Treatment Effect Estimation with Adjustment Feature Selection

Haotian Wang
College of Computer, National University of Defense Technology
wanghaotian13@nudt.edu.cn

Kun Kuang*
Institute of Artificial Intelligence, Zhejiang University
kunkuang@zju.edu.cn

Haoang Chi
Intelligent Game and Decision Lab, Defense Innovation Institute
haoangchi618@gmail.com

Longqi Yang
Defense Innovation Institute
yanglongqi19@nudt.edu.cn

Mingyang Geng
College of Computer, National University of Defense Technology
gengmingyang13@nudt.edu.cn

Wanrong Huang
College of Computer, National University of Defense Technology
huangwanrong12@nudt.edu.cn

Wenjing Yang†
College of Computer, National University of Defense Technology
wenjing.yang@nudt.edu.cn

ABSTRACT

In causal inference, it is common to select a subset of observed covariates, named the adjustment features, to be adjusted for estimating the treatment effect. For real-world applications, the abundant covariates are usually observed, which contain extra variables partially correlating to the treatment (treatment-only variables, e.g., instrumental variables) or the outcome (outcome-only variables, e.g., precision variables) besides the confounders (variables that affect both the treatment and outcome). In principle, unbiased treatment effect estimation is achieved once the adjustment features contain all the confounders. However, the performance of empirical estimations varies a lot with different extra variables. To solve this issue, variable separation/selection for treatment effect estimation has received growing attention when the extra variables contain instrumental variables and precision variables.

In this paper, assuming no mediator variables exist, we consider a more general setting by allowing for the existence of post-treatment and post-outcome variables rather than instrumental and precision variables in observed covariates. Our target is to separate the treatment-only variables from the adjustment features. To this end, we establish a metric named **O**ptimal **A**djustment **F**eatures (OAF), which empirically measures the asymptotic variance of the estimation. Theoretically, we show that our OAF metric is minimized if and only if adjustment features consist of the confounders and outcome-only variables, i.e., the treatment-only variables are perfectly separated. As optimizing the OAF metric is a combinatorial optimization problem, we introduce Reinforcement Learning (RL) and adopt the policy gradient to search for the optimal adjustment set. Empirical results on both synthetic and real-world datasets demonstrate that (a) our method successfully searches the optimal adjustment features and (b) the searched adjustment features achieve a more precise estimation of the treatment effect.

CCS CONCEPTS

• **Computing methodologies** → **Causal reasoning and diagnostics**.

KEYWORDS

treatment effect estimation, covariate separation, confounder balancing

ACM Reference Format:
Haotian Wang, Kun Kuang, Haoang Chi, Longqi Yang, Mingyang Geng, Wanrong Huang, and Wenjing Yang. 2023. Treatment Effect Estimation with Adjustment Feature Selection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599531

1 INTRODUCTION

Causal inference [13, 18], which refers to inferring the variation of potential outcomes by intervening treatments, is a fundamental research area in decision-making [7, 35, 39] and interpretable artificial intelligence [14, 38]. Under the potential outcome framework [13], we aim to estimate the average effect of intervening the (binary) treatment **T** on the outcome **Y** given a set of covariates, as shown in Figure 1c. For example, a researcher attempts to assess the average treatment effect (ATE) of a drug (**T**) on patients' recovery (**Y**) from population data given some patients' characteristics. One fundamental problem of causal inference is the non-random treatment

*Kun Kuang and Haotian Wang contributed equally to this research.
†Wenjing Yang is the corresponding author.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599531

assignment between the control and treated groups, where the treatment is assigned with some explicit/implicit assignment policy manifested as correlations with some predictive covariates called confounders ($\mathbf{X}$ in Figure 1c) [32, 39]. As a consequence, direct regression of $\mathbf{T}$ on $\mathbf{Y}$ will introduce systematic bias without considering diverse treatment assignments across different groups [13]. To overcome this issue, the randomized control trial (RCT) provides the golden standard [5], while the ethical problems or the expensive practical cost become the obstacle to performing RCT in realistic cases. Fortunately, observational studies provide the practical alternative to infer the treatment effect from the non-randomized data [13, 39].

Despite remarkable progress, an important but easily overlooked problem raises in realistic applications: the collected covariates usually contain extra variables aside from the confounders $\mathbf{X}$, which profoundly affects the treatment effect estimation[9, 15] (as shown Figure 1b and 1c). Recalling the drug-recovery example, the drug analyzer often collects the covariates $\mathbf{U}$ to be as abundant as enough such that all the confounders $\mathbf{X}$ (e.g., gender or age) are observed ($\mathbf{X} \subseteq \mathbf{U}$). Meanwhile, extra variables are also collected into $\mathbf{U}$ besides $\mathbf{X}$, which is often divided into two types: (a) the treatment-only variables $\mathbf{I}$, denoting the extra variables partially correlating to the treatment $\mathbf{T}$ (e.g., income); (b) the outcome-only variables $\mathbf{Z}$, denoting the extra variables partially correlating to the outcome $\mathbf{Y}$ (e.g., living environment). According to previous literature [6, 8, 20], adjusting $\mathbf{I}$ will decrease the precision, while adjusting $\mathbf{Z}$ will benefit the estimation. Therefore, even though the estimation is unbiased ($\mathbf{X}$ belongs to the adjustment set), the choice of different adjustment features selected from the covariates still plays a vital role in determining the performance of ATE estimation.

However, due to the lack of prior guidance, it is a common practice to include each observed covariate into the adjustment feature set [21, 22], which we call the brute-force approach. Due to the (potential) large asymptotic variance, the brute-force approach is inefficient with poor empirical performance in some real-world cases [9]. To overcome this issue, previous approaches [9, 15] have attempted to separate the confounders from the precision variables (pre-outcome variables in Figure 1b, a special case of $\mathbf{Z}$ in Figure 1c) or instrumental variables (pre-treatment variables in Figure 1b, a special case of $\mathbf{I}$ in Figure 1c). However, two drawbacks prevent these strategies to be applied in realistic scenes. To be first, these settings only consider pre-treatment and pre-outcome variables, e.g., instrumental and precision variables, as shown in Figure 1b. Such kind of methods fails in more general settings when post-treatment and post-outcome variables exist in observed covariates. Second, these approaches are heuristics as they cannot clarify what adjustment features are expected by their methods and how the selected adjustment features affect the estimation, while our approach is well supported by semi-parametric inference theory [20].

In this paper, we consider a more general problem setting as shown in Figure 1c by allowing $\mathbf{I}$ and $\mathbf{Z}$ to be pre-treatment/post-treatment and pre-outcome/post-outcome variables or both. To facilitate the efficiency analysis, we pose a prior assumption that there are no mediator variables (no variables are lying on the path from T to Y). To support the efficiency (variance) analysis, such an assumption is necessary as in previous works [20]. We target to separate the treatment-only variables $\mathbf{I}$ from the confounders $\mathbf{X}$ and outcome-only variables $\mathbf{Z}$ for more efficient ATE estimation.

To achieve this target, we draw inspiration from semi-parametric inference [20, 27] and establish a computationally tractable metric named **Optimal Adjustment Features (OAF)**, which empirically characterizes the asymptotic variance of the ATE estimation. Theoretically, in the non-parametric regime, we show that our OAF metric decreases within the supplementation of $\mathbf{Z}$ or the deletion of $\mathbf{I}$ into the adjustment set. Therefore, the minimization of the variance metric implies that the optimal adjustment feature set ($\{\mathbf{Z}, \mathbf{X}\}$) is selected, i.e., the estimator achieves a minimal asymptotic variance. As our OAF varies discretely within the change of adjustment features, we treat its minimization as a combinatorial optimization problem. Regarding optimization efficiency, we introduce reinforcement learning (RL) and propose a policy gradient-based optimization framework named OAF by **Policy Gradient (OAFP)**. More specifically, we construct the actor with an encoder-decoder model [4] to generate the binary feature mask on the original covariates, where the feature mask serves as the differentiable policy. On the other hand, the OAF metric plays the role of the reward function to guide the policy gradient (e.g., the update of the feature mask). In summary, our contributions are highlighted as follows:

i We propose a computational tractable metric, named OAF, to measure the optimality of the adjustment features for treatment effect estimation with a non-parametric theoretical guarantee;
ii We design an RL-based combinatorial optimization framework, named OAFP, to optimize the proposed OAF metric and generate the corresponding feature mask for selecting adjustment features;
iii Extensive results on both synthetic and real-world datasets verify that: (a) our method can efficiently search the optimal adjustment features, (b) the searched adjustment features significantly improves the precision of treatment effect estimation.

2 RELATED WORK

2.1 Confounder Balancing

To estimate ATE/CATE, statistical methods focus on balancing the confounder across different groups via diverse strategies, including reweighting [15], matching [25] or covariate alignment [2]. To overcomearehe model misspecification for the high-dimensional data, a bunch of machine learning methods is further combined to capture the non-linear relationships among variables [16, 19, 21, 27, 29, 33, 39]. In detail, the representative non-parametric approach is to discretely fit the potential outcome using a regression tree or random forest (e.g., CF tree or CF forest) [29]. The typical semi-parametric approaches includes TMLE [27], doubly-robust methods [14] and DragonNet [22], which is asymptotically unbiased and efficient. The mainstream of deep methods models the confounder balancing as the domain adaptation problem, which learns the group invariant representation by minimizing the distribution divergence across different treatment arms [21, 33]. Besides, some methods also use sample-wise reweighting to make treatment and confounder independent in the representation space [19].

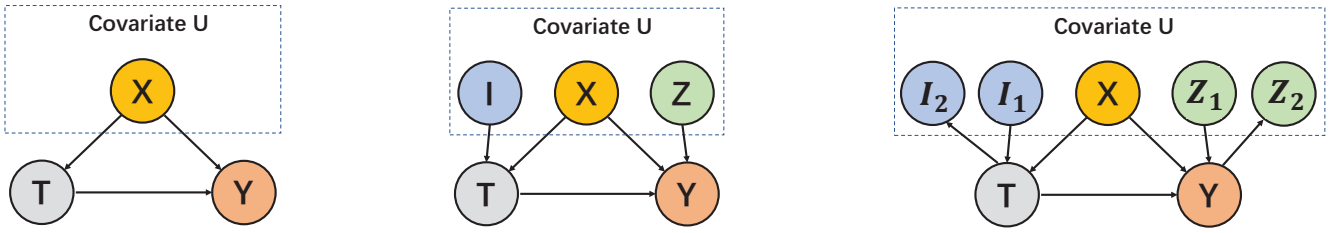


Figure 1: The distinction among the settings, where arrows and dashed lines refer to the causal relationship and the correlation, respectively. (a) Setting of brute-force adjustment, where each covariate is considered as the confounder for adjustment. (b) Setting of the previous separation approach, which only allows I and Z to be pre-treatment and pre-outcome variables. (c) Setting of our approach, which I and Z to be pre-treatment/post-treatment and pre-outcome/post-outcome or both. Meanwhile, the correlations between I, Z and X are allowed as well.

2.2 Covariate Separation

Recent methods have already noticed the problem of separating confounder from the instrumental/precision variables [9, 15]. For instance, [15] proposed a data-driven variance reduction approach [15] named DVD to separate the confounders from the precision variables, while DVD does not consider the treatment-only variables. To overcome this gap, [9] introduces the instrumental variables with non-linear deep networks to achieve disentanglement in the representation space. However, our paper contrasts the above-mentioned methods from three aspects: (a) they only consider the instrumental variables and precision variables, while we allow a much broader setting for **I** and **Z** in Figure 1c, (b) they are lack of theoretical understanding on how their methods achieve variable separation for better ATE estimation, while our method is well supported by the semi-parametric inference theory; (c) [9] achieves disentanglement in the representation space; while our methods directly separate **I** from {**Z**,**X**} among the original covariates.

2.3 Reinforcement Learning for Combinatorial Optimization

Beyond sequential decision-making tasks (e.g., MDP), recent advances on reinforcement Learning (RL) has brought new opportunity for combinatorial optimization (CO) problems [4, 37]. Originally, [4] proposed a policy-gradient-based framework to solve the Travelling Salesman Problem (TSP). Traditional approximation methods for NP-hard CO often require some parametric assumption, such as sub-modular objectives [26]. By contrast, the RL-based framework fits arbitrary objective functions such that they could solve more general CO problems. For instance, [37] has adopted the RL-based CO method for causal discovery by differentiably searching the causal DAG. Our paper adopts the RL-based CO framework to minimize the proposed variance metric. Although the metric is a function of different sets, it does not share some ideal properties such as sub-modular. Hence, traditional greedy methods cannot be applied to optimize our metric, while the RL-based CO framework is a proper candidate.

3 PROBLEM SETUP

Notations For concreteness, we consider the estimation of the average effect of a binary treatment. Suppose the data we own is generated independently and identically: $\{Y_i, U_i, T_i\}_{i=1}^{n} \sim P^j$, where P^j,n and $\mathbf{U}$ refer to the underlying joint distribution density, the sample size, and the collected covariates, respectively. Following notations in [13], we define the potential outcome under the treatment arm $\mathbf{T} = \mathbf{t}$ as $\mathbf{Y}(\mathbf{t})$ (We use upper-case (e.g. $\mathbf{T}$) to denote random variables, and lower-case (e.g. $\mathbf{t}$) for realizations.). Then the average treatment effect (ATE) equals to the expected difference between the treated ($\mathbf{T} = 1$) and the control ($\mathbf{T} = 0$) groups: $\gamma(P) = \mathbb{E}[\mathbf{Y}(\mathbf{T} = 1) - \mathbf{Y}(\mathbf{T} = 0)]$, where we refer ATE as $\gamma(P)$ for the convenience of later analysis. Given the collected covariates $\mathbf{U} = \{\mathbf{Z}, \mathbf{X}, \mathbf{I}\}$, one has to select $\mathbf{V} \subseteq \mathbf{U}$ as the adjustment feature set for ATE estimation. To facilitate the efficiency analysis, we pose a prior assumption that there is no mediator variables (no variables are lying on the path from T to Y).

Basic Assumptions To guarantee the validity of $\mathbf{V}$, three prior assumptions should be satisfied: [a] **Stable Unit Treatment Value**: $Y_i(\mathbf{t})$ for sample i is independent of the treatment assignments on sample $j \neq i$; [b] **Unconfoundedness**: $Y(\mathbf{t}) \perp\!\!\!\perp \mathbf{T} \mid \mathbf{V}$; [c] **Overlap**: For arbitrary $\mathbf{V} \in \mathcal{V}$, $p(\mathbf{t} \mid \mathbf{V})$ for $\mathbf{t} \in \{0, 1\}$, where $\mathcal{V}$ is the domain of $\mathbf{V}$. When the above-mentioned assumptions are mentioned, the selected $\mathbf{V}$ supports the unbiased estimation of ATE via diverse methods. For instance, the outcome regression (stratification) estimate $\gamma(P) = m_{\mathbf{V}}^{\mathbf{T}=1}(\mathbf{Y}) - m_{\mathbf{V}}^{\mathbf{T}=0}(\mathbf{Y})$, where $m_{\mathbf{V}}^{\mathbf{T}=\mathbf{t}}(\mathbf{Y}) = \mathbb{E}[\mathbf{Y} \mid \mathbf{T} = \mathbf{t}, \mathbf{V}]$ refers to the conditional outcome. Alternatives include using the propensity score $\pi^{\mathbf{T}}(\mathbf{V}) = P(\mathbf{T} = \mathbf{t} \mid \mathbf{V})$ for inverse-reweighting. The adjustment set $\mathbf{V}$ satisfying the above three principles is valid, and invalid otherwise.

No Mediator Assumption Notably, we allow post-treatment and post-outcome variables to be contained in the observed covariates U in this paper, which breaks the constraints of previous works [9, 15, 31] and can deal with more general problem settings. However, it does not mean any post-treatment variables are allowed. To be specific, variables that are simultaneously the descendants of T and ancestors of Y, e.g., the mediator variables, are not allowed [20]. In other words, in our causal graph, no variables lie on any path from T to Y. Posing such an assumption is necessary to

derive the efficiency analysis, while also bringing limitations to our methods. However, compared to previous research, our method has made a breakthrough in allowing post-treatment and post-outcome variables.

4 ESTABLISHING THE VARIANCE METRIC

4.1 Semi-parametric Inference for ATE Estimation.

Beyond estimating the whole underlying distribution P, previous literature in semi-parametric inference [27] concerns estimating the ATE parameter γ as a function of the underlying density P. Moreover, we denote the estimated density from $\{Y_i, U_i, T_i\}_{i=1}^n$ as $\hat{P}$ (via diverse machine learning methods) and the empirical distribution of P as P_n. In the case that γ is pathwise differentiable to P (this holds for ATE) and the underlying statistical model is convex, the following convergence result is obtained through Central Limit Theorem (CLT) once one of $\pi^{\mathbf{T}}(\mathbf{V})$ and $m_{\mathbf{V}}^{\mathbf{T}}$ is consistent:

$$\sqrt{n}\left(\gamma\left(\widehat{P}\right)-\gamma(P)\right) \xrightarrow{d} N(0, \mathrm{Var}[D^{\mathrm{eff}}(\mathbf{V})]), \tag{1}$$

where $\xrightarrow{d}$ refers to the convergence in distribution. The function $D^{\mathrm{eff}}(\mathbf{V})$ of $\mathbf{V}$ denotes the efficient influence curve [27], which has an unique expression [12]:

$$\frac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V})}(\mathbf{Y}-m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}))+m_{\mathbf{V}}^{\mathbf{T}=1}(\mathbf{Y})-m_{\mathbf{V}}^{\mathbf{T}=0}(\mathbf{Y})-\gamma(P). \tag{2}$$

The above conclusion reflects two critical intuitions: (a) the Cauchy–Schwarz inequality and Cramer-Rao bound [27] guarantees that $D^{\mathrm{eff}}(\mathbf{V})$ achieves the efficient estimation (with optimal asymptotic variance as $\mathrm{Var}[D^{\mathrm{eff}}(\mathbf{V})]$) of γ concerning each $\mathbf{V}$; (b) different $\mathbf{V}$ determines different $\mathrm{Var}[D^{\mathrm{eff}}(\mathbf{V})]$, which further determines the ATE estimation. However, these pure theoretical results suffer from two drawbacks: (a) they require prior causal graphs to guide the choice of adjustment sets, which is not realistic for practical applications; (b) the formulation of $\mathrm{Var}[D^{\mathrm{eff}}(\mathbf{V})]$ is too complicated for computation, which requires further simplification.

4.2 Optimality of Adjustment Features

Previous theoretical research has already established the connection between the optimality of the adjustment features $\mathbf{V}$ and the minimization of asymptotic variance: $\mathbf{Var}[D^{\mathbf{eff}}(\mathbf{V})]$ **is minimized if and only if** $\mathbf{V}=\{\mathbf{X},\mathbf{Z}\}$ [8, 20]. Although all these methods share unbiased estimation, their empirical performance is distinct from each other. This is caused by that different estimators use different adjustment features with asymptotic variances such that they have diverse empirical bias (see in Lemma 4).

Overall, we judge whether a set of features is optimal if and only if the corresponding estimator achieves the minimal asymptotic variance [20]. Notably, if several sets of features share the minimal asymptotic variance, they are all considered as optimal adjustment features.

4.3 Theoretical Properties of OAF Metric

Different from [20], we adopt the decomposed version (in Chapter 6.2 in [28]) of the efficient influence curve (2) as $D_d^{\mathrm{eff}}(\mathbf{V})$, which is computationally tractable and also satisfies the linear asymptotic results in (1):

$$D_d^{\mathrm{eff}}(\mathbf{V})=\frac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V})}(\mathbf{Y}-m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y})), \tag{3}$$

where the validity of such decomposition is supported in the following lemma:

Lemma 4.1 (Validity of $D_d^{\mathrm{eff}}(\mathbf{V})$). *Similar to D^{eff}, $\widehat{\gamma}(P)$ is asymptotically linear with D_d^{eff}, and $\sqrt{n}\left(\gamma\left(\widehat{P}\right)-\gamma(P)\right) \xrightarrow{d} N(0, Var[D_d^{eff}(\mathbf{V})])$.*

Moreover, we strengthen the viewpoint that the asymptotic variance is critical for the precision of estimating ATE in the case of finite samples using the following proposition:

Lemma 4.2. *Suppose that the cumulative distribution function F_n of $\gamma\left(\widehat{P}\right)-\gamma(P)$ is continuous within the sample size n increasing, then for any $\alpha \geq 0$ and n,*

$$P(|\gamma\left(\widehat{P}\right)-\gamma(P)| \geq \alpha) \leq \delta_n + 1 - F\left(\frac{\sqrt{n}\alpha}{\sqrt{Var[D_d^{eff}(\mathbf{V})]}}\right), \tag{4}$$

where F refers to the cumulative distribution function of the normal distribution $N(0,1)$ and $\delta_n = \sup|F_n - F|$ describes the pointwise convergence of $\{F_n\}$ to F with increasing n. According to the above lemma, we conclude that smaller $\mathrm{Var}[D_d^{\mathrm{eff}}(\mathbf{V})]$ implies the smaller right-side in (4), which further results in more precise $\gamma\left(\widehat{P}\right)$. Therefore, choosing different adjustment features $\mathbf{V}$ from the covariate set $\mathbf{U}$ determines different asymptotic variance $\mathrm{Var}[D_d^{\mathrm{eff}}(\mathbf{V})]$, which further affects the precision of ATE estimation. Naturally, we propose our metric named **Optimal Adjustment Features (OAF)**, as a functional of the adjustment features $\mathbf{V} \mapsto R_+$ ::

$$\begin{aligned}\mathbf{R}^{\mathrm{OAF}}(\mathbf{V}) &= \mathrm{Var}[D_d^{\mathrm{eff}}(\mathbf{V})] \\ &= \mathrm{Var}[\frac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V})}(\mathbf{Y}-m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}))].\end{aligned} \tag{5}$$

Nevertheless, one might be still confused about how $\mathcal{R}^{\mathrm{OAF}}$ varies within $\mathbf{V}$ changing. We provide theoretical insights to answer this problem using the following theorem:

Theorem 4.3 (Connections between $\mathcal{R}^{\mathrm{OAF}}$ and $\mathbf{V}$). *We denote the selected features for adjustment as $\mathbf{V} \subseteq \{\mathbf{X}\cup\mathbf{I}\cup\mathbf{Z}\}$. Meanwhile, we denote the optimal adjustment set as $\mathbf{V}_0 = \{\mathbf{X}\cup\mathbf{Z}\}$. Then the optimality of our reward is stated from the following three sub-theorems:*

(a) *If $\mathbf{V}$ is a valid adjustment set, then $\mathbf{R}^{OAF}(\mathbf{V}') \leq \mathcal{R}^{OAF}(\mathbf{V})$ holds for $\mathbf{V}' = \mathbf{V}\cup\mathbf{Z}'$, where $\mathbf{Z}' \subseteq \mathbf{Z}$.*
(b) *If $\mathbf{V}$ is a valid adjustment set, then $\mathbf{R}^{OAF}(\mathbf{V}) \leq \mathcal{R}^{OAF}(\mathbf{V}')$ holds for any $\mathbf{V}' = \mathbf{V}\cup\mathbf{I}'$, where $\mathbf{I}' \subseteq \mathbf{I}$.*
(c) *We assume that the $\{\mathbf{X}\cup\mathbf{I}\cup\mathbf{Z}\}$ contains all the parents of $\mathbf{Y}$, which implies that $\mathbf{Z}$ contains all the outcome-precision variables of $\mathbf{Y}$. Then $\mathbf{R}^{OAF}(\mathbf{V_0}) \leq \mathcal{R}^{OAF}(\mathbf{V}')$ holds for any $\mathbf{V}'$ which is not a valid adjustment set.*

Remark. Overall, $\mathbf{R}^{\mathrm{OAF}}(\mathbf{V}) = \mathrm{Var}[D_d^{\mathrm{eff}}(\mathbf{V})]$ achieves the minimum if $\mathbf{V} = \{\mathbf{X}\cup\mathbf{Z}\}$. Meanwhile, we argue that if $\mathbf{R}^{\mathrm{OAF}}(\mathbf{V}) = \mathrm{Var}[D_d^{\mathrm{eff}}(\mathbf{V})]$ then $\mathbf{V} = \{\mathbf{X}\cup\mathbf{Z}\}$ is the optimal adjustment features. To be specific, $\mathbf{V}$ must equal to $\{\mathbf{X},\mathbf{Z}\}$ when $\mathbf{R}^{\mathrm{OAF}}(\mathbf{V})$ achieves the minimum in the case that all the inequalities in Theorem 4.3

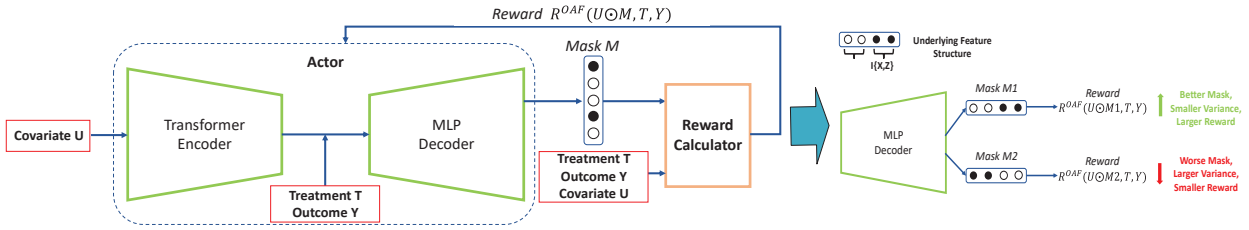


Figure 2: The framework of our OAFP method.

strictly hold (otherwise $\mathbf{R}^{\mathrm{OAF}}(\{\mathbf{X},\mathbf{Z}\}) < \mathcal{R}^{\mathrm{OAF}}(\mathbf{V})$ contradicts the assumption). The case that some equalities hold is meaningless since Lemma 4 implies that any valid adjustment features achieve the minimal asymptotic variance is optimal for ATE estimation. Finally, we claim that the proposed $\mathbf{R}^{\mathrm{OAF}}(\mathbf{V})$ achieves the minimum if and only if $\mathbf{V} = \{\mathbf{X},\mathbf{Z}\}$ are the optimal adjustment features.

4.4 Empirical Estimation for Computation

Recalling the empirical data $\{Y_i, U_i, T_i\}_{i=1}^n$, it is necessary to find an unbiased estimation of $\mathbf{R}^{\mathrm{OAF}}(\mathbf{V}) = \mathrm{Var}[D_d^{\mathrm{eff}}(\mathbf{V})]$ in the case of finite samples. Fortunately, the M-estimation theory [24] provides the empirical sandwich estimator as an unbiased solution. Although the influence curve approach is more general than the M-estimator approach, they are equivalent in the case of ATE estimation [24]. More specifically, supposing that $\widehat{\pi^{\mathbf{T}}(\mathbf{V})}$ and $\widehat{m_{\mathbf{V}}^{\mathbf{T}=\mathbf{1}}(\mathbf{Y})}$ represents the estimated propensity score and the conditional outcome, respectively, the corresponding empirical M-estimator can be written as $\widehat{\phi}(\gamma) = \frac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\widehat{\pi^{\mathbf{T}}(\mathbf{V})}}(\mathbf{Y} - \widehat{\pi^{\mathbf{T}}(\mathbf{V})})$, where the "sandwitch" terms can be further calculated as $\widehat{A}(\gamma) = I$ (I is identity matrix) and $\widehat{B}(\gamma) = \frac{1}{n}\sum_{i=1}^n \widehat{\phi}_{i=1}(\gamma)^2$. Finally, the empirical estimation of our metric, namely $\widehat{\mathbf{R}}^{\mathrm{OAF}}(\mathbf{V})$, is derived as follows:

$$\begin{aligned}\widehat{\mathbf{R}}^{\mathrm{OAF}}(\mathbf{V}) &= \widehat{A}(\gamma)\widehat{B}(\gamma)\widehat{A}(\gamma)^T \\ &= \frac{1}{n}\sum_{i=1}^n \left(\frac{\mathcal{I}(\mathbf{t}_i=1)-\mathcal{I}(\mathbf{t}_i=0)}{\widehat{\pi^{\mathbf{t}_i}(\mathbf{v}_i)}}(\mathbf{y}_i - \widehat{m_{\mathbf{V}}^{\mathbf{T}=\mathbf{t}_i}(\mathbf{Y})})\right)^2. \end{aligned} \tag{6}$$

Remark In fact, the term in (6) is similar to the additional term of doubly-robust methods (e.g., AIPW) or the iteration term in TMLE, as both TMLE and AIPW tune the estimator or estimated distributions to compensate for the term $P_n D_d^{\mathrm{eff}}(\mathbf{V})$ [12] such that the error term convergences to a zero-mean Gaussian distribution.

5 POLICY-GRADIENT BASED COMBINATORIAL OPTIMIZATION

As mentioned above, the empirical variance metric $\widehat{\mathbf{R}}^{\mathrm{OAF}}(\mathbf{V},\mathbf{T},\mathbf{Y})$ in (6) varies discretely with different adjustment features $\mathbf{V}$, where we rewrite $\widehat{\mathbf{R}}^{\mathrm{OAF}}(\mathbf{V},\mathbf{T},\mathbf{Y})$ here to strengthen the point that the calculation of $\widehat{\mathcal{R}}$ depends on $\mathbf{T},\mathbf{Y}$ as well. Hence, it is difficult to optimize $\widehat{\mathbf{R}}^{\mathrm{OAF}}(\mathbf{V},\mathbf{T},\mathbf{Y})$ in a differentiable approach. As an alternative, we consider the minimization of $\widehat{\mathcal{R}}^{\mathrm{OAF}}$ as a combinatorial optimization problem.

Motivated by the recent advances in neural combinatorial search area [4, 37], we introduce the reinforcement learning (RL) to efficiently search $\mathbf{V}$. To this end, we define the binary feature mask $\mathbf{M}$ on the original covariates $\mathbf{U} = \{\mathbf{Z},\mathbf{X},\mathbf{I}\}$ such that the ultimate goal is to find $\mathbf{M}$ corresponding to the optimal adjustment features $\{\mathbf{Z},\mathbf{X}\}$. We suppose the policy for mask generation is $q_\Phi(\cdot \mid \{\mathbf{T},\mathbf{Y},\mathbf{U}\})$, where Φ is the network parameter. Then the expected reward is defined to be our training objective as follows:

$$J(\psi \mid \mathbf{s}) = \mathbb{E}_{\mathbf{M}\sim q_\Phi(\cdot\mid\{\mathbf{T},\mathbf{Y},\mathbf{U}\})} - \mathcal{R}^{\mathrm{OAF}}(\mathbf{U}\odot\mathbf{M},\mathbf{T},\mathbf{Y}), \tag{7}$$

where s is the joint of t, y, u, and we use the notation $\odot$ to denote the selection of $\mathbf{V} = \mathbf{U}\odot\mathbf{M}$. In detail, we adopt the policy gradient method with variance reduction (reinforcement) to optimize the objective in (7), where the total framework is named **OAF by Policy Gradient (OAFP)** (as shown in Figure 2). Previous work for combinatorial optimization adopts the parametric approach by building a critic network to estimate the reward and reduce the variance [4, 37]. However, the critic can estimate the reward accurately only when the reward design is relatively simple (e.g., the traveling salesman problem [4]). By contrast, the $\widehat{\mathbf{R}}^{\mathrm{OAF}}(\mathbf{V},\mathbf{T},\mathbf{Y})$ in our problem is more complex, which is calculated upon two estimators $\widehat{\pi^{\mathbf{T}}(\mathbf{V})}$ and $\widehat{m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y})}$. Therefore, we alternatively use the non-parametric approach as reinforcement [30] to calculate the gradient of $J(\psi \mid \mathbf{s})$ concerning ψ. Moreover. we also add an entropy regularization term to encourage the exploration of the actor during the search process [37].

Regarding the implementations, we follow previous paradigms [4] and build the actor-network in the encoder-decoder architecture, as shown in Figure 2. The encoder is a multi-block transformer and the decoder is a Multi-layer-perception (MLP) perception. We leave the detailed settings of the actor-network in the appendix. To improve the efficiency during the optimization, we sample K arrays $\{B_1, B_2, \ldots, B_K\}$ as a batch, where $B_i = \{\mathbf{t}_i, \mathbf{u}_i, \mathbf{y}_i\}_{i=1}^{n_b}$ with n_b as the sample size for each array. As such operation implies the computation of $\widehat{\mathbf{R}}^{\mathrm{OAF}}(\mathbf{V},\mathbf{T},\mathbf{Y})_i$ for each B_i, calculating the reward becomes more time-consuming than updating the actor-network, especially in the case that $\widehat{\pi^{\mathbf{T}}(\mathbf{V})}$ and $\widehat{m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y})}$ are non-linear estimators. To alleviate this problem, we training $\widehat{\pi^{\mathbf{T}}(\mathbf{V})}$ and $\widehat{m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y})}$ in parallel with multiples processes.

Remark Our OAFP framework is designed for searching the adjustment feature set, which is a combinatorial optimization problem. Notably, although RL is typically applied in a sequential decision-making context, we do not use it to tackle any sequential problem. Alternatively, RL is also widely adopted to achieve non-sequential combinatorial optimization problems, such as causal discovery or the Travelling Salesman Problem [4, 37]. Detailed procedure of our OAFP framework is provided in the Algorithm 1.

Algorithm 1 Training procedure of OAFP Framework

Input The dataset $\mathcal{D} = \{t_i^p, y_i^p, u_i^p\}_{i=1}^n$, the transformer encoder network ϕ, and the MLP decoder network ψ, the iterations number $\mathcal{I}$.
for itr = 1 to $\mathcal{I}$ **do**
  Sampling $\{t_i^p, y_i^p, u_i^p\}_{i=1}^{n_b}$ for a mini-batch;
  Compute the mask $M = \psi(\phi(\mathbf{t}, \mathbf{y}, \mathbf{u}))$;
  Compute the selected features $\mathbf{v} = \mathbf{M} \odot \mathbf{u}$;
  Estimate $\widehat{\pi^{\mathbf{T}}(\mathbf{v})}$ and $\widehat{m_{\mathbf{v}}^{\mathbf{T}}(\mathbf{Y})}$;
  Calculate the reward with $\widehat{\pi^{\mathbf{T}}(\mathbf{v})}$ and $\widehat{m_{\mathbf{v}}^{\mathbf{T}}(\mathbf{Y})}$ as in (6);
  Update the whole network $\{\phi, \psi\}$ based on the policy gradient in (7);
end for
Output Select adjustment features $\mathbf{v}$ based on the final mask M.

6 EXPERIMENTS

6.1 Benchmarks

To evaluate the effectiveness of the proposed method, we conduct experiments on three datasets including the synthetic data, the semi-synthetic IHDP dataset [11] and the real-world Twins dataset [1], respectively. Details are present in the appendix for saving space.

Synthetic. Our synthetic datasets are generated according to the following process, which takes as input the total sample size N, the feature dimension d of the covariate U. In general, we first generate the pre-treatment part of $\mathbf{Z}$, the confounders $\mathbf{X}$ with the post-treatment part of $\mathbf{I}$. Then $\mathbf{Y}$ and $\mathbf{T}$ are generated, where the post-treatment part of $\mathbf{I}$ and the post-outcome part of $\mathbf{Z}$ are further generated. To be specific, we first generate $\mathbf{X}$ with feature size d_x, the pre-treatment $\mathbf{I}^e$ with size d_{Ie} and the pre-outcome $\mathbf{Z}^e$ with size d_{Ze}:

$$\mathbf{X}_1, \cdots, \mathbf{X}_{d_x}, \mathbf{Z}_1^e, \cdots, \mathbf{Z}_{d_{ze}}^e, \mathbf{I}_1^e, \cdots, \mathbf{I}_{d_{Ie}}^e \overset{iid}{\sim} \mathcal{N}(0,1) \tag{8}$$

The treatment $\mathbf{T}$ is then sampled from the logistic transformation of $\mathbf{I}^e$ and $\mathbf{X}$ as

$$\mathbf{T} \sim \text{ Bernoulli } \left(1/\left(1+\exp\left(-\left(\mathbf{I}^T\mathbf{X} + \mathbf{I}^T\mathbf{I}^e\right)\cdot r\right)\right)\right), \tag{9}$$

where $r = \frac{d_x + d_{Ie}}{20}$ is the scaling factor. Meanwhile, following previous protocols [15], the outcome $\mathbf{Y}$ is generated under both the linear and non-linear setting. More specifically, the linear generation of $\mathbf{Y}$ is

$$Y = \mathbf{X}\beta_{xy} + \mathbf{z}^e\beta^{zy} + \mathbf{T} + \sigma^Y, \tag{10}$$

where the non-linear generation is

$$Y = \mathbf{X}\beta_{xy} + \sum_{i=1}^{d_{ze}} \mathbf{z}_i^e \mathbf{z}_{i+1}^e \cdot \beta_i^{zy} + \mathbf{T} + \sigma^Y, \tag{11}$$

where the term $i+1$ is modulated by d_{ze}. Furthermore, the post-treatment variables $\mathbf{I}^o$ and the post-outcome variables $\mathbf{Z}^o$ are generated as $\mathbf{I}^o = \beta^{Io}\mathbf{T} + \sigma^{Io}$ and $\mathbf{Z}^o = \beta^{Zo}\mathbf{Y} + \sigma^{Zo}$. Overall, the ATE for the synthetic dataset is 1 and the covariate is $\mathbf{U} = \{\mathbf{X}, \mathbf{Z}^e, \mathbf{Z}^o, \mathbf{I}^e, \mathbf{I}^o\}$. To increase the challenging of separating $\mathbf{I}$ from $\{\mathbf{Z}, \mathbf{X}\}$, we set $d_{Io} = 0.3d$, $d_{Ie} = 0.2d$, $d_X = 0.3d$, $d_{Zo} = 0.1d$ and $d_{Ze} = 0.1d$ by enlarging the ratio of $\mathbf{I}$. Besides, the sample size N is set to 2000. We set the coefficient $\beta^{xy} = 4$ and $\beta^{zy} = -2$ for more significant difference between the effects of $\mathbf{X}$ and $\mathbf{Z}^e$ on $\mathbf{Y}$. Meanwhile, we sample β^{Io}, β^{Zo} from $U(0,1)$, together with σ^Y, σ^{Io} and σ^{Zo} sampled from $N(0,2)$.

IHDP. Based on the original RCT data, the selection bias is introduced by [11] via removing a non-random subset of the treated population. The resulting dataset contains 747 instances (608 control, 139 treated) with 25 covariates collected from the real-world [22]. We follow the classical surface-B setting in [11] to generate the IHDP dataset with the real-world 25 covariates. In detail, we set 5 continuous covariates (as all) as the confounders $\mathbf{X}$. Meanwhile, we randomly select half of the rest 20 discrete variables as $\mathbf{I}$, with the rest as $\mathbf{Z}$. To this end, we select $\mathbf{Z}$ or $\mathbf{I}$ from the Bernouli distribution $B(0.5)$ for each discrete variable. The effect coefficients of $\mathbf{X}$ and $\mathbf{Z}$ on $\mathbf{Y}$, namely β_{xy} and β_{zy}, is generated in the same protocol in [11]. The effect coefficients of $\mathbf{I}$ and $\mathbf{X}$ on $\mathbf{T}$, namely β_{it} and β_{xt}, are generated from $U(-2,2)$ as the uniform distributions. Furthermore, the $\mathbf{Y}_1$, $\mathbf{Y}_0$ and $\mathbf{T}$ are generated as follows:

$$\left\{\begin{array}{l} \mathbf{Y}_1 = \beta_{xy}^T\mathbf{X} + \beta_{zy}^T\mathbf{Z} - \omega + N(0,1), \\ \mathbf{Y}_0 = \exp(\beta_{xy}^T\mathbf{X} + \beta_{zy}^T\mathbf{Z}) + N(0,1), \\ \mathbf{T} \sim \text{ Bernoulli } \left(1/\left(1+\exp\left(-\left(\beta_{xt}^T\mathbf{X} + \beta_{it}^T\mathbf{I}\right)\right)\right)\right) \end{array}\right\}, \tag{12}$$

where ω refers to the term to keep the Average Treatment Effect on the Treated (ATT) close to 4 [11]. As the covariates $\mathbf{X}$ are fixed, we do not distinguish pre-outcome and post-outcome variables in IHDP.

Twins. The original Twins dataset is derived from all twins born in the USA between the year 1989 and 1991 [1]. Following previous protocol [22], we consider 28 variables related to parents, pregnancy, and birth, where the outcome is the children's mortality after one year. We introduce 5 pre-treatment $\mathbf{I}^e$ and 5 post-treatment $\mathbf{I}^o$ by adding them to covariates, resulting in the 38-dimension covariates. In detail, we sample $\mathbf{I}^e$ from $N(0,1)$. Then $\mathbf{T}$ is sampled from the Bernoulli-logistic approach as follows:

$$\mathbf{T} \sim \text{ Bernoulli } \left(1/\left(1+\exp\left(-\left(\beta_{xt}^T\mathbf{X} + \beta_{it}^T\mathbf{I}^e\right) + N(0,0.5)\right)\right)\right), \tag{13}$$

with β_{it} and β_{xt} sampled from $U(-2,2)$. Moreover, $\mathbf{I}^o$ is generated as $\mathbf{I}^o = \beta^{Io}\mathbf{T} + \sigma^{Io}$, with $\beta^{Io} \sim U(-2,2)$ and $\sigma^{Io} \sim N(0,0.5)$.

6.2 Baselines and Implementations

Baselines The baselines we compared in this paper can be summarized into three classes:

(a) Statistical methods, which include the direct difference method (Direct) [15], the inverse propensity score reweighting (IPW) [3], Augmented IPW (AIPW) [28] which is doubly robust to model misspecifications and the TMLE method [28] which starts from an initial distribution and iteratively updates the estimation;
(b) Machine Learning methods including the DragonNet [22] which is a deep version of TMLE, Generative adversarial Network (GANITE) [34] which designs a GAN-style framework to balance the confounders, the Bayesian regression Tree (BART) [11] which is a typical nonparametric with

uncertainty estimations, and the orthogonal regularized network (DNOUT) [10];

(c) Previous covariate disentanglement/separation methods including the LASSO regularized AIPW (AIPW-L), the DVD method in [15] which is a data-driven separation approach, the TEDVAE approach [36] which achieves promising disentanglement using autoencoder, the DR-CFR method in [9] which utilizes the mutual information to build a disentangle framework, and the multi-environment invariant method NICE [23].

Implementations of baselines For baselines we have compared in this paper, we exactly follow the optimal hyper-parameters with the original network architectures in their open-source implementations. Notably, the NICE [23] method requires multiple environments to support the identification of optimal adjustment features, where the different environment is generated using a distinct causal graph. For our problem, to adapt the NICE method, we randomly split the training data into three environments to simulate the heterogeneous training domains.

Implementations of our method. Roughly speaking, we implement both the linear and the non-linear versions of our method, respectively. For the linear implementation, we implement the $\widehat{\pi^{\mathbf{T}}(\mathbf{V})}$ as the logistic regression and $\widehat{m^{\mathbf{T}}_{\mathbf{V}}(\mathbf{Y})}$ as the linear regression to search the adjustment features. The downstream estimator for ATE estimation is the doubly-robust AIPW. For the non-linear implementation, we build a two-layer MLP as $\widehat{\pi^{\mathbf{T}}(\mathbf{V})}$ with a four-layer MLP as $\widehat{m^{\mathbf{T}}_{\mathbf{V}}(\mathbf{Y})}$for searching features, with the DragonNet as the downstream estimator for estimating ATE. To ease the notation, we name our method OAFP_L implemented for the linear case, and OAFP_N implemented for the non-linear case. Our implementation in Python will be released to the public once accepted.

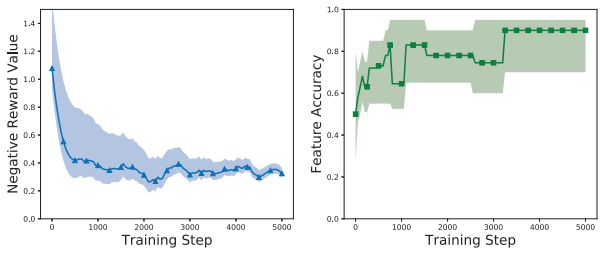
Metric. We mainly focus on two metrics: the bias of ATE and the accuracy of feature selection. The former metric is quantified by $\epsilon_{\text{ATE}} = |\text{ATE} - \widehat{\text{ATE}}|$, where $\text{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i^1 - \frac{1}{N}\sum_{j=1}^{N} Y_j^0$ is the underlying truth. Notably, as the underlying ATE for Twins is close to zero (0.025), we report the relative error as $\epsilon_{\text{ATE}} = \frac{|\text{ATE}-\widehat{\text{ATE}}|}{\text{ATE}}$ for Twins dataset. For the latter metric, we use $\text{Acc} = \frac{|\widehat{\mathbf{M}}-\mathbf{M}_0|_1}{d}$ to measure the feature accuracy, where $\widehat{\mathbf{M}}$ refers to the optimized feature mask and $\mathbf{M}^0$ refers to the ground truth feature mask with $\mathbf{M}_i^0 = 0$ when $\mathbf{U}_i \in \mathbf{I}$ and $\mathbf{M}_i^0 = 1$ otherwise.

6.3 Results and Analysis

In this section, we first propose three questions on the evaluation of the proposed OAFP method:

(a) Whether OAFP searches the adjustment features accurately;
(b) Whether the adjustment features searched by our OAFP achieve better ATE estimation compared to baselines;
(c) Whether the search process of OAFP is efficient on the time cost.

6.3.1 Results and Analysis on Searching Adjustment Features. To answer the first question, we report results on Results on feature search (Fs_Acc) with the relative error (R_err) in Table 4, where $\text{R\_err} = \frac{|R(\widehat{\mathbf{V}})-R(\mathbf{V})|}{R(\mathbf{V})}$ measure the relative distance between the optimal variance metric as $R(\mathbf{V})$ and the metric for our searched features $\widehat{\mathbf{V}}$ as $R(\widehat{\mathbf{V}})$. Notably, we search non-linear synthetic data, IHDP, and twins using the non-linear OAFP-N, while the linear synthetic dataset is searched using OAFP-L. The feature accuracy in Table 4 reflects that our method OAFP_L and OAFP_N successfully search the optimal adjustment features $\{\mathbf{Z}, \mathbf{X}\}$ in linear and non-linear cases, respectively. Meanwhile, the relative error of the variance metric R_err in Table 4 also reflects that our searched adjustment features achieve the empirical asymptotic variance close to the optimal one achieved by $\{\mathbf{Z}, \mathbf{X}\}$. Moreover, we provide an intuitive illustration of how the reward $\widehat{\mathcal{R}}$ and the feature accuracy vary in the training process during the search process in Figure 3a and Figure 3b, respectively. As shown in 3a, the average reward converges stably under the threshold at around 4000 steps, where the corresponding feature accuracy also achieves 95% after 3000 steps. Besides, the reasons behind that the feature accuracy cannot achieve 100% can be attributed to (a) error between the empirical $\widehat{\mathcal{R}}$in (6) and $\mathcal{R}$; (b) the effect of some covariates are too small in the underlying structural equation such that their existence or not is less important.



a Reward **b Accuracy**

Figure 3: Reward and feature accuracy curves in the setting of non-linear synthetic data with $d = 20$.

6.3.2 Results and Analysis on ATE estimation. We then report the downstream results on ATE estimation in Table 2, Table 3 (results on linear simulation is present in appendix), respectively. For results on non-linear simulation, our method, OAFP_L and OAFP_N, achieve significant improvement in the estimation performance compared to other baselines. Results on the semi-synthetic IHDP and the real-world Twins further verify the superiority of our method. Meanwhile, the poor performance for methods without covariate separation also strengthens our view that the existence of treatment-only variables $\mathbf{I}$ will hurt the ATE performance in finite-sample cases. **Notably, our linear implementation OAFP_L performs less accurately than some deep methods (e.g., DragonNet) due to the model misspecification problem for the IHDP dataset.**

6.3.3 Results on the efficiency of our method. To verify how our RL framework improves the searching efficiency, we compare the search process of our OAFP to that of the brute-force approach (traversing the powerset of $\mathbf{U}$ and find the minimal $\mathcal{R}$) in Table 5. Obviously, it is meaningful for us to design the RL framework as it significantly reduces the time cost for searching the optimal adjustment features. (The brute-force approach is even impossible when the feature dimension is larger than 20.)

Table 1: Linear simulation results. The metrics are Mean±STD over 10 repeated experiments.

Linear Simulation							
Settings		In_sample Prediction			Out_of_sample Prediction		
Feature Dimension		20	40	80	20	40	80
Statistical	Direct	5.33±0.53	6.88±0.69	8.66±0.85	5.54±1.41	7.36±1.26	6.25±1.84
	IPW	0.68±2.20	0.93±2.90	1.11±2.59	0.87±2.33	1.51±3.11	2.28±3.87
	AIPW	0.30±0.64	0.93±2.90	1.11±2.59	0.27±2.00	0.76±1.96	1.39±1.41
	TMLE	0.25±0.07	0.58±0.11	0.61±0.05	0.48±0.10	0.60±0.13	0.65±0.11
Machine	DragonNet	0.05±0.48	0.29±0.15	0.49±0.37	0.93±0.42	0.90±0.37	1.05±0.86
	GANITE	0.86±0.00	0.97±0.00	1.01±0.00	0.99±0.00	1.00±0.00	1.01±0.00
	DNOUT	0.46±0.03	0.51±0.03	0.64±0.14	0.40±0.02	0.49±0.04	0.62±0.11
	BART	1.02±0.12	2.13±0.15	2.56±0.61	1.51±1.00	1.99±0.28	2.87±0.57
Decomposed	AIPW_L	0.50±0.41	0.57±0.49	0.64±0.35	0.60±0.34	0.67±0.38	0.91±0.52
	DVD	1.02±0.15	0.74±0.46	0.84±0.14	1.06±0.04	0.70±0.00	0.91±0.00
	DR-CFR	0.66±0.26	0.44±0.18	0.30±0.13	0.69±0.12	0.68±0.08	0.46±0.06
	TEDVAE	0.31±0.01	0.41±0.02	0.54±0.02	0.32±0.03	0.46±0.03	0.51±0.02
	NICE	1.02±0.09	1.05±0.11	1.34±1.08	0.90±0.18	0.96±0.13	1.18±0.84
Ours	**OAFP_L**	**0.03±0.01**	**0.01±0.01**	**0.08±0.02**	**0.14±0.03**	**0.15±0.07**	**0.12±0.08**
	OAFP_N	**0.01±0.02**	**0.02±0.01**	**0.01±0.01**	**0.06±0.08**	**0.03±0.05**	**0.02±0.03**

Table 2: Non-Linear simulation results. The metrics are Mean±STD over 10 repeated experiments.

Settings		In_sample Prediction			Out_of_sample Prediction		
Feature Dimension		20	40	80	20	40	80
Statistical	Direct	4.69±0.62	7.09±0.68	8.92±0.76	5.23±0.41	6.28±1.41	9.28±1.32
	IPW	0.99±4.50	1.27±3.13	4.36±2.37	1.33±1.92	2.22±5.39	4.51±3.33
	AIPW	1.32±1.95	0.99±0.27	2.35±0.83	0.21±1.19	0.55±0.47	3.88±1.23
	TMLE	0.42±0.11	0.59±0.07	0.62±0.02	0.50±0.12	0.66±0.18	0.81±0.20
Machine	DragonNet	0.19±0.19	0.20±0.14	0.57±0.38	0.99±0.16	0.84±0.70	0.87±1.02
	GANITE	0.80±0.01	0.87±0.01	0.99±0.01	0.99±0.01	1.08±0.01	1.10±0.01
	DNOUT	0.47±0.01	0.62±0.04	0.92±0.09	0.50±0.02	0.61±0.05	0.95±0.09
	BART	0.92±0.20	2.03±0.27	2.89±0.98	0.92±0.20	2.25±0.16	2.98±1.10
Decomposed	AIPW_L	0.59±0.10	0.66±0.05	0.89±0.10	0.54±0.29	0.74±0.13	0.96±0.22
	DVD	0.95±0.03	0.83±0.01	0.76±0.01	1.06±0.08	0.64±0.01	1.05±0.73
	DR-CFR	0.88±0.08	1.18±0.16	2.08±0.69	1.28±0.08	1.69±0.73	1.52±0.51
	TEDVAE	0.37±0.01	0.43±0.02	0.55±0.03	0.38±0.03	0.49±0.04	0.60±0.02
	NICE	1.08±0.32	1.24±0.60	1.81±0.22	1.10±0.37	1.23±0.41	1.93±0.35
Ours	**OAFP_L**	**0.03±0.13**	**0.12±0.10**	**0.23±0.13**	**0.24±0.22**	**0.20±0.13**	**0.32±0.34**
	OAFP_N	**0.01±0.10**	**0.09±0.07**	**0.13±0.11**	**0.15±0.09**	**0.16±0.07**	**0.14±0.08**

Table 5: Comparison on the time cost (hours) of searching for features.

Method	Syn_20_l	Syn_40_l	Syn_20_n	Syn_40_n
Ours	0.22	0.27	28.055	41.94
Brute-force	11.65	$1.83 \cdot 10^7$	$2.94 \cdot 10^3$	$4.61 \cdot 10^9$

Comparison with a simple greedy method. To show the necessity and superiority of our RL-based optimization framework, we design another greedy-based feature selection baseline. To be specific, the greedy baseline starts from setting the whole set as the adjustment features. Then adding a variable decreases $\mathbf{R}^{OAF}$, then the baseline judges it as a variable from $\mathbf{V}$. Otherwise, the baseline judges it as a variable from $\mathbf{I}$. Finally, the baseline ends up at steps equal to the feature size. As shown in the following Table 6, we observe that the greedy baseline achieves very poor feature selection result compared to the optimal adjustment set $\{\mathbf{Z}, \mathbf{X}\}$. This is inherent to the theoretical fact that the optimization objective $\mathbf{R}^{OAF}$ is not a sub-modular function. As a consequence, the greedy baseline does not benefit from the near-optimal result [26].

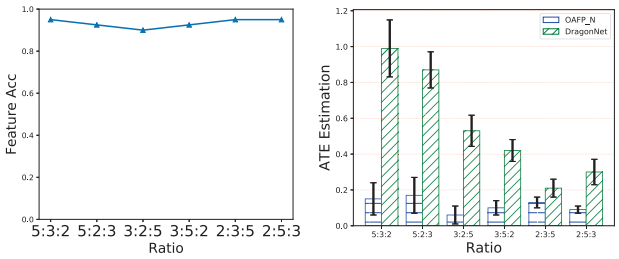
Ablation Studies. We conduct ablation studies on data simulation to examine the impact of variations in the ratio of $\mathbf{I}$. For our study, we use non-linear synthetic data with a total dimension of $d = 20$. The original ratio of $\mathbf{I}$:$\mathbf{X}$:$\mathbf{Z}$ is $5:3:2$, and we simulate five additional cases by adjusting the ratio: $5:2:3$, $3:2:5$, $3:5:2$, $2:3:5$, and $2:5:3$. Figure 4 shows the results, demonstrating that our method accurately selects the optimal adjustment set and estimates the ATE. Interestingly, as the ratio of $\mathbf{I}$ decreases, the

Table 3: Results on IHDP and Twins datasets. The metrics are Mean±STD over 10 repeated experiments. The best performance is marked in bold.

Benchmark		IHDP		Twins	
Settings		In_sample	Out_of_sample	In_sample	Out_of_sample
Statistical	Direct	3.36±3.70	3.70±3.36	1.50±0.03	4.34±0.14
	IPW	3.48±5.92	3.48±5.91	1.79±0.05	9.29±0.39
	AIPW	1.82±2.99	1.82±2.99	1.79±0.05	9.29±0.39
	TMLE	2.71±1.80	2.52±1.07	1.76±0.02	4.01±0.02
Machine	DragonNet	1.19±1.04	1.37±0.95	1.05±0.01	1.03±0.01
	GANITE	5.40±0.04	5.60±0.01	15.60±0.08	19.6±0.19
	DOUT	3.16±1.41	3.08±1.26	2.04±0.02	2.20±0.02
	BART	3.12±2.42	3.28±2.60	0.95±0.03	0.97±0.03
Decomposed	AIPW_L	1.85±2.64	1.85±2.64	1.03±0.03	3.35±0.12
	DVD	2.79±0.82	0.73±0.03	1.42±0.01	7.78±0.05
	DR-CFR	2.45±1.05	1.74±1.00	3.64±0.03	6.00±0.01
	TEDVAE	0.46±0.04	0.52±0.05	0.71±0.02	0.72±0.02
	NICE	2.75±3.91	2.68±2.25	42.92±0.02	53.12±1.84
Ours	**OAFP_L**	**1.14±0.47**	**1.24±0.33**	**0.65±0.02**	**1.98±0.06**
	OAFP_N	**0.28±0.07**	**0.29±0.09**	**0.30±0.01**	**0.43±0.01**

Table 4: Results on feature search, where S-20-l refers to the synthetic data with 20 features in the linear setting.

Dataset	Fs_Acc	R_err
S-20-l	95.0%	0.02
S-40-l	92.5%	0.04
S-80-l	90.0%	0.11
S-20-n	95.0%	0.06
S-40-n	95.0%	0.10
S-80-n	90.0%	0.13
IHDP	92.0%	0.11
Twins	94.7%	0.13



a Feature Accuracy **b ATE Estimation**

Figure 4: Results with different variable ratios.

performance gap between the original DragonNet and our OAFP_N narrows, supporting our claim that including **I** is detrimental for ATE estimation.

Table 6: Comparison on the greedy baseline (%), where Syn_20_l refers to synthetic data with 20 covariates generated under the linear setting. For each method, we report the feature selection accuracies on each dataset.

Method	Syn_20_l	Syn_40_l	Syn_20_n	Syn_40_n
Ours	98	96	94	90
Greedy	51	32	47	10

7 FUTURE WORKS AND CONCLUSION

This paper addresses the problem of estimating average treatment effect (ATE) from observational studies. We explore the benefits of separating treatment-only variables (**I**) and outcome-only variables (**Z**) from the collected covariates (**U**), along with the confounder X, using semi-parametric inference. We propose a variance metric to evaluate adjustment features and develop an RL-based framework for efficient optimization. Experimental results confirm the effectiveness of our method in identifying optimal features and accurately estimating ATE. However, two challenges require further attention: (a) Relaxing the no mediator assumption, which constrains the applicability of our method and calls for extending the analysis to mediators. (b) Estimating individualized treatment effects (ITE) remains an open and challenging problem due to the lack of a closed form of efficient influence curve.

7.1 Acknowledgement

This work was supported in part by National Natural Science Foundation of China (No.91948303-1, 62006207, U20A20387), Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), and Zhejiang Province Natural Science Foundation (LQ21F020020). We would also like to thank Dr. Hao Zou for productive discussions.

REFERENCES

[1] Douglas Almond, Kenneth Y Chay, and David S Lee. 2005. The costs of low birth weight. *The Quarterly Journal of Economics* 120, 3 (2005), 1031–1083.

[2] Susan Athey, Guido W Imbens, and Stefan Wager. 2018. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80, 4 (2018), 597–623.

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

[4] Irwan Bello, Hieu Pham, Quoc V Le, Mohammad Norouzi, and Samy Bengio. 2016. Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940* (2016).

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

[6] William G Cochran. 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* (1968), 295–313.

[7] Carlos Fernández-Loría and Foster Provost. 2022. Causal decision making and causal effect estimation are not the same… and why it matters. *INFORMS Journal on Data Science* (2022).

[8] Jinyong Hahn. 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* (1998), 315–331.

[9] Negar Hassanpour and Russell Greiner. 2019. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*.

[10] Tobias Hatt and Stefan Feuerriegel. 2021. Estimating average treatment effects via orthogonal regularization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 680–689.

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

[12] Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. 2022. Demystifying statistical learning based on efficient influence functions. *The American Statistician* (2022), 1–13.

[13] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[14] Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems* 33 (2020), 265–277.

[15] Kun Kuang, Peng Cui, Hao Zou, Bo Li, Jianrong Tao, Fei Wu, and Shiqiang Yang. 2020. Data-driven variable decomposition for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[16] Bryan Lim. 2018. Forecasting treatment responses over time using recurrent marginal structural networks. *advances in neural information processing systems* 31 (2018).

[17] Safoora Masoumi and Saeid Shahraz. 2022. Meta-analysis using Python: a hands-on tutorial. *BMC medical research methodology* 22, 1 (2022), 1–8.

[18] Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress* 19, 2 (2000).

[19] Zhaozhi Qian, Alicia Curth, and Mihaela van der Schaar. 2021. Estimating Multi-cause Treatment Effects via Single-cause Perturbation. *Advances in Neural Information Processing Systems* 34 (2021), 23754–23767.

[20] Andrea Rotnitzky and Ezequiel Smucler. 2020. Efficient Adjustment Sets for Population Average Causal Treatment Effect Estimation in Graphical Models. *J. Mach. Learn. Res.* 21, 188 (2020), 1–86.

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

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

[23] Claudia Shi, Victor Veitch, and David M Blei. 2021. Invariant representation learning for treatment effect estimation. In *Uncertainty in Artificial Intelligence*. PMLR, 1546–1555.

[24] Leonard A Stefanski and Dennis D Boos. 2002. The calculus of M-estimation. *The American Statistician* 56, 1 (2002), 29–38.

[25] Elizabeth A Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25, 1 (2010), 1.

[26] Stratis Tsirtsis and Manuel Gomez Rodriguez. 2020. Decisions, counterfactual explanations and strategic behavior. *Advances in Neural Information Processing Systems* 33 (2020), 16749–16760.

[27] Mark J Van der Laan, Sherri Rose, et al. 2011. *Targeted learning: causal inference for observational and experimental data*. Vol. 10. Springer.

[28] Mark J Van Der Laan and Daniel Rubin. 2006. Targeted maximum likelihood learning. *The international journal of biostatistics* 2, 1 (2006).

[29] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.

[30] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3 (1992), 229–256.

[31] Anpeng Wu, Junkun Yuan, Kun Kuang, Bo Li, Runze Wu, Qiang Zhu, Yue Ting Zhuang, and Fei Wu. 2022. Learning decomposed representations for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[32] Pengzhou Wu and Kenji Fukumizu. 2021. \beta-Intact-VAE: Identifying and Estimating Causal Effects under Limited Overlap. *arXiv preprint arXiv:2110.05225* (2021).

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

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

[35] Shengyu Zhang, Dong Yao, Zhou Zhao, Tat-Seng Chua, and Fei Wu. 2021. Causerec: Counterfactual user sequence synthesis for sequential recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 367–377.

[36] Weijia Zhang, Lin Liu, and Jiuyong Li. 2021. Treatment effect estimation with disentangled latent factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 10923–10930.

[37] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. 2019. Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477* (2019).

[38] Yueting Zhuang, Ming Cai, Xuelong Li, Xiangang Luo, Qiang Yang, and Fei Wu. 2020. The next breakthroughs of artificial intelligence: The interdisciplinary nature of AI. *Engineering* 6, 3 (2020), 245.

[39] Hao Zou, Bo Li, Jiangang Han, Shuiping Chen, Xuetao Ding, and Peng Cui. 2022. Counterfactual Prediction for Outcome-Oriented Treatments. In *International Conference on Machine Learning*. PMLR, 27693–27706.

A THEORETICAL PROOF

Proof of Lemma 4.1.

$$\begin{aligned}
&\mathbb{E}[D_d^{\text{eff}}(\mathbf{V})] \\
&= \mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V})}(\mathbf{Y}-m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}))\right] \\
&= \mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=1)}{\pi^{\mathbf{T}}(\mathbf{V})}(\mathbf{Y}-m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}))\right]-\mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V})}(\mathbf{Y}-m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}))\right],
\end{aligned}$$

We then expand the first term as follows:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=1)}{\pi^{\mathbf{T}}(\mathbf{V})}(\mathbf{Y}-m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}))\right] \\
&= \mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=1)}{\pi^{\mathbf{T}=1}(\mathbf{V})}(\mathbf{Y}(1)-m_{\mathbf{V}}^{\mathbf{T}=1}(\mathbf{Y}))\right] \\
&= \mathbb{E}_{\mathbf{V}}\mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=1)}{\pi^{\mathbf{T}=1}(\mathbf{V})}(\mathbf{Y}(1)-m_{\mathbf{V}}^{\mathbf{T}=1}(\mathbf{Y}))\mid\mathbf{V}\right] \\
&= \mathbb{E}_{\mathbf{V}}\left\{\mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=1)}{\pi^{\mathbf{T}=1}(\mathbf{V})}\mid\mathbf{V}\right]\mathbb{E}\left[(\mathbf{Y}(1)-m_{\mathbf{V}}^{\mathbf{T}=1}(\mathbf{Y}))\mid\mathbf{V}\right]\right\} \\
&= 0,
\end{aligned}$$

where the first equality is due to the consistency of $\mathbf{Y}$, the second equality is due to the tower property of expectation. Meanwhile, the third equality is due to the fact that $\mathbf{Y}(\mathbf{t}) \perp\!\!\!\perp \pi^{\mathbf{T}}(\mathbf{V}) \mid \mathbf{V}$. Finally, $\mathbb{E}\left[(\mathbf{Y}(1)-m_{\mathbf{V}}^{\mathbf{T}=1}(\mathbf{Y}))\mid\mathbf{V}\right]=0$ derives the last equality. Then (b) follows immediately:

$$\text{Var}[D_d^{\text{eff}}(\mathbf{V})] = \mathbb{E}[(D_d^{\text{eff}}(\mathbf{V}))^2] - (\mathbb{E}[D_d^{\text{eff}}(\mathbf{V})])^2 = \mathbb{E}[(D_d^{\text{eff}}(\mathbf{V}))^2].$$

□

Lemma A.1 (Validity of $D_d^{\text{eff}}(\mathbf{V})$). *Similar to D^{eff}, $\widehat{\gamma}(P)$ is asymptotically linear with D_d^{eff}, and $\sqrt{n}\left(\gamma\left(\widehat{P}\right)-\gamma(P)\right) \xrightarrow{d} N(0, Var[D_d^{eff}(\mathbf{V})])$.*

Proof. First, recalling the the original derivation of D^{eff}, a previous proved conclusion is provided such that the estimator $\widehat{\gamma}(P)$ is asymptotically linear with influence curve as D^{eff} [28]:

$$\gamma\left(\widehat{P}\right)-\gamma(P) = \frac{1}{n}\sum_{i=1}^{n} D_i^{\text{eff}}(\mathbf{V}) + \mathcal{O}(\frac{1}{\sqrt{n}}). \tag{14}$$

Moreover, the decomposition of D^{eff} is also proposed in [28]:

$$\begin{cases} D_d^{\text{eff}} = D^{\text{eff1}} = \dfrac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V})}(\mathbf{Y}-m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y})) \\ D^{\text{eff2}} = m_{\mathbf{V}}^{\mathbf{T}=1}(\mathbf{Y}) - m_{\mathbf{V}}^{\mathbf{T}=0}(\mathbf{Y}) - \gamma(P), \end{cases} \tag{15}$$

where the second term equals to zero under the integral of the empirical distribution P_n: $P_n D^{\text{eff2}} = 0$ [27]. Thus we conclude that $\widehat{\gamma}(P)$ is asymptotically linear with influence curve as D^{eff2}: $\frac{1}{n}\sum_{i=1}^{n} D_i^{\text{eff}}(\mathbf{V}) = \frac{1}{n}\sum_{i=1}^{n} D_i^{\text{eff2}}(\mathbf{V}) + \mathcal{O}(1)$. Meanwhile, combined with previous conclusion in Lemma that $\mathbb{E}[D_d^{\text{eff}}(\mathbf{V})]=0$, we have the following derivation:

$$\begin{aligned}
&\lim_{n\to+\inf}\sqrt{n}\left(\gamma\left(\widehat{P}\right)-\gamma(P)\right) \\
&= \lim_{n\to+\inf}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(D_i^{\text{eff1}}(\mathbf{V})+D_i^{\text{eff2}}(\mathbf{V})\right)+\sqrt{n}\mathcal{O}(\frac{1}{\sqrt{n}})\right\} \\
&= \lim_{n\to+\inf}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i^{\text{eff1}}(\mathbf{V})\right\},
\end{aligned} \tag{16}$$

Then the claim follows from the CLT. □

Proof of Lemma 4.2. We first claim that although we suppose the continuity of the CDF, the similar conclusion can be extended to CDFs with non-left-continuous points as well. As the term δ_n controls the convergence of the series $\{F_i\}_{i=1}^n$ to F, results in Lemma A.1 imply that for any α in the support of P, the following inequality holds:

$$|F(\alpha)-F_n(\alpha)| \le \delta_n \Longrightarrow 1-F_n(\alpha) \le 1-F(\alpha)+\delta_n. \tag{17}$$

where F is the CDF of the normal distribution $N(0, \text{Var}[D_d^{\text{eff}}(\mathbf{V})])$. Meanwhile, we observe that $N(0, \text{Var}[D_d^{\text{eff}}(\mathbf{V})]) \overset{d}{=} Z*\sqrt{\text{Var}[D_d^{\text{eff}}(\mathbf{V})]}$, where $\overset{d}{=}$ refers to the in-distribution equality and $Z \sim N(0,1)$. Therefore, the above inequality can be derived as follows:

$$\begin{aligned}
P(X \ge \alpha) &\le \delta_n + 1 - P(X \ge \alpha) \\
&\le \delta_n + 1 - P(Z * \sqrt{\text{Var}[D_d^{\text{eff}}(\mathbf{V})]} \ge \alpha),
\end{aligned} \tag{18}$$

where the final conclusion is obtained when we further let $X = \sqrt{n}|\gamma\left(\widehat{P}\right)-\gamma(P)|$ and $\alpha_0 = \sqrt{n}\alpha$. □

Proof of Theorem 4.3. **(a)** Some of the techniques in our proof here are inspired by [14]. First, $\mathbf{V}' = \mathbf{V} \cup \mathbf{Z}'$ implies that $\mathbf{Z}' \perp\!\!\!\perp \mathbf{T} \mid \mathbf{V}$. Then $\pi^{\mathbf{T}}(\mathbf{V}) = \pi^{\mathbf{T}}(\mathbf{V}')$ holds, which further derives the following equation:

$$D_d^{\text{eff}}(\mathcal{V}) = D_d^{\text{eff}}(\mathcal{V}') + \underbrace{\frac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V})}(m_{\mathbf{V}'}^{\mathbf{T}}(\mathbf{Y})-m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}))}_{D_M}.$$

Then we obtain the fact that $\mathbb{E}[D_M D_d^{\text{eff}}(\mathbf{V}')]=0$:

$$\mathbb{E}[D_M D_d^{\text{eff}}(\mathbf{V}')] = \mathbb{E}_{\mathbf{V}'}\mathbb{E}[D_M D_d^{\text{eff}}(\mathbf{V}') \mid \mathbf{V}'] = 0,$$

where the first equality is due to the tower property, and the second equality is due to the fact that $\mathbb{E}[\mathbf{Y}(\mathbf{t}) - m_{\mathbf{V}'}^{\mathbf{T}}(\mathbf{Y}) \mid \mathbf{V}'] = 0$. Meanwhile, we derive the expectation of the term D_M as follows:

$$\begin{aligned}
&\mathbb{E}[D_M] \\
&= \mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V})}(m_{\mathbf{V}'}^{\mathbf{T}}(\mathbf{Y})-m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}))\right] \\
&= \underbrace{\mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=1)}{\pi^{\mathbf{T}}(\mathbf{V})}(m_{\mathbf{V}'}^{\mathbf{T}}(\mathbf{Y})-m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}))\right]}_{D_M^1} - \underbrace{\mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V})}(m_{\mathbf{V}'}^{\mathbf{T}}(\mathbf{Y})-m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}))\right]}_{D_M^2},
\end{aligned}$$

where the term D_M^1 is then simplified as follows:

$$\begin{aligned}
D_M^1 &= \mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=1)}{\pi^{\mathbf{T}=1}(\mathbf{V})}(m_{\mathbf{V}'}^{\mathbf{T}=1}(\mathbf{Y})-m_{\mathbf{V}}^{\mathbf{T}=1}(\mathbf{Y}))\right] \\
&= \mathbb{E}_{\mathbf{V}}\left[\mathbb{E}\left[(m_{\mathbf{V}'}^{\mathbf{T}=1}(\mathbf{Y})-m_{\mathbf{V}}^{\mathbf{T}=1}(\mathbf{Y}))\mid\mathbf{V}\right]\mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=1)}{\pi^{\mathbf{T}=1}(\mathbf{V}')}\mid\mathbf{V}\right]\right],
\end{aligned}$$

where the term $\mathbb{E}\left[(m_{\mathbf{V}'}^{\mathbf{T}=1}(\mathbf{Y})-m_{\mathbf{V}}^{\mathbf{T}=1}(\mathbf{Y}))\mid\mathbf{V}\right]=0$ due to the fact that $\mathbb{E}[m_{\mathbf{V}'}^{\mathbf{T}=1}(\mathbf{Y}) \mid \mathbf{V}] = m_{\mathbf{V}}^{\mathbf{T}=1}(\mathbf{Y})$. The simplification of D_M^2 is

similar to that of D_M^1. Thus we derive that $\mathbb{E}[D_M] = 0$. Finally, we derive the formulation of $\mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V}))$ as follows:

$$\begin{aligned}
&\mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V})) \\
&= \mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V}')) + \mathrm{Var}(D_M) \\
&= \mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V}')) + \mathbb{E}\left[D_M^2\right] \\
&= \mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V}')) + \mathbb{E}\left[\left(\frac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V})}(m_{\mathbf{V}'}^{\mathbf{T}}(\mathbf{Y}) - m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}))\right)^2\right] \\
&= \mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V}')) + \mathbb{E}_{\mathbf{V}}\left[\mathrm{Var}(m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}) \mid \mathbf{V})\left(\frac{1}{p(\mathbf{T}=1 \mid \mathbf{V})} + \frac{1}{p(\mathbf{T}=0 \mid \mathbf{V})}\right)\right],
\end{aligned}$$

where the last equality is due to the fact that $\mathbb{E}[m_{\mathbf{V}'}^{\mathbf{T}}(\mathbf{Y}) \mid \mathbf{V}] = m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y})$ with some algebra on the term $\mathbb{E}\left[\left(\frac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V})}\right)^2 \mid \mathbf{V}\right]$.

(b) First, $\mathbf{V}' = \mathbf{V} \cup \mathbf{I}'$ and $\mathbf{I}' \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{V}, \mathbf{T}$ imply that $m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y}) = m_{\mathbf{V}'}^{\mathbf{T}}(\mathbf{Y})$ holds. Then we derive the following decomposition of $\mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V}'))$:

$$\mathrm{Var}(\mathbb{E}\left[D_d^{\mathrm{eff}}(\mathbf{V}') \mid \mathbf{T}, \mathbf{V}, \mathbf{Y}\right]) + \mathbb{E}\left[\mathrm{Var}\left(D_d^{\mathrm{eff}}(\mathbf{V}') \mid \mathbf{T}, \mathbf{V}, \mathbf{Y}\right)\right], \quad (19)$$

where the term $\mathbb{E}\left[D_d^{\mathrm{eff}}(\mathbf{V}') \mid \mathbf{T}, \mathbf{V}, \mathbf{Y}\right]$ is simplified as follows:

$$\begin{aligned}
&\mathbb{E}\left[D_d^{\mathrm{eff}}(\mathbf{V}') \mid \mathbf{T}, \mathbf{V}, \mathbf{Y}\right] \\
&= \mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V}')}(\mathbf{Y} - m_{\mathbf{V}'}^{\mathbf{T}}(\mathbf{Y})) \mid \mathbf{T}, \mathbf{V}, \mathbf{Y}\right] \\
&= \mathbb{E}\left[\frac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V}')}(\mathbf{Y} - m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y})) \mid \mathbf{T}, \mathbf{V}, \mathbf{Y}\right] \\
&= \sum_{\mathbf{t}\in\{0,1\}} (2\mathbf{t}-1)\mathcal{I}(\mathbf{T}=\mathbf{t})(\mathbf{Y} - m_{\mathbf{V}}^{\mathbf{t}}(\mathbf{Y}))\mathbb{E}\left[\frac{1}{\pi^{\mathbf{T}}(\mathbf{V}')} \mid \mathbf{T}, \mathbf{V}\right] \\
&= D_d^{\mathrm{eff}}(\mathbf{V}).
\end{aligned}$$

Then, we apply the results in (19) and simplify the expression of $\mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V}'))$ as follows:

$$\begin{aligned}
&\mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V}')) \\
&= \mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V})) + \mathbb{E}\left[\mathrm{Var}\left(D_d^{\mathrm{eff}}(\mathbf{V}') \mid \mathbf{T}, \mathbf{V}, \mathbf{Y}\right)\right] \\
&= \mathrm{Var}(D_d^{\mathrm{eff}}(\mathcal{V})) + \mathbb{E}\left[\mathrm{Var}\left(\frac{\mathcal{I}(\mathbf{T}=1)-\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}}(\mathbf{V}')}(\mathbf{Y} - m_{\mathbf{V}'}^{\mathbf{T}}(\mathbf{Y})) \mid \mathbf{T}, \mathbf{V}, \mathbf{Y}\right)\right] \\
&= \mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V})) + \mathbb{E}\left[\sum_{\mathbf{t}\in\{0,1\}} \mathcal{I}(\mathbf{T}=\mathbf{t})\left(\mathbf{Y} - m_{\mathbf{V}'}^{\mathbf{T}}(\mathbf{Y})\right)^2 \mathrm{Var}\left(\frac{1}{\pi^{\mathbf{T}}(\mathbf{V}')} \mid \mathbf{T}, \mathbf{V}, \mathbf{Y}\right)\right].
\end{aligned}$$

[(c)] Once the variable set $\mathbf{Z}$ contains all the parents of $\mathbf{Y}$, we can write the structural equation of $\mathbf{Y}$ as $\mathbf{Y} = f_{\mathbf{Y}}(\mathbf{T}, \mathbf{X}, \mathbf{Z})$. Then the proposed OAF metric of $\mathbf{V}_0$, i.e., can be derived as follows:

$$\underbrace{\mathbb{E}\left[\left(\frac{\mathcal{I}(\mathbf{T}=1)}{\pi^{\mathbf{T}=1}(\mathbf{V}_0)}(\mathbf{Y} - m_{\mathbf{V}_0}^{\mathbf{T}=1}(\mathbf{Y}))\right)^2\right]}_{K_1} + \underbrace{\mathbb{E}\left[\left(\frac{\mathcal{I}(\mathbf{T}=0)}{\pi^{\mathbf{T}=0}(\mathbf{V}_0)}(\mathbf{Y} - m_{\mathbf{V}_0}^{\mathbf{T}=0}(\mathbf{Y}))\right)^2\right]}_{K_2},$$

where we further expand K_1 as follows:

$$\begin{aligned}
K_1 &= \mathbb{E}_{\mathbf{V}_0}\left[\mathbb{E}\left[\left(\frac{\mathcal{I}(\mathbf{T}=1)}{\pi^{\mathbf{T}=1}(\mathbf{V}_0)}\right)^2 \mid \mathbf{V}_0\right]\mathbb{E}\left[\left(\mathbf{Y}(\mathbf{t}=1) - m_{\mathbf{V}_0}^{\mathbf{T}}(\mathbf{Y})\right)^2 \mid \mathbf{V}_0\right]\right] \\
&= \mathbb{E}_{\mathbf{V}_0}\left[\mathbb{E}\left[\left(\frac{\mathcal{I}(\mathbf{T}=1)}{\pi^{\mathbf{T}=1}(\mathbf{V}_0)}\right)^2 \mid \mathbf{V}_0\right]\mathbb{E}\left[(\mathbf{Y}(\mathbf{t}=1) - f_{\mathbf{Y}}(\mathbf{T}, \mathbf{X}, \mathbf{Z}))^2 \mid \mathbf{V}_0\right]\right] \\
&= 0,
\end{aligned}$$

where the second equality holds due to the fact that $\mathbf{Z}$ contains all the parents of $\mathbf{Y}$. Similar to above derivation, we obtain that $K_2 = 0$. Furthermore, we conclude that $\mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V}_0)) = 0$, which indicates that $\mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V}_0)) \le \mathrm{Var}(D_d^{\mathrm{eff}}(\mathbf{V}'))$.

□

B DETAILS ON IMPLEMENTATION

Details on estimating $\widehat{\pi^{\mathbf{T}}(\mathbf{V})}$ and $\widehat{m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y})}$ For the linear implementation OAFP_L, we build $\widehat{\pi^{\mathbf{T}}(\mathbf{V})}$ and $\widehat{m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y})}$ using the linear regression and logistic regression without any regularization tricks. Meanwhile, our implementation on the downstream estimator, namely the AIPW estimator, follows the Zepid package [17]. For the nonlinear implementation OAFP_N, we build $\widehat{\pi^{\mathbf{T}}(\mathbf{V})}$ and $\widehat{m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y})}$ with two deep networks. The regression network for $\widehat{\pi^{\mathbf{T}}(\mathbf{V})}$ consists of four MLP layers with the activation function as *ELU*, and the score network consists of three MLP layers with *ELU* as the activation function for the first two layers and *Sigmoid* for the last layer. The optimizer we choose for $\widehat{\pi^{\mathbf{T}}(\mathbf{V})}$ and $\widehat{m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y})}$ is the Adam optimizer [], where the learning rate is 0.001 and 0.0005, respectively. Notably, we split an extra validation set from the training data such that $\widehat{\pi^{\mathbf{T}}(\mathbf{V})}$ and $\widehat{m_{\mathbf{V}}^{\mathbf{T}}(\mathbf{Y})}$ are evaluated on the validation part. Besides, we implement OAFP_L on a single Tesla V100 gpu. For OAFP_N, we compute $\widehat{\mathcal{R}}$ on a 8-gpu Tesla V100 cluster, where each batch array B_i is trained in a single process in parallel.

Details on our OAFP framework Our implementation follows prior work on neural combinatorial search [4, 37]. The encoder is a Transformer and the decoder is a multi-layer MLP. The Transformer takes the covariates U as input (total size: $\mathcal{R}^{K\times d\times n_b}$). The encoder's output has the same shape. We combine $\mathbf{T}$ and $\mathbf{Y}$ with the learned representation of $\mathbf{U}$ and feed them to the MLP decoder. The decoder uses sigmoid functions to sample the binary feature mask. We use $n_b = 512$ and $K = 64$ throughout our experiments. The Transformer encoder has two blocks and the MLP decoder has two linear layers with ReLU activation. We apply the reinforce approach [30] to reduce actor variance, using an exponential average of past rewards with a scaling factor of 0.99. The entropy term's hyper-parameter is set to 1.

Details on the downstream estimators AIPW and DragonNet We implemented two downstream estimators, AIPW and DragonNet, using the zepid package for AIPW and the original open-source implementation for DragonNet. AIPW employs linear regression for outcome regression and logistic regression for propensity score estimation. These estimators are combined using semi-parametric approaches [12]. DragonNet, as a deep version of AIPW, integrates score prediction and outcome prediction into an end-to-end network with target regularization to satisfy the estimation equation [22]. DragonNet's architecture consists of 4-layer MLD layers activated by the *ELU* function. It has three heads: a single-layer MLP with *sigmoid* as the score head and two three-layer MLPs with *ELU* as the outcome heads. The Adam optimizer with an initial learning rate of 0.001 is used for DragonNet.