debias): The sample weights is computed by density ratio estimation described above to remove the correlation between treatment and confounders in data. Then the predictive model is trained on the re-weighted dataset.
- SCIGAN (Bica et al., 2020a): It use generative adversarial networks to impute the outcome of counterfactual treatment of the samples and augment the dataset.



(a) Linear setting (b) Exponential setting (c) Logit setting

Figure 2. Simulation results(out-of-sample setting, the results under within-sample setting is similar and thus omitted)under different settings where the degree of selection bias α is varied. The curves present the mean value of regret in 10 repeated experiments. The shaded region presents the interval $[\text{mean} - \text{std}, \text{mean} + \text{std}]$ of the regret.

Then it trains the predictive model on the augmented dataset. The original model is designed for the combination of multi-level treatment and continuous treatment. Therefore, we constrain the number of multi-level treatment to be one for the compatibility with our setting.

- RMNet (Tanimoto et al., 2021): It learns the balanced representation of confounders and treatment to remove the correlation between treatment and confounders in data. Furthermore, it minimizes the classification error of whether or not the treatment in dataset is relatively good for each sample. We use the model version that is designed based on Wasserstein distance. The original model is designed for multi-dimensional binary treatment, we extend it to continuous treatment.

- IPS-BanditNet (Joachims et al., 2018) We use deep neural networks consisting of fully-connected layers as the parameterized policy π_θ, which takes confounders as input and output the treatment. The objective function is the standard estimator in Kallus & Zhou (2018).

- BCRI (Wang et al., 2019) We implement the policy π_θ of the softmax version with the deep neural networks based predictive model.

Evaluation Metric We evaluate the treatment selection performance under two different settings (Shalit et al., 2017). One is within-sample, where the metric is the average regret of the samples in the observational dataset. Formally, $Regret_{in} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{\mathbf{x}_i}(\rho^*(\mathbf{x}_i)) - Y_{\mathbf{x}_i}(\rho^f(\mathbf{x}_i))\right)$ for methods which directly selects treatments based on predictive model and $Regret_{in} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{\mathbf{x}_i}(\rho^*(\mathbf{x}_i)) - \mathbb{E}_{\mathbf{t}\sim\pi_\theta(\mathbf{t}|\mathbf{x}_i)}[Y_{\mathbf{x}_i}(\mathbf{t})]\right)$ for policy-based methods. Conversely, the other setting is out-of-sample, where the metric is averaged over the new samples $\{\mathbf{x}_i^{tes}\}_{1\le i\le n_{tes}}$ for which no factual outcome is observed.

The results of $\rho^f(\mathbf{X})$ and some $\rho^*(\mathbf{X})$ do not have closed-form solution. To numerically compute it, we sample q points consecutively with equal intervals in $[a, b]$ and select the one with largest true or predicted outcome. Formally,

$$\rho^f(\mathbf{X}) = \underset{\mathbf{t}\in\{a, a+\frac{b-a}{q-1},\ldots,b\}}{\arg\max} f(\mathbf{X}, \mathbf{t}). \tag{12}$$

We set the number of search points $q = 1001$.

5.2. Synthetic Dataset

Data Generation We generate the synthetic datasets under different settings. We first generate the confounders $\mathbf{X} = (x_1, x_2, ..., x_d)$, where each element is generated by $u_j \overset{iid}{\sim} \mathcal{N}(0, 1), x_j = |u_j|$. To generate outcome, we set two parameter vectors $\mathbf{v}_1 \in \mathbb{R}^{d\times 1}$ and $\mathbf{v}_2 \in \mathbb{R}^{d\times 1}$. They are sampled by firstly sampling each element of $\mathbf{u}_i \in \mathbb{R}^{d\times 1}$ from $\text{Unif}(0, 1)$, then setting $\mathbf{v}_i = \mathbf{u}_i/||\mathbf{u}_i||$, where $||\cdot||$ is Euclidean norm. Mimicking the demand curve in marketing (Besbes & Zeevi, 2015), we define the potential outcome to be $Y_{\mathbf{X}}(\mathbf{t}) = g(\mathbf{X}, \mathbf{t}) \cdot \mathbf{t}$, where the function $g(\mathbf{X}, \mathbf{t})$ is of the forms as following:

- Linear: $g(\mathbf{X}, \mathbf{t}) = max(-\mathbf{v}_2^\intercal\mathbf{X} \cdot \mathbf{t} + 1.8\mathbf{v}_1^\intercal\mathbf{X}, 0)$, where $\rho^*(\mathbf{X}) = min(0.9\mathbf{v}_1^\intercal\mathbf{X}/\mathbf{v}_2^\intercal\mathbf{X}, r)$
- Exponential: $g(\mathbf{X}, \mathbf{t}) = e^{-\mathbf{v}_2^\intercal\mathbf{X}\cdot\mathbf{t}+\mathbf{v}_1^\intercal\mathbf{X}}$, where $\rho^*(\mathbf{X}) = min(1/\mathbf{v}_2^\intercal\mathbf{X}, r)$
- Logit: $g(\mathbf{X}, \mathbf{t}) = 2/(1+e^{-\mathbf{v}_2^\intercal\mathbf{X}\cdot\mathbf{t}+\mathbf{v}_1^\intercal\mathbf{X}})$, where $\rho^*(\mathbf{X})$ is numerically calculate by Equation 12.

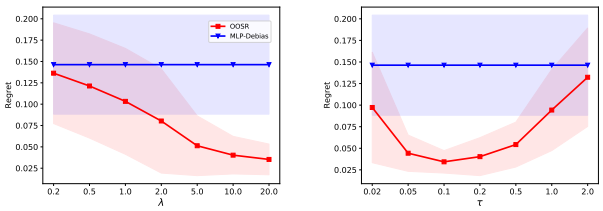
We assign treatment for each sample with a similar policy as in Bica et al. (2020b). Specifically, we sample t_i from a beta distribution. Supposed the treatment space is $[0, r]$, the treatment follows $\frac{t_i}{r} \sim \text{Beta}(\alpha, \beta)$. $\alpha \ge 1$ controls the degree of selection bias. When $\alpha = 1$, it is a uniform distribution. $\beta = \frac{\alpha-1}{\rho^*(\mathbf{x}_i)/2r} + 2 - \alpha$ guarantees that the mode of

treatment assignment distribution is $\rho^*(\mathbf{x}_i)/2$. After sampling the treatment, the factual outcome is the corresponding potential outcome $y_i = Y_{\mathbf{x}_i}(t_i) = g(\mathbf{x}_i, t_i) \cdot t_i$.

In the simulations, we set the dimension of confounders $d = 5$ and the treatment boundary $r = 2.0$ for Linear setting and $r = 3.0$ for Exponential and Logit settings. A sample set of size 10000 is randomly generated as held-out test-set to compute the out-of-sample metric.

Results We conduct experiments under different settings. For each setting, we repeated the experiments and train each models for 10 times. Then we calculate the mean value and standard deviation of the treatment selection regrets. The results of experiments where the sample size is varied are presented in Table 1. And the results of experiments with varying degree of selection bias is shown in Figure 2.

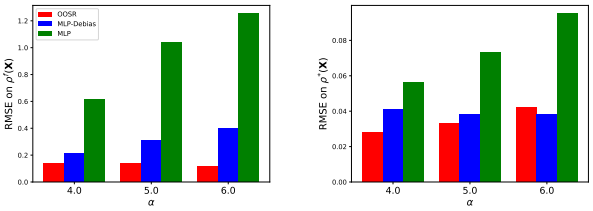
The overall trend is that with the sample size increasing the treatment selection regret become smaller. And with more severe selection bias (i.e. larger α), the regret becomes larger. Among the different methods, we can observe that directly trained MLP suffers from the selection bias in observational data and results in large treatment selection regret. The MLP-Debias method eliminates the correlation between treatments and confounders and achieve lower regret than vanilla MLP method. SCIGAN impute the counterfactual outcome and augment the observational dataset to remove selection bias. The predictive model trained on the augmented dataset can achieve improved regret. The RMNet is designed for multi-dimensional binary treatment setting, however, after adapted to the continuous treatment setting, it still have competitive performance under different settings. The policy-based methods also performs well in the experiments. Since the policy in BCRI is defined to be stochastic, it results in sub-optimal performance compared to our method. Our method fit the outcome curve based on the outcome-oriented sample weights and outperforms the other methods.



(a) Fix $\tau = 0.2$ (b) Fix $\lambda = 10.0$

Figure 3. Parameter analysis on λ and τ. We conduct experiments under Logit setting, while fixing $\alpha = 5.0$ and $n = 4000$.

Parameter Analysis The hyper-parameters are set to be $\lambda = 10.0$ and $\tau = 0.2$ in the experiments above. We also analyze the influence of parameter λ and τ on regret measured under the out-of-sample setting. The results are presented in Figure 3. From the results, we can observe that larger λ contributes to better performance. Because it strengthens the local outcome prediction on outcome-oriented treatments. When $\lambda > 10.0$, the regret is stable and do not change much. The regret become smaller with τ increasing at first, since extremely small τ makes the loss function less smooth. When τ increase further, the performance significantly drops. This is because the strength of outcome prediction optimization on the outcome-oriented treatments becomes weak.



(a) (b)

Figure 4. Comparison of outcome prediction on $\rho^f(\mathbf{X})$ and $\rho^*(\mathbf{X})$. We conduct experiments under Logit setting, while fixing $n = 4000$ and varying α.

Abalation study We measure the outcome prediction error (i.e. RMSE) on the treatment $\rho^f(\mathbf{X})$ and $\rho^*(\mathbf{X})$ (i.e. the two term in Equation 4) under Logit setting. From the results in Figure 4, we can observe that the outcome prediction error on $\rho^f(\mathbf{X})$ is the domination of the r.h.s in Equation 4 and our method significantly reduce it. And the outcome prediction error on treatment $\rho^*(\mathbf{X})$ is also suppressed to some extent. This phenomenon means the necessity of setting large λ and is consistent with the observation in parameter analysis.

5.3. Semi-synthetic Dataset

Data Generation The confounder feature is obtained from a real-world dataset TCGA (Weinstein et al., 2013). We choose the 10 columns with largest variance in the raw TCGA dataset as the confounder matrix $\mathbf{X}$. In semi-synthetic datasets, we set the treatment boundary as $r = 1.0$. To generate the treatments and outcomes, we sample three vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ as in synthetic datasets. The outcome generation process is similar to that in Bica et al. (2020b). The two simulated outcome curve is listed below:

- $Y_{\mathbf{X}}(\mathbf{t}) = \mathbf{v}_1^\intercal \mathbf{X} + (12\mathbf{v}_2^\intercal \mathbf{X} - 2) \cdot \mathbf{t} - (12\mathbf{v}_3^\intercal \mathbf{X} - 2) \cdot \mathbf{t}^2$, where $\rho^*(\mathbf{X}) = min((12\mathbf{v}_2^\intercal \mathbf{X} - 2)/(24\mathbf{v}_3^\intercal \mathbf{X} - 4), 1.0)$
- $Y_{\mathbf{X}}(\mathbf{t}) = \mathbf{v}_1^\intercal \mathbf{X} + 12\mathbf{t} \cdot \left(\mathbf{t} - 0.75\frac{\mathbf{v}_2^\intercal \mathbf{X}}{\mathbf{v}_3^\intercal \mathbf{X}}\right)^2$, where $\rho^*(\mathbf{X}) = \begin{cases} \frac{\mathbf{v}_2^\intercal \mathbf{X}}{\mathbf{v}_3^\intercal \mathbf{X}}/4 & \frac{\mathbf{v}_2^\intercal \mathbf{X}}{\mathbf{v}_3^\intercal \mathbf{X}} \geq 1.0 \\ 1.0 & \frac{\mathbf{v}_2^\intercal \mathbf{X}}{\mathbf{v}_3^\intercal \mathbf{X}} < 1.0 \end{cases}$

Table 2. The experimental results on semi-synthetic datasets of different methods. The metrics are Mean±STD over 10 repeated experiments. The best performance is marked bold.

Setting 1: Varying the degree of selection bias α								
α	$\alpha = 6.0$		$\alpha = 6.5$		$\alpha = 7.0$		$\alpha = 7.5$	
Methods	Within-S.	Out-of-S.	Within-S.	Out-of-S.	Within-S.	Out-of-S.	Within-S.	Out-of-S.
MLP	1.547±0.001	1.532±0.001	1.547±0.001	1.532±0.001	1.547±0.001	1.532±0.001	1.547±0.001	1.532±0.001
SCIGAN	0.251±0.006	0.254±0.006	0.387±0.008	0.392±0.008	0.551±0.010	0.556±0.009	0.785±0.013	0.792±0.013
RMNet	0.546±0.360	0.550±0.363	0.545±0.440	0.548±0.445	0.686±0.542	0.685±0.537	0.551±0.250	0.549±0.249
IPS-BanditNet	0.260±0.030	0.259±0.030	0.265±0.052	0.266±0.053	0.272±0.030	0.275±0.030	0.288±0.037	0.291±0.037
BCRI	0.091±0.063	0.093±0.061	0.121±0.088	0.124±0.090	0.186±0.039	0.187±0.038	0.502±0.176	0.499±0.171
MLP-Debias	0.040±0.014	0.039±0.014	0.202±0.071	0.204±0.071	0.276±0.083	0.278±0.086	0.346±0.090	0.352±0.093
OOSR	**0.016**±0.005	**0.015**±0.005	**0.096**±0.051	**0.097**±0.051	**0.125**±0.042	**0.127**±0.041	**0.187**±0.052	**0.190**±0.053
Setting 2: Varying the degree of selection bias α								
α	$\alpha = 4.0$		$\alpha = 4.5$		$\alpha = 5.0$		$\alpha = 5.5$	
Methods	Within-S.	Out-of-S.	Within-S.	Out-of-S.	Within-S.	Out-of-S.	Within-S.	Out-of-S.
MLP	0.100±0.064	0.098±0.058	0.210±0.058	0.195±0.054	0.192±0.068	0.182±0.063	0.279±0.073	0.266±0.071
SCIGAN	0.064±0.037	0.066±0.040	0.139±0.082	0.143±0.082	0.148±0.057	0.154±0.056	0.209±0.095	0.212±0.089
RMNet	0.154±0.064	0.159±0.065	0.145±0.068	0.149±0.070	0.165±0.129	0.169±0.128	0.189±0.080	0.192±0.075
IPS-BanditNet	0.509±0.044	0.496±0.045	0.491±0.033	0.473±0.034	0.580±0.125	0.569±0.133	0.623±0.156	0.608±0.155
BCRI	0.132±0.034	0.152±0.036	0.243±0.139	0.254±0.135	0.267±0.141	0.279±0.132	0.313±0.107	0.320±0.106
MLP-Debias	0.028±0.019	0.028±0.020	0.122±0.073	0.113±0.065	0.112±0.089	0.107±0.086	0.171±0.072	0.160±0.067
OOSR	**0.015**±0.014	**0.016**±0.015	**0.105**±0.079	**0.100**±0.071	**0.098**±0.096	**0.095**±0.093	**0.154**±0.066	**0.143**±0.062

As in the synthetic datasets, we assign treatment for each sample from a beta distribution. The treatment follows $t_i \sim \mathrm{Beta}(\alpha, \beta)$. $\alpha \geq 1$ controls the degree of selection bias and $\beta = \frac{\alpha-1}{\rho^*(\mathbf{x}_i)/2} + 2 - \alpha$. After assigning the treatment to each sample, the factual outcome is the corresponding potential outcome.

We randomly split 33% of the sample as the held-out test-set to compute out-of-sample metric.

Results We vary the degree of selection bias α and repeatedly conduct experiments for 10 times for each setting. The results are reported in Table 2.

The overall results is quitely consistent with the simulations. The directly trained MLP is prone to selection bias and results in large regret. The MLP-Debias, SCIGAN, RMNet, BCRI improve the performance based on the vanilla MLP in different degree. Under the setting 2, the performance of IPS-BanditNet significantly declines. The reason may be that for many samples, the optimal treatment is at boundary (i.e. $\rho^*(\mathbf{X}) = 1.0$) where less treatments is sampled in the observational dataset. Our proposed OOSR method attempt to optimize the outcome prediction further on the outcome-oriented treatments and achieve the best performance among the different methods.

6. Conclusion

In this paper, we studied the problem of learning counterfactual outcome for treatment selection. Under the continuous treatment setting, we theoretically analyze that the treatment selection regret is connected to prediction error on two treatment points, which is true/pseudo-optimal treatments rather than the whole treatment space. To improve treatment selection, we propose Outcome-oriented Sample Re-weighting (OOSR) method which strengthens the outcome prediction on the outcome-oriented treatment region to optimize the upper bound of regret. Extensive experimental results on the synthetic datasets and semi-synthetic datasets reveal the effectiveness of our method. The interesting direction of future work is the more delicate analysis on the upper bound of regret and optimization algorithm of the objective function.

Acknowledgements

This work was supported in part by National Key R&D Program of China (No. 2018AAA0102004), National Natural Science Foundation of China (No. 62141607, U1936219), and Beijing Academy of Artificial Intelligence (BAAI). Bo Li's research was supported by the National Natural Science Foundation of China (No.72171131); the Tsinghua University Initiative Scientific Research Grant (No. 2019THZWJC11); Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grants 2020AAA0108400 and 2020AAA0108403. We thank the reviewers for their insightful comments and suggestions.

References

Alaa, A. M. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian

processes. *Advances in Neural Information Processing Systems*, 30, 2017.

Alaa, A. M. and van der Schaar, M. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pp. 129–138. PMLR, 2018.

Arbour, D., Dimmery, D., and Sondhi, A. Permutation weighting. In *International Conference on Machine Learning*, pp. 331–341. PMLR, 2021.

Assaad, S., Zeng, S., Tao, C., Datta, S., Mehta, N., Henao, R., Li, F., and Duke, L. C. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pp. 1972–1980. PMLR, 2021.

Besbes, O. and Zeevi, A. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science*, 61(4):723–739, 2015.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *International Conference on Learning Representations*, 2020a.

Bica, I., Jordon, J., and van der Schaar, M. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in neural information processing systems (NeurIPS)*, 2020b.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pp. 81–88, 2007.

Booth, C. and Tannock, I. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. *British journal of cancer*, 110(3):551–555, 2014.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. Domain separation networks. *Advances in neural information processing systems*, 29:343–351, 2016.

Charles, D., Chickering, M., and Simard, P. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14, 2013.

den Boer, A. V. and Sierag, D. D. Decision-based model selection. *European Journal of Operational Research*, 290(2):671–686, 2021.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *ICML*, 2011.

Fernández-Loría, C. and Provost, F. Causal decision making and causal effect estimation are not the same… and why it matters. *INFORMS Journal on Data Science*, 2022.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Hansen, B. E. Lecture notes on nonparametrics. *Lecture notes*, 2009.

Hassanpour, N. and Greiner, R. Counterfactual regression with importance sampling weights. In *IJCAI*, pp. 5880–5887, 2019.

Hassanpour, N. and Greiner, R. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020.

Joachims, T., Swaminathan, A., and De Rijke, M. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*, 2018.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

Johansson, F. D., Kallus, N., Shalit, U., and Sontag, D. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

Kallus, N. Balanced policy evaluation and learning. *Advances in neural information processing systems*, 31, 2018.

Kallus, N. and Zhou, A. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pp. 1243–1251. PMLR, 2018.

Kohavi, R. and Longbotham, R. Unexpected results in online controlled experiments. *ACM SIGKDD Explorations Newsletter*, 12(2):31–35, 2011.

Kuang, K., Cui, P., Li, B., Jiang, M., Yang, S., and Wang, F. Treatment effect estimation with data-driven variable decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Lim, B., Alaa, A., and van der Schaar, M. Forecasting treatment responses over time using recurrent marginal structural networks. *NeurIPS*, 18:7483–7493, 2018.

Qian, Z., Curth, A., and van der Schaar, M. Estimating multi-cause treatment effects via single-cause perturbation. *Advances in Neural Information Processing Systems*, 34, 2021.

Qin, J. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

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, 1983.

Rubin, D. B. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pp. 1151–1172, 1984.

Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M., and Karlen, W. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5612–5619, 2020.

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

Sondhi, A., Arbour, D., and Dimmery, D. Balanced off-policy evaluation in general action spaces. In *International Conference on Artificial Intelligence and Statistics*, pp. 2413–2423. PMLR, 2020.

Strehl, A. L., Langford, J., Li, L., and Kakade, S. M. Learning from logged implicit exploration data. In *NIPS*, 2010.

Su, Y., Wang, L., Santacatterina, M., and Joachims, T. Cab: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learning*, pp. 6005–6014. PMLR, 2019.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pp. 814–823. PMLR, 2015a.

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015b.

Tanimoto, A., Sakai, T., Takenouchi, T., and Kashima, H. Regret minimization for causal inference on large treatment space. In *International Conference on Artificial Intelligence and Statistics*, pp. 946–954. PMLR, 2021.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR, 2016.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Van Der Laan, M. J. and Rubin, D. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

Wang, L., Bai, Y., Bhalla, A., and Joachims, T. Batch learning from bandit feedback through bias corrected reward imputation. In *ICML Workshop on Real-World Sequential Decision Making*, 2019.

Wang, Y.-X., Agarwal, A., and Dudık, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pp. 3589–3597. PMLR, 2017.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45:1113–1120, 2013.

Xu, R., Zhang, X., Cui, P., Li, B., Shen, Z., and Xu, J. Regulatory instruments for fair personalized pricing. In *Proceedings of the ACM Web Conference 2022*, pp. 4–15, 2022.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Zhang, X., Zhou, L., Xu, R., Cui, P., Shen, Z., and Liu, H. Domain-irrelevant representation learning for unsupervised domain generalization. *arXiv preprint arXiv:2107.06219*, 2021.

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

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Zou, H., Cui, P., Li, B., Shen, Z., Ma, J., Yang, H., and He, Y. Counterfactual prediction for bundle treatment. *Advances in Neural Information Processing Systems*, 33, 2020.

A. Proof

In this section, we give the proof of the proposition in the main paper.

Proposition A.1. ***(Restated)*** *With the confounders* $\mathbf{X}$*, treatments* $\mathbf{t}$*, potential outcome function* $Y_{\mathbf{X}}(t)$ *defined as above, the treatment selection regret (i.e. Equation 1) of counterfactual prediction model* f *satisfies the following inequality:*

$$Regret \leq \sqrt{\mathbb{E}_{\mathbf{X}}[(Y_{\mathbf{X}}(\rho^f(\mathbf{X})) - f(\mathbf{X}, \rho^f(\mathbf{X})))^2]} + \sqrt{\mathbb{E}_{\mathbf{X}}[(Y_{\mathbf{X}}(\rho^*(\mathbf{X})) - f(\mathbf{X}, \rho^*(\mathbf{X})))^2]}$$

Proof. According to the definition of $\rho^f(\mathbf{X})$, we have $f(\mathbf{X}, \rho^f(\mathbf{X})) \geq f(\mathbf{X}, \rho^*(\mathbf{X}))$. Then,

$$\begin{aligned} Regret &= \mathbb{E}_{\mathbf{X}}\left[Y_{\mathbf{X}}(\rho^*(\mathbf{X})) - Y_{\mathbf{X}}(\rho^f(\mathbf{X}))\right] \\ &\leq \mathbb{E}_{\mathbf{X}}\left[Y_{\mathbf{X}}(\rho^*(\mathbf{X})) - Y_{\mathbf{X}}(\rho^f(\mathbf{X})) + f(\mathbf{X}, \rho^f(\mathbf{X})) - f(\mathbf{X}, \rho^*(\mathbf{X}))\right] \\ &= \mathbb{E}_{\mathbf{X}}\left[Y_{\mathbf{X}}(\rho^*(\mathbf{X})) - f(\mathbf{X}, \rho^*(\mathbf{X}))\right] + \mathbb{E}_{\mathbf{X}}\left[f(\mathbf{X}, \rho^f(\mathbf{X})) - Y_{\mathbf{X}}(\rho^f(\mathbf{X}))\right] \end{aligned}$$

Based on mean-value inequality, we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{X}}\left[Y_{\mathbf{X}}(\rho^*(\mathbf{X})) - f(\mathbf{X}, \rho^*(\mathbf{X}))\right] + \mathbb{E}_{\mathbf{X}}\left[f(\mathbf{X}, \rho^f(\mathbf{X})) - Y_{\mathbf{X}}(\rho^f(\mathbf{X}))\right] \\ &\leq \sqrt{\mathbb{E}_{\mathbf{X}}\left[\left(Y_{\mathbf{X}}(\rho^*(\mathbf{X})) - f(\mathbf{X}, \rho^*(\mathbf{X}))\right)^2\right]} + \sqrt{\mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X}, \rho^f(\mathbf{X})) - Y_{\mathbf{X}}(\rho^f(\mathbf{X}))\right)^2\right]} \end{aligned}$$

□

Proposition A.2. ***(Restated)*** *Given the treatment space* $\mathcal{T} = [a, b]$*, if we assume the predictive loss function* $\mathcal{G}(\mathbf{X}, \mathbf{t}) = (Y_{\mathbf{X}}(\mathbf{t}) - f(\mathbf{X}, \mathbf{t}))^2$ *is L-Lipschitz on* $\mathbf{t}$*, then we have:*

$$\begin{aligned} &\mathbb{E}_{\mathbf{X}}\left[(Y_{\mathbf{X}}(\rho^*(\mathbf{X})) - f(\mathbf{X}, \rho^*(\mathbf{X})))^2\right] \leq \mathbb{E}_{\mathbf{X}}\left[\frac{1}{b-a}\int_{\mathbf{t}=a}^{b} \mathcal{G}(\mathbf{X}, \mathbf{t}) d\mathbf{t}\right] + L \cdot \frac{b-a}{2} \\ = \quad &\mathbb{E}_{\mathbf{X}, \mathbf{t} \sim p(\mathbf{X}, \mathbf{t})}\left[\frac{\mathcal{G}(\mathbf{X}, \mathbf{t})}{(b-a)p(\mathbf{t}|\mathbf{X})}\right] + L \cdot \frac{b-a}{2} \approx \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - f(\mathbf{x}_i, t_i))^2}{(b-a)p(t_i|\mathbf{x}_i)} + L \cdot \frac{b-a}{2} \end{aligned}$$

Proof. Because $\mathcal{G}(\mathbf{X}, \mathbf{t})$ is L-Lipschitz, $\mathcal{G}(\mathbf{X}, \rho^*(\mathbf{X})) \leq \mathcal{G}(\mathbf{X}, \mathbf{t}) + L \cdot |\rho^*(\mathbf{X}) - \mathbf{t}|$. Therefore, we have

$$\begin{aligned} \int_{\mathbf{t}=a}^{b} \mathcal{G}(\mathbf{X}, \rho^*(\mathbf{X})) d\mathbf{t} &\leq \int_{\mathbf{t}=a}^{b} \mathcal{G}(\mathbf{X}, \mathbf{t}) + L \cdot |\rho^*(\mathbf{X}) - \mathbf{t}| d\mathbf{t} \\ \Rightarrow (b-a)\mathcal{G}(\mathbf{X}, \rho^*(\mathbf{X})) &\leq \int_{\mathbf{t}=a}^{b} \mathcal{G}(\mathbf{X}, \mathbf{t}) + \frac{L}{2}(b-a)^2 \\ \Rightarrow \mathbb{E}_{\mathbf{X}}\left[(Y_{\mathbf{X}}(\rho^*(\mathbf{X})) - f(\mathbf{X}, \rho^*(\mathbf{X})))^2\right] &\leq \mathbb{E}_{\mathbf{X}}\left[\frac{1}{b-a}\int_{\mathbf{t}=a}^{b} \mathcal{G}(\mathbf{X}, \mathbf{t}) d\mathbf{t}\right] + L \cdot \frac{b-a}{2} \end{aligned}$$

Additionally,

$$\mathbb{E}_{\mathbf{X}}\left[\frac{1}{b-a}\int_{\mathbf{t}=a}^{b} \mathcal{G}(\mathbf{X}, \mathbf{t}) d\mathbf{t}\right] = \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbf{t}=a}^{b} p(\mathbf{t}|\mathbf{X}) \frac{\mathcal{G}(\mathbf{X}, \mathbf{t})}{(b-a)p(\mathbf{t}|\mathbf{X})} d\mathbf{t}\right] = \mathbb{E}_{\mathbf{X}, \mathbf{t} \sim p(\mathbf{X}, \mathbf{t})}\left[\frac{\mathcal{G}(\mathbf{X}, \mathbf{t})}{(b-a)p(\mathbf{t}|\mathbf{X})}\right].$$

Since the observational data $\{(\mathbf{x}_i, t_i)\}_{1 \leq i \leq n}$ is sampled from $p(\mathbf{X}, \mathbf{t})$, $\mathbb{E}_{\mathbf{X}, \mathbf{t} \sim p(\mathbf{X}, \mathbf{t})}\left[\frac{\mathcal{G}(\mathbf{X}, \mathbf{t})}{(b-a)p(\mathbf{t}|\mathbf{X})}\right]$ can be approximated by $\frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - f(\mathbf{x}_i, t_i))^2}{(b-a)p(t_i|\mathbf{x}_i)}$. □

Proposition A.3. ***(Restated)*** *Assume the function is parameterized by* θ*, that is* f_θ*, and the functions* $\mathcal{A}(f_\theta)$ *and* $\mathcal{B}(f_\theta)$ *are differentiable and strictly convex on* θ*,* θ^* *is the global minimum point of* $\sqrt{\mathcal{A}(f_\theta)} + \sqrt{\mathcal{B}(f_\theta)}$*, then there exists* $\gamma \in \mathbb{R}^+$ *such that*

$$\theta^* = \arg\min_{\theta} \gamma\mathcal{A}(f_\theta) + \mathcal{B}(f_\theta)$$

Proof. Since θ^* is the global minimum point of $\sqrt{\mathcal{A}(f_\theta)}+\sqrt{\mathcal{B}(f_\theta)}$, we have

$$\frac{\partial\sqrt{\mathcal{A}(f_\theta)}+\sqrt{\mathcal{B}(f_\theta)}}{\partial\theta}\Big|_{\theta=\theta^*}=0\Rightarrow\frac{\sqrt{\mathcal{B}(f_{\theta^*})}}{\sqrt{\mathcal{A}(f_{\theta^*})}}\frac{\partial\mathcal{A}(f_\theta)}{\partial\theta}+\frac{\partial\mathcal{B}(f_\theta)}{\partial\theta}\Big|_{\theta=\theta^*}=0,$$

Letting $\gamma=\frac{\sqrt{\mathcal{B}(f_{\theta^*})}}{\sqrt{\mathcal{A}(f_{\theta^*})}}$,

$$\frac{\partial\gamma\mathcal{A}(f_\theta)+\mathcal{B}(f_\theta)}{\partial\theta}\Big|_{\theta=\theta^*}=0.$$

Because the functions $\mathcal{A}(f_\theta)$ and $\mathcal{B}(f_\theta)$ are differentiable and strictly convex on θ, θ^* is also the global minimum point of $\gamma\mathcal{A}(f_\theta)+\mathcal{B}(f_\theta)$. □

B. Experimental Details

To allow for fair comparison, we ensure the different methods share the same backbone of predictive models. The predictive model is a neural networks with two hidden layers of size 20. The policy networks in IPS-BanditNet is of the same architecture. We use the ELU activation function. The predictive models are trained by SGD optimizer for 60000 iterations in synthetic experiments, 100000 iterations in setting 1 of semi-synthetic experiments and 300000 iterations in setting 2 of semi-synthetic experiments. The policy networks are trained for 4000 epochs. For our algorithm, in each experiments, the length of the first stage is 40% of the training process, and the length of each round in the second stage is 5% of the training process. Since the treatment space is bounded in the experiments, we truncate and normalize the kernel as in Kallus & Zhou (2018).

Since validation is a difficult problem in counterfactual prediction task, we select hyper-parameter in a indirect way. Firstly, we train a neural network $g(\mathbf{X},\mathbf{t})$ using the re-weighted dataset which is removed selection bias. Then we treat $g(\mathbf{X},\mathbf{t})$ as the ground truth of potential outcome and update the dataset $y_i'=g(\mathbf{x}_i,t_i)$. Using the updated dataset $\{(\mathbf{x}_i,t_i,y_i')\}_{1\le i\le n}$ and "ground truth" $g(\mathbf{X},\mathbf{t})$, we select hyper-parameters by grid searching. We choose $\lambda=10.0$ and $\tau=0.2$.

C. Smoothness of Outcome Curve

We define error term $\mathcal{E}(\mathbf{X},\mathbf{t})=(Y_{\mathbf{X}}(\mathbf{t})-f(\mathbf{X},\mathbf{t}))^2$. Then the approximation target of Equation 5 can be written as $\mathbb{E}_{\mathbf{X}}[\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))]$. We can have the following results on the bias of estimator in Equation 7.

Proposition C.1. *The estimation bias of Equation 7 is* $Bias(\mathcal{A}(f))=\frac{\kappa_2(K)\tau^2}{2}\mathbb{E}_{\mathbf{X}}\left[\frac{\partial^2}{\partial\mathbf{t}^2}\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))\right]+O(\tau^2)$*, where* $\kappa_2(K)=\int_u K(u)u^2du$.

Proof. Following the proof in Theorem 1 of Kallus & Zhou (2018), we have

$$\begin{aligned}\mathcal{E}(\mathbf{X},\mathbf{t}) &= \mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))+(\mathbf{t}-\rho^f(\mathbf{X}))\left(\tfrac{\partial}{\partial\mathbf{t}}\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))\right)+\tfrac{(\mathbf{t}-\rho^f(\mathbf{X}))^2}{2}\left(\tfrac{\partial^2}{\partial\mathbf{t}^2}\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))\right)+O((\mathbf{t}-\rho^f(\mathbf{X}))^2).\\ \mathbb{E}\left[\mathcal{A}(f)\right] &= \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbf{t}}p(\mathbf{t}|\mathbf{X})\frac{K((\rho^f(\mathbf{X})-\mathbf{t})/\tau)}{\tau p(\mathbf{t}|\mathbf{X})}\mathcal{E}(\mathbf{X},\mathbf{t})d\mathbf{t}\right]=\mathbb{E}_{\mathbf{X}}\left[\int_{\mathbf{t}}\frac{K((\rho^f(\mathbf{X})-\mathbf{t})/\tau)}{\tau}\mathcal{E}(\mathbf{X},\mathbf{t})d\mathbf{t}\right].\end{aligned}\tag{13}$$

Letting $u=(\mathbf{t}-\rho^f(\mathbf{X}))/\tau$, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{A}(f)\right] &= \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbf{t}}K(u)\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X})+\tau u)du\right]\\ &= \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbf{t}}K(u)\left(\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))+u\tau\frac{\partial}{\partial\mathbf{t}}\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))+\frac{(u\tau)^2}{2}\left(\frac{\partial^2}{\partial\mathbf{t}^2}\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))\right)+O(\tau^2)\right)du\right]\\ &= \int_u K(u)du\mathbb{E}_{\mathbf{X}}[\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))]+\int_u K(u)udu\cdot\mathbb{E}_{\mathbf{X}}[\tau\frac{\partial}{\partial\mathbf{t}}\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))]\\ &+ \frac{\kappa_2(K)\tau^2}{2}\mathbb{E}_{\mathbf{X}}\left[\frac{\partial^2}{\partial\mathbf{t}^2}\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))\right]+O(\tau^2)\\ &= \mathbb{E}_{\mathbf{X}}[\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))]+\frac{\kappa_2(K)\tau^2}{2}\mathbb{E}_{\mathbf{X}}\left[\frac{\partial^2}{\partial\mathbf{t}^2}\mathcal{E}(\mathbf{X},\rho^f(\mathbf{X}))\right]+O(\tau^2).\end{aligned}$$

Table 3. The experimental results on synthetic datasets of OOSR methods. We set the sample size $n = 4000$ and $\alpha = 6.0$.

Varying the intercept constant h.								
h	$h = 0.05$		$h = 0.10$		$h = 0.15$		$h = 0.20$	
Methods	Within-S.	Out-of-S.	Within-S.	Out-of-S.	Within-S.	Out-of-S.	Within-S.	Out-of-S.
OOSR	0.058±0.024	0.063±0.027	0.171±0.208	0.159±0.210	0.396±0.372	0.405±0.376	0.646±0.364	0.659±0.366

□

From the results in C.1, we can see that the smoothness of error curve (i.e. $\frac{\partial^2}{\partial \mathbf{t}^2}\mathcal{E}(\mathbf{X}, \rho^f(\mathbf{X}))$) is highly related to the approximation error. When the slope of the outcome curve changes substantially even the outcome curve is discontinuous, the error curve also suffers from the same problem. Then our objective function can not approximate the regret upper bound well, and this may affect the effectiveness of our method. To empirically demonstrate it, we modify the outcome curve in the Linear setting of synthetic datasets as following:

$$Y_{\mathbf{X}}(\mathbf{t}) = \begin{cases} max(-\mathbf{v}_2^\intercal \mathbf{X} \cdot \mathbf{t} + 1.8\mathbf{v}_1^\intercal \mathbf{X}, 0) \cdot \mathbf{t} - h & \mathbf{t} < 0.8\rho^*(\mathbf{X}) \\ max(-\mathbf{v}_2^\intercal \mathbf{X} \cdot \mathbf{t} + 1.8\mathbf{v}_1^\intercal \mathbf{X}, 0) \cdot \mathbf{t} & \mathbf{t} \geq 0.8\rho^*(\mathbf{X}) \end{cases}$$

The other mechanisms keep unchanged. Therefore, the true optimal treatments and corresponding treatment outcome also keep unchanged. We evaluate the performance of our method varying the intercept constant h. The results shown in Table 3 reveal that larger h makes the performance of our method worse.

D. Pseudo-code for Our Algorithm

Algorithm 1 Outcome-oriented Sample Re-weighting(OOSR)

Input: observational dataset $\{(\mathbf{x}_i, t_i, y_i)\}_{1 \leq i \leq n}$, learning rate η, the number of iterations T_1 in the first stage, the number of iterations $T_2 * m$ in the second stage.
Output: the predictive model $f_\theta^{(m)}$
Estimate the inverse propensity score $\frac{1}{\hat{p}(t_i|\mathbf{x}_i)}$ based on density ratio estimation.
Initialize the sample weights $w_i^{(0)} \leftarrow \frac{1}{(b-a)\hat{p}(t_i|\mathbf{x}_i)}$ and parameters of model $\theta^{(0)}$
for $i = 1$ **to** T_1 **do**
  Sample batch $B = \{(\mathbf{x}_i^B, t_i^B, y_i^B)\}_{1 \leq i \leq |B|}$
  $\mathcal{L}^{(0)} \leftarrow \frac{1}{n}\sum_{i=1}^{|B|} w_{i^B}^{(0)} \cdot (f_\theta^{(0)}(\mathbf{x}_i^B, t_i^B) - y_i^B)^2$
  Update $\theta^{(0)} \leftarrow \theta^{(0)} - \eta \frac{\partial \mathcal{L}^{(0)}}{\partial \theta^{(0)}}$
end for // The first stage finishes
for $j = 1$ **to** m **do**
  Update weights $w_i^{(j)} \leftarrow \frac{1+\lambda K\left((\rho^{f_\theta^{(j-1)}}(\mathbf{x}_j) - t_j)/\tau\right)}{(b-a)\hat{p}(t_j|\mathbf{x}_j)}$
  Initialize $\theta^{(j)} \leftarrow \theta^{(j-1)}$
  for $i = 1$ **to** T_2 **do**
    Sample batch $B = \{(\mathbf{x}_i^B, t_i^B, y_i^B)\}_{1 \leq i \leq |B|}$
    $\mathcal{L}^{(j)} \leftarrow \frac{1}{n}\sum_{i=1}^{|B|} w_{i^B}^{(j)} \cdot (f_\theta^{(j)}(\mathbf{x}_i^B, t_i^B) - y_i^B)^2$
    Update $\theta^{(j)} \leftarrow \theta^{(j)} - \eta \frac{\partial \mathcal{L}^{(j)}}{\partial \theta^{(j)}}$
  end for
end for // The second stage finishes
return the predictive model $f_\theta^{(m)}$