EasyChair Preprint

№ 13093

Trustworthy Policy Learning Under the Counterfactual No-Harm Criterion

Haoxuan Li, Chunyuan Zheng, Yixiao Cao, Zhi Geng, Yue Liu and Peng Wu

April 25, 2024

Trustworthy Policy Learning under the Counterfactual No-Harm Criterion

Haoxuan Li [1] **Chunyuan Zheng** [2] **Yixiao Cao** [1] **Zhi Geng** [3] **Yue Liu** [4] **Peng Wu** [3]

Abstract

Trustworthy policy learning has significant importance in making reliable and harmless treatment decisions for individuals. Previous policy learning approaches aim at the well-being of subgroups by maximizing the utility function (e.g., conditional average causal effects, post-view click-through&conversion rate in recommendations), however, individual-level counterfactual no-harm criterion has rarely been discussed. In this paper, we first formalize the counterfactual no-harm criterion for policy learning from a principal stratification perspective. Next, we propose a novel upper bound for the fraction negatively affected by the policy and show the consistency and asymptotic normality of the estimator. Based on the estimators for the policy utility and harm upper bounds, we further propose a policy learning approach that satisfies the counterfactual no-harm criterion, and prove its consistency to the optimal policy reward for parametric and non-parametric policy classes, respectively. Extensive experiments are conducted to show the effectiveness of the proposed policy learning approach for satisfying the counterfactual no-harm criterion.

1. Introduction

Policy learning determines the individuals who should be treated based on their covariates (Murphy, 2003), and it is important that a decision made by an algorithm can be trusted by humans (Floridi, 2019; Kaur et al., 2022). Specifically, trustworthy policy learning requires that the learned policy has beneficence, non-maleficence, autonomy, justice, and explicability (Thiebes et al., 2021; Floridi, 2019; Kaur et al., 2022), and many counterfactual-based metrics are proposed to quantify the policy's trustworthiness (Kusner et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Wu et al., 2019), which makes the algorithm try to understand, for the individuals, what the outcome would be if an alternative intervention had been implemented (Pearl, 2009).

Nevertheless, the counterfactual harmlessness of policy learning is rarely discussed, which would prevent an active intervention on the individuals from having worse outcomes than the natural state without the intervention (Richens et al., 2022). This also serves as the basic principle of the Hippocratic oath (Sokol, 2013) that *"First do no harm"*, and similar principles can be found from Lin (2006); Mill (1966); Asimov (2004). Towards this end, previous studies employ group causal effects to define the utility to learn individualized treatment policies (Bertsimas et al., 2016; Kitagawa & Tetenov, 2018; Athey & Wager, 2021), however, they can only maximize the average benefit of *subgroups*, without satisfying the counterfactual no-harm for *individuals.*

In this paper, we formally discuss the cause of counterfactual harm from a principal stratification perspective (Frangakis & Rubin, 2002), by dividing the units into groups by the joint value of the potential outcomes. We then formalize the utility functions of the conditional average treatment effect (CATE)-based (Chipman et al., 2010; Johansson et al., 2016; Shalit et al., 2017; Wager & Athey, 2018; Künzel et al., 2019; Shi et al., 2019) and the recommendation-based (Ma et al., 2018; Zhang et al., 2020; Wang et al., 2022) policy learning and discuss the explicit solutions of the optimal policy. Unfortunately, neither of them is able to satisfy the individual counterfactual no-harm, which is summarized as pursuing only the maximal causal effect gain of the subpopulation is not sufficient to achieve reliable and no-harm decision making for individuals.

The basic challenge for satisfying the counterfactual no-harm criterion from subgroups to individuals is that, since each unit can be only assigned with one treatment, we always observe the corresponding potential outcome, but not both, which is also known as the fundamental problem of causal inference (Holland, 1986; Morgan & Winship, 2015). We follow Kallus (2022b) to consider the fraction negatively affected (FNA), and further propose a metric to quantify the fraction harmed by the policy. Specially, we extend Li

[1]Center for Data Science, Peking University [2]Department of Mathematics, University of California, San Diego [3]School of Mathematics and Statistics, Beijing Technology and Business University [4]Center for Applied Statistics and School of Statistics, Renmin University of China. Correspondence to: Yue Liu <liuyue_stats@ruc.edu.cn>, Peng Wu <pengwu@btbu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

& Pearl (2019) and Kallus (2022b) to give upper bounds of the counterfactual harm, which are strictly tighter under mild assumptions. Notably, the proposed estimators of upper bounds are consistent and asymptotically normal under weaker assumptions compared to Kallus (2022b), and are convenient for policy learning, especially in optimization.

Next, we turn to the question that how to bridge the CATE and the cost function with the counterfactual no harm criterion? From a policy learning perspective, we demonstrate that larger CATE or cost would contribute to counterfactual harmless, which also has a guiding significance in practice.

To learn the optimal policies satisfying the counterfactual no-harm criterion, we propose estimators for the policy utility and the upper bound of policy harm, respectively, and further propose a policy learning approach. Moreover, we prove the consistency results, when the policies are parametric (also known as policy gradient) and nonparametric, respectively. To the best of our knowledge, this is the first paper to propose policy learning approaches that satisfy the counterfactual no-harm criterion and to prove its consistency to the optimal counterfactual harmless policy reward.

The contributions of this paper are summarized as follows.

- We formally discuss the counterfactual no-harm criterion for policy learning from a principal stratification perspective and show that common CATE-based and recommendation-based policy learning do not satisfy the criterion.
- We propose a metric to quantify the fraction harmed by the policy, and a novel estimator for its upper bound, and prove its consistency and asymptotic normality.
- Based on the estimators for the upper bounds and policy reward, we further propose policy learning approaches that satisfy the counterfactual no-harm criterion and prove its consistency to the optimal policy reward for parametric and non-parametric policy classes, respectively.
- Extensive experiments are conducted to show the effectiveness of the proposed policy learning approaches for satisfying the counterfactual no-harm criterion.

2. Related Work

Trustworthy Policy Evaluation and Learning. Policy learning aims to determine the individuals who should be treated that maximizes the utility function based on their covariates (Murphy, 2003). Previous studies employ group causal effects to define the utility to learn individualized treatment policies, using regression based (Bertsimas et al., 2016), reweighted based (Kitagawa & Tetenov, 2018), and doubly robust methods (Athey & Wager, 2021).

In addition to utility maximization, trustworthy policy learning requires that the learned policy has beneficence, non-maleficence, autonomy, justice, and explicability (Thiebes et al., 2021; Floridi, 2019; Kaur et al., 2022), and many counterfactual-based metrics are proposed to quantify the policy's trustworthiness (Kusner et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Ben-Michael et al., 2022). In this paper, we focus on policy learning under the counterfactual no-harm criterion, which has rarely been discussed.

Heterogeneous Treatment Effects and No-Harm Criterion. Heterogeneous treatment effects, also known as the conditional average treatment effects (CATEs), describe the average treatment effects on subgroups with specific covariates, which plays a crucial role in such domains as precision medicine (Jaskowski & Jaroszewicz, 2012) and decision making (Guelman et al., 2015). Many approaches have been proposed for the estimation of CATE, such as Bayesian Additive Regression Trees (BART) (Chipman et al., 2010), Balancing Neural Network (BNN) (Johansson et al., 2016), CounterFactual Regression (CFR) (Shalit et al., 2017), Perfect Match (PM) (Schwab et al., 2018), Causal Forest (CF) (Wager & Athey, 2018), X-learner (Künzel et al., 2019), and DragonNet (Shi et al., 2019).

However, the observation-based utilities and CATE do not necessarily satisfy the no-harm criterion, especially under the *individual* sense. This is intuitively due to that CATE-based policy learning only seeks to maximize the average effect under (sub)groups (see Section 4 for the formal discussions). Towards this end, Richens et al. (2022) propose a formal definition of harm and benefit using causal models. Li & Pearl (2019) and Ben-Michael et al. (2022) consider the utilities depend on unobserved outcomes in binary outcomes case. Kallus (2022b) propose the sharp bounds on the fractions that are negatively affected, and Kallus (2022a) study the conditional value at risk (CVaR) for the continuous outcomes. In this paper, we extend Li & Pearl (2019) and Kallus (2022b) to give an upper bound of the counterfactual harm by the policy, the proposed upper bound is strictly tighter under mild assumptions, as well as has many desirable properties. We also propose estimation methods for policy learning satisfying the counterfactual no-harm criterion, and show the consistency and asymptotic normality.

3. Preliminaries

3.1. Notation and Setup

In this paper, we consider the case of binary treatment. Suppose a simple random sampling of n units from a super population $\mathbb{P}$, for each unit i, the covariate and the assigned treatment are denoted as $X_i \in \mathcal{X} \subset \mathbb{R}^m$ and $T_i \in \mathcal{T} = \{0, 1\}$, respectively, where $T_i = 1$ means receiving treatment, while $T_i = 0$ means not receiving treatment and maintaining a natural state. Let $Y_i \in \mathcal{Y} = \{0, 1\}$ be the corresponding binary outcome. Without loss of generality, we assume that the larger outcome is preferable. To study the counterfactual

Table 1. The units are divided into four subgroups from a principal stratification perspective, according to $(Y(0), Y(1))$, named "useless treatment group", "useful treatment group", "harmful treatment group", and "harmless treatment group", respectively.

NOTATION	GROUP	$Y(0)$	$Y(1)$
$Y_{0,0}$	USELESS TREATMENT	0	0
$Y_{0,1}$	USEFUL TREATMENT	0	1
$Y_{1,0}$	HARMFUL TREATMENT	1	0
$Y_{1,1}$	HARMLESS TREATMENT	1	1

no-harm criterion for individuals, we adopt the potential outcome framework (Rubin, 1974; Neyman, 1990) in causal inference. Specifically, let $Y_i(0)$ and $Y_i(1)$ be the outcome of unit i had this unit receive treatment $T_i = 0$ and $T_i = 1$, respectively. Since each unit can be only assigned with one treatment, we always observe the corresponding outcome be either $Y_i(0)$ or $Y_i(1)$, but not both, which is also known as the fundamental problem of causal inference (Holland, 1986; Morgan & Winship, 2015).

We assume that the observation for unit i is $Y_i = (1 - T_i)Y_i(0) + T_iY_i(1)$. In other words, the observed outcome is the potential outcome corresponding to the assigned treatment, which also known as the consistency assumption in the causal literature. We assume that the stable unit treatment value assumption (STUVA) assumption holds, i.e., there should not be alternative forms of the treatment and interference between units. Furthermore, we follow Li & Pearl (2022) and Kallus (2022b) to assume that the strong ignorability assumption holds, i.e., $(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i|X_i$ and let $\eta < \mathbb{P}(T_i = 1|X_i = x) < 1 - \eta$, where η is a constant between 0 and $1/2$.

To evaluate treatment assignments or learned policies, causal effects are widely adopted. For unit i, the individual treatment effect (ITE) is defined as $\text{ITE}_i = Y_i(1) - Y_i(0)$, where $\text{ITE}_i > 0$ indicates that the treatment $T_i = 1$ is beneficial for individual i, and vice versa. The conditional average treatment effect (CATE) is defined as

$$\tau(x) = \mathbb{E}[\text{ITE}_i|X_i = x] = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x],$$

that is, the difference in the conditional mean outcomes between treatments given covariate. For simplification, we drop the subscript i for a generic unit hereafter.

3.2. Principal Stratification Method

In contrast to dividing units into groups by the observed characteristics, principal stratification method (Frangakis & Rubin, 2002) divides units into groups by the *joint value of the potential outcomes* from a counterfactual perspective. It provides more informative description of the individual risk, and has been widely adopted in survival analysis (Zhang & Rubin, 2003; Imai, 2008; Ding et al., 2011) and mediation analysis (Frangakis & Rubin, 1999; Gallop et al., 2009; Jiang et al., 2016).

Specifically, we follow Ben-Michael et al. (2022) to define the groups of $(Y(0) = 0, Y(1) = 0)$, $(Y(0) = 0, Y(1) = 1)$, $(Y(0) = 1, Y(1) = 0)$, $(Y(0) = 1, Y(1) = 1)$ as the *useless treatment group*, *useful treatment group*, *harmful treatment group*, and *harmless treatment group*, respectively. For simplification, we denote the labels of the four groups as $Y_{0,0}, Y_{0,1}, Y_{1,0}$, and $Y_{1,1}$ correspondingly, as shown in Table 1. Let $\mathbb{P}(Y_{0,0}|X = x)$, $\mathbb{P}(Y_{0,1}|X = x)$, $\mathbb{P}(Y_{1,0}|X = x)$ and $\mathbb{P}(Y_{1,1}|X = x)$ be the probability that units with covariate $X = x$ belong to each group. Then $\tau(x)$ is

$$\begin{aligned}\tau(x) &= \mathbb{E}(Y(1) - Y(0)|X = x) \\ &= (0 - 0)\mathbb{P}(Y_{0,0}|X = x) + (1 - 0)\mathbb{P}(Y_{0,1}|X = x) \\ &+ (0 - 1)\mathbb{P}(Y_{1,0}|X = x) + (1 - 1)\mathbb{P}(Y_{1,1}|X = x) \\ &= \mathbb{P}(Y_{0,1}|X = x) - \mathbb{P}(Y_{1,0}|X = x),\end{aligned}$$

that is, the **difference** between the probabilities of belonging to the *useful group* $Y_{0,1}$ and *harmful group* $Y_{1,0}$ in the subpopulation of $X = x$. Whereas the principal stratification interests in the **values** of $\mathbb{P}(Y_{0,1}|X = x)$ and $\mathbb{P}(Y_{1,0}|X = x)$.

Remarkably, compared to CATE, the principal stratification provides a more fine-grained and informative description of the individuals. However, even with the strong ignorability assumption, we are still unable to obtain unbiased estimates of all the $\mathbb{P}(Y_{0,0}|X = x)$, $\mathbb{P}(Y_{0,1}|X = x)$, $\mathbb{P}(Y_{1,0}|X = x)$ and $\mathbb{P}(Y_{1,1}|X = x)$, which poses a serious challenge to assess the individual risk of a learned policy.

4. Counterfactual No-Harm Criterion and the Relation to Trustworthy Policy Learning

4.1. Counterfactual No-Harm Criterion

Trustworthy policy learning requires that the learned policy pursue both beneficence and non-maleficence (Thiebes et al., 2021). However, many previous studies have been devoted to maximizing group utility, while have ignored the counterfactual no-harm requirement on the individual level.

For instance, for seriously ill patients, one can give either an (active) therapeutic intervention $T = 1$ or maintain a (conservative) natural state $T = 0$. However, in any case, the treatment assigned to an individual should not be harmful, i.e., no active treatment $T = 1$ should be given to individuals with $(Y(0) = 1, Y(1) = 0)$, since these patients could have had a more favorable outcome under the natural state $T = 0$. This also serves as the basic principle of the Hippocratic oath (Sokol, 2013) that *"First do no harm"*, and similar principles can be found from the environmental policy (Lin, 2006), the foundations of classical liberalism (Mill, 1966), and Asimov's laws of robotics (Asimov, 2004).

Given that policy learning based on (conditional) average causal effects seeks to maximize the average utility of the (sub)population rather than the individual, we argue that these approaches may be overly aggressive and thus harm a large number of individuals. For example, consider a policy that can be useful for 50% of patients but will harm 30% of patients, while an alternative policy can only be useful for 15% of patients but no harm. It is clear that the latter is more applicable when considering counterfactual no-harm requirements, whereas those policies only considering to maximize CATEs would prefer the former.

4.2. Previous CATE-Based Policy Learning Does Not Meet the Counterfactual No-Harm Criterion

Let $\pi : \mathcal{X} \to [0, 1]$ be a policy that maps from the individual context $X = x$ to the probability of the treatment $T = 1$ to be assigned. For a general policy learning under the counterfactual criterion, let $U = U(X, T, Y(T))$[1] be the utility function and the policy reward $R(\pi)$ is

$$R(\pi) = \mathbb{E}[\pi(X)U(X, 1, Y(1)) + (1 - \pi(X))U(X, 0, Y(0))].$$

The policy learning aims to learn an optimal policy π^* that maximizes the policy reward $\pi^* = \arg\max_{\pi\in\Pi} R(\pi)$.

For the observed outcome-based decision making rule, the utility function is defined as $U(X, T, Y) = Y$. More generally, given the bounded cost function $c(X)$ of imposing active treatment $T = 1$ compared to no treatment $T = 0$, the utility function is $U(X, T, Y; c) = Y - Tc(X)$. Let the policy reward be $R(\pi; c)$, and the optimal policy be $\pi^*(x; c) = \arg\max_{\pi\in\Pi} R(\pi; c)$.

By substituting the utility in the policy reward, we have

$$\begin{aligned} R(\pi; c) &= \mathbb{E}[(Y(1) - c(X))\pi(X) + Y(0)(1 - \pi(X))] \\ &= \mathbb{E}[(Y(1) - Y(0) - c(X))\pi(X) + Y(0)], \end{aligned}$$

and the optimal policy is

$$\pi^*(x; c) = \begin{cases} 1, & \mathbb{E}[Y(1) - Y(0) \mid X = x] = \tau(x) > c(x) \\ 0, & \mathbb{E}[Y(1) - Y(0) \mid X = x] = \tau(x) < c(x), \\ d, & \mathbb{E}[Y(1) - Y(0) \mid X = x] = \tau(x) = c(x) \end{cases}$$

where d is any value between 0 and 1, and π^* would always impose a treatment intervention $T = 1$ for individuals whose $\tau(x)$ is greater than the cost $c(x)$ and vice versa, which is same as CATE-based policy learning. From a principle stratification prospective, that is equivalent to

$$\pi^*(x; c) = \begin{cases} 1, & \mathbb{P}(Y_{0,1}|X = x) - \mathbb{P}(Y_{1,0}|X = x) > c(x) \\ 0, & \mathbb{P}(Y_{0,1}|X = x) - \mathbb{P}(Y_{1,0}|X = x) < c(x). \\ d, & \mathbb{P}(Y_{0,1}|X = x) - \mathbb{P}(Y_{1,0}|X = x) = c(x) \end{cases}$$

[1]Under the consistency assumption in Section 3.1, we write $U = U(X, T, Y(T)) = U(X, T, Y)$ thereafter for simplification.

Therefore, one can conclude that the optimal policies do not satisfy the counterfactual no-harm criterion. The reason is that such policies only focus on the difference between $\mathbb{P}(Y_{0,1}|X = x)$ and $\mathbb{P}(Y_{1,0}|X = x)$, and fail to control $\mathbb{P}(Y_{1,0}|X = x)$ itself and may assign harmful treatments. In particular, when both $\mathbb{P}(Y_{0,1}|X = x)$ and $\mathbb{P}(Y_{1,0}|X = x)$ are large, the optimal policy might still prefer to assign the active treatment $T = 1$, which results in a harmful decision making for the individuals.

4.3. Previous Recommendation Policy Learning Does Not Meet the Counterfactual No-Harm Criterion

In contrast to CATE-based policy learning, an alternative branch is personalized recommendation, which plays an crucial role in practice. For advertising agencies, they gain profit only when the ad is being recommended to the user $T = 1$ and converts $Y = 1$ (Ma et al., 2018; Zhang et al., 2020; Wang et al., 2022). Formally, the utility function is $U(X, T, Y) = TY - Tc(X)$, where $c(X)$ is the cost of placing an advertisement $T = 1$. Then we have

$$R(\pi; c) = \mathbb{E}[(Y(1) - c(X))\pi(X)],$$

and the optimal policy $\pi^*(x; c) = \arg\max_{\pi\in\Pi} R(\pi; c)$ is

$$\pi^*(x; c) = \begin{cases} 1, & \mathbb{P}(Y_{0,1}|X = x) + \mathbb{P}(Y_{1,1}|X = x) > c(x) \\ 0, & \mathbb{P}(Y_{0,1}|X = x) + \mathbb{P}(Y_{1,1}|X = x) < c(x). \\ d, & \mathbb{P}(Y_{0,1}|X = x) + \mathbb{P}(Y_{1,1}|X = x) = c(x) \end{cases}$$

One can see that this would lead to a more serious violation of the counterfactual no-harm criterion compared to the polices learned in Section 4.2, which is also empirically verified in Section 7. In fact, the optimal policies only care about the sum of $\mathbb{P}(Y_{0,1}|X = x)$ and $\mathbb{P}(Y_{1,1}|X = x)$, i.e., the users for whom conversion $Y(1) = 1$ would occur under the active recommendation $T = 1$. Such policies never take into account the harmful treatment population $\mathbb{P}(Y_{1,0}|X = x)$, which would lead to a more aggressive recommendation policy and cause potential user churn.

5. Proposed Sharp Bounds of the Counterfactual No-Harm Criterion

In the previous section, we found that both CATE-based and recommendation-based policy learning fail to satisfy the counterfactual no-harm criterion, since they do not care how many individuals will be negatively affected by the learned policy. However, we cannot explicitly identify the individuals who are negatively affected by the treatment intervention, because of the fundamental problem of causal inference – that we never observe the two potential outcomes $(Y(0), Y(1))$ at the same time. We follow Kallus (2022b) to consider the fraction negatively affected (FNA), i.e., FNA =

$\mathbb{P}(Y(0)=1,Y(1)=0)=\mathbb{P}(Y_{1,0})$, and let $\text{FNA}(x)$ be the FNA with given covariates $X=x$ that

$$\text{FNA}(x)=\mathbb{P}(Y(0)=1,Y(1)=0|X=x).$$

Given a policy $\pi\in\Pi$, we further propose $\text{FNA}(\pi)$ as the fraction harmed by the policy π that

$$\text{FNA}(\pi)=\mathbb{E}(\mathbb{P}(Y(0)=1,Y(1)=0|X)\pi(X)).$$

In Proposition 5.1, we discuss a general upper bound for $\text{FNA}(x)$ and $\text{FNA}(\pi)$, respectively.

Proposition 5.1 (Tight upper bounds)**.** *(a) The tight upper bound of FNA(x), named $w_{\text{FNA}}(x)$, is*

$$\min\{\mathbb{P}(Y=1|T=0,X=x),\mathbb{P}(Y=0|T=1,X=x)\};$$

(b) Given a policy $\pi\in\Pi$, the tight upper bound of the FNA(π) is $\mathbb{E}[w_{\text{FNA}}(X)\pi(X)]$.

The upper bounds $w_{\text{FNA}}(x)$ and $\mathbb{E}[w_{\text{FNA}}(X)\pi(X)]$ in Proposition 5.1 are tight, that is, the best we could infer given infinite data, and they are reached when $\mathbb{P}(Y=1|T=0,X=x)=\mathbb{P}(Y=0|T=1,X=x)=1$. Besides, it does not require any additional assumptions, which can be regarded as a special case in Li & Pearl (2022) and Kallus (2022b). However, this bound is wide and inconvenient for our policy learning (see the discussions after Theorem 5.2). By further assuming that $Y(0)$ and $Y(1)$ are non-negatively correlated given $X=x$, we give narrower bounds in Theorem 5.2, and discuss the convenience as well as the theoretical results in the following. Note that the assumption is empirically reasonable as well as easily satisfied. For example, in medical scenarios where $T=1$ indicates receiving active treatment, a patient's health status affects both $Y(0)$ and $Y(1)$ (Efron & Feldman, 1991); for a teacher-incentive program where $T=1$ indicates receiving financial incentives, a teacher's knowledge level and intend to teach affects both $Y(0)$ and $Y(1)$ (Duflo et al., 2012).

Theorem 5.2 (Main result 1)**.** *(a) If $Y(0)$ and $Y(1)$ are non-negatively correlated given $X=x$, the tight upper bound of the FNA(x), named $u_{\text{FNA}}(x)$, is*

$$\mathbb{P}(Y=1|T=0,X=x)\mathbb{P}(Y=0|T=1,X=x);$$

(b) Given a policy $\pi\in\Pi$, the tight upper bound of the FNA(π) is $\mathbb{E}[u_{\text{FNA}}(X)\pi(X)]$.

Notably, the conclusion in Theorem 5.2 gives the tightest-possible upper bounds (see Remark 5.3) and are narrower than the upper bounds in Proposition 5.1 (see Remark 5.4).

Remark 5.3 (Tightest-Possible (i.e., Sharp) Bounds). The upper bounds $u_{\text{FNA}}(x)$ are tight, and are reached when $Y(0)$ and $Y(1)$ are conditional independent for $x\in\mathcal{X}$.

Remark 5.4 (Tighter Bounds). The upper bounds $u_{\text{FNA}}(x)$ are tighter than that of $w_{\text{FNA}}(x)$ for $x\in\mathcal{X}$, and $\mathbb{E}[u_{\text{FNA}}(X)\pi(X)]\leq\mathbb{E}[w_{\text{FNA}}(X)\pi(X)]$ for $\pi\in\Pi$.

Moreover, the upper bounds in Theorem 5.2 require only mild assumptions to guarantee the asymptotic normality of the estimates, while the upper bounds in Proposition 5.1 require stronger assumptions, namely the sharpness margin condition in Kallus (2022b). We further claim that the upper bounds in Theorem 5.2 are convenient for policy learning, especially for optimization, with better smoothness and differentiability, compared to the upper bounds in Proposition 5.1 where minimization operators exist.

In the end of this section, we formally discuss the relation between the CATEs $\tau(x)$ and the upper bounds of $\text{FNA}(x)$ in Theorem 5.5. Given that CATEs are the finest magnitudes that can be identified via a data-driven way, Theorem 5.5 has important implications for guiding the policy learning that satisfies the counterfactual no-harm criterion.

Theorem 5.5 (Relation between CATEs and upper bounds)**.** *For the upper bounds $w_{\text{FNA}}(x)$ in Proposition 5.1 and $u_{\text{FNA}}(x)$ in Theorem 5.2, for all $x\in\mathcal{X}$, we have*

$$w_{\text{FNA}}(x)\leq\frac{1-\tau(x)}{2},\quad \textit{and}\quad u_{\text{FNA}}(x)\leq\frac{(1-\tau(x))^2}{4}.$$

Theorem 5.5 states that, for units whose CATE $\tau(x)$ tends to be 1, the probability that they are negatively affected by the treatment $T=1$ tends to be 0, i.e., the treatment is no-harm and safe. In fact, in real medical scenarios, physicians treat patients if they are confident that $\tau(x)$ is sufficiently large, and both $w_{\text{FNA}}(x)$ and $u_{\text{FNA}}(x)$ are small from Theorem 5.5. An alternative observation is that physicians treat patients who would die if untreated, i.e., $\mathbb{P}(Y=1|T=0,X=x)$ is small, which would also lead to small $w_{\text{FNA}}(x)$ and $u_{\text{FNA}}(x)$ from the formulas.

In Corollary 5.6, we further discuss the relation between the cost function $c(x)$ and the counterfactual harm upper bounds of the optimal policies π^* in Section 4.2.

Corollary 5.6 (Relation to the cost)**.** *For the upper bound $w_{\text{FNA}}(\pi)$ in Proposition 5.1 and $u_{\text{FNA}}(\pi)$ in Theorem 5.2, the optimal policies π^* in Section 4.2 satisfy*

$$w_{\text{FNA}}(\pi^*)\leq\mathbb{E}\Big[\frac{1-c(X)}{2}\pi^*(X)\Big],\ \textit{and}$$
$$u_{\text{FNA}}(\pi^*)\leq\mathbb{E}\Big[\frac{(1-c(X))^2}{4}\pi^*(X)\Big].$$

Corollary 5.6 shows that increasing the cost function $c(x)$ reduces the counterfactual harm of the optimal policies π^* in Section 4.2. This is because the optimal policies π^* tend to be more conservative as $c(x)$ increases, and thus fewer units are being actively treated with $T=1$. Notably, given

CATE $\tau(x)$ and the cost function $c(x)$, the $u_{\text{FNA}}(x)$ and $u_{\text{FNA}}(\pi^*)$ always lead to tighter counterfactual harm upper bounds than $w_{\text{FNA}}(x)$ and $w_{\text{FNA}}(\pi^*)$ in the right hand side (RHS) of Theorem 5.5 and Corollary 5.6.

6. Trustworthy No-Harm Policy Learning

Denote π^* as the optimal target policy satisfying the counterfactual no-harm criterion

$$\begin{array}{ll} \max_{\pi\in\Pi} & R(\pi; c, \rho) \\ \text{subject to} & u_{\text{FNA}}(\pi) \leq \lambda, \end{array} \tag{1}$$

where λ is a pre-specified level of allowed harm, and

$$R(\pi; c, \rho) = \mathbb{E}[\pi(X)\{Y(1) - c(X)\} + \rho Y(0)\{1 - \pi(X)\}]$$

for $\rho \in [0, 1]$, which is a general form of policy reward for different utility functions given in Sections 4.2 and 4.3. For example, $R(\pi; c, 1) = R(\pi)$ for $U(X, T, Y) = Y - Tc(X)$, and $R(\pi; c, 0) = R(\pi)$ for $U(X, T, Y) = TY - Tc(X)$.

Let $\hat{\pi}^*$ be the learned policy of π^*, derived by optimizing the empirical form of Eq. (1),

$$\begin{array}{ll} \max_{\pi\in\Pi} & \hat{R}(\pi; c, \rho) \\ \text{subject to} & \hat{u}_{\text{FNA}}(\pi) \leq \lambda, \end{array} \tag{2}$$

where $\hat{R}(\pi; c, \rho)$ and $\hat{u}_{\text{FNA}}(\pi)$ are the corresponding estimators of $R(\pi; c, \rho)$ and $u_{\text{FNA}}(\pi)$, obtained as follows.

Let $e(x) := \mathbb{P}(T = 1|X = x)$, $\mu_t(x) := \mathbb{E}[Y|T = t, X = x]$ for $t = 0, 1$, and

$$\varphi_\pi(Z; e, \mu_0, \mu_1) = \left(\frac{T(Y - \mu_1(X))}{e(X)} + \mu_1(X) - c(X)\right)\pi(X) + \rho\left(\frac{(1 - T)(Y - \mu_0(X))}{1 - e(X)} + \mu_0(X)\right)(1 - \pi(X)),$$

$$\psi_\pi(Z; e, \mu_0, \mu_1) = \left(\frac{(1 - T)(Y - \mu_0(X))}{1 - e(X)} + \mu_0(X)\right)\pi(X) - \left(\frac{T(Y - \mu_1(X))}{e(X)} + \mu_1(X)\right)\mu_0(X)\pi(X),$$

where $Z = (T, X, Y)$, then $R(\pi; c, \rho)$ and $u_{\text{FNA}}(\pi)$ can be unbiasedly estimated by φ_π and ψ_π from Lemma 6.1.

Lemma 6.1. *$\forall\pi \in \Pi$, $R(\pi; c, \rho) = \mathbb{E}[\varphi_\pi(Z; e, \mu_0, \mu_1)]$ and $u_{\text{FNA}}(\pi) = \mathbb{E}[\psi_\pi(Z; e, \mu_0, \mu_1)]$.*

Denote $\hat{e}(x)$ and $\hat{\mu}_t(x)$ for $t = 0, 1$ as the estimators of $e(x)$ and $\mu_t(x)$, respectively, using the sample-splitting (Wager & Athey, 2018; Chernozhukov et al., 2018) technique (See appendix for details). From Lemma 6.1, it is natural to define the estimators of $R(\pi; c, \rho)$ and $u_{\text{FNA}}(\pi)$ as

$$\hat{R}(\pi; c, \rho) = \frac{1}{n}\sum_{i=1}^{n}\varphi_\pi(Z_i; \hat{e}, \hat{\mu}_0, \hat{\mu}_1),$$

$$\hat{u}_{\text{FNA}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\psi_\pi(Z_i; \hat{e}, \hat{\mu}_0, \hat{\mu}_1),$$

which are augmented inverse probability weighting (AIPW)-like estimators (Robins et al., 1994; 1995).

Theorem 6.2. *Suppose that $||\hat{e}(x) - e(x)||_2 \cdot ||\hat{\mu}_t(x) - \mu_t(x)||_2 = o_{\mathbb{P}}(n^{-1/2})$ for all $x \in \mathcal{X}$ and $t \in \{0, 1\}$,*

(a) $\hat{R}(\pi, c; \rho)$ is consistent and asymptotically normal

$$\sqrt{n}\{\hat{R}(\pi, c; \rho) - R(\pi, c; \rho)\} \to N(0, \sigma_1^2),$$

where $\sigma_1^2 = \mathbb{V}[\varphi_\pi(Z; e, \mu_0, \mu_1)]$;

(b) if $\mu_0(x) = \mu_0(x; \phi)$ is a parametric model, $\hat{u}_{\text{FNA}}(\pi)$ is consistent and asymptotically normal

$$\sqrt{n}\{\hat{u}_{\text{FNA}}(\pi) - u_{\text{FNA}}(\pi)\} \to N(0, \sigma_2^2),$$

where

$$\sigma_2^2 = \mathbb{V}\Big[\psi_\pi(Z; e, \mu_0, \mu_1) - s(X)\mathbb{E}\Big\{\frac{\partial\mu_0(X; \phi)}{\partial\phi}\mu_1(X)\pi(X)\Big\}\Big],$$

and $s(X)$ is the influence function of estimator of ϕ.

Theorem 6.2 shows the consistency and asymptotically normality of $\hat{R}(\pi; c, \rho)$ and $\hat{u}_{\text{FNA}}(\pi)$ under mild assumptions. Based on it, we can derive the convergence rates of $R(\pi^*; c, \rho) - R(\hat{\pi}^*; c, \rho)$ and $R(\pi^*; c, \rho) - \hat{R}(\hat{\pi}^*; c, \rho)$, which are the regret of the learned policy, and error of the estimated reward of learned policy, respectively.

Theorem 6.3 (Main result 2)**.** *Suppose that for all $\pi \in \Pi$, $\pi(x) = \pi(x; \theta)$ is a continuously differentiable and convex function with respect to θ, where $\theta \in \Theta$ is a compact set, under the assumptions in Theorem 6.2, then we have*

(a) The expected reward of the learned policy is consistent, and $R(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho) = O_{\mathbb{P}}(1/\sqrt{n})$;

(b) The estimated reward of the learned policy is consistent, and $\hat{R}(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho) = O_{\mathbb{P}}(1/\sqrt{n})$.

Theorem 6.3(a) shows that the regret of the learned policy has a convergence rate of order $1/\sqrt{n}$, and Theorem 6.3(b) shows that the estimated reward of learned policy $\hat{R}(\hat{\pi}^*)$ is a $\sqrt{n}$-consistent estimator of the optimal harmless policy reward $R(\pi^*)$ for parametric policy classes under mild assumptions, which are widely widely adopted in practice (Puterman, 2014; Sutton & Barto, 2018).

Theorem 6.4 (Main result 3)**.** *Suppose that Π is a $\mathbb{P}$-G-C class, $\hat{\mu}_t(x)$ and $\hat{e}(x)$ are uniformly consistent estimators of $\mu_t(x)$ and $e(x)$ for $t = 0, 1$, respectively, and $a\pi \in \Pi$ for any $\pi \in \Pi$ and $0 < a < 1$, then we have (a) $R(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho) \xrightarrow{\mathbb{P}} 0$; and (b) $\hat{R}(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho) \xrightarrow{\mathbb{P}} 0$.*

In contrast to policy gradient learning, if we relax the parametric restriction on the policy class and extend it to the $\mathbb{P}$-Glivenko-Cantelli ($\mathbb{P}$-G-C) class (van der Vaart & Wellner, 1996), then both $R(\hat{\pi}^*; c, \rho)$ and $\hat{R}(\hat{\pi}^*; c, \rho)$ remain consistent estimators of $R(\pi^*; c, \rho)$ under mild assumptions, as concluded in Theorem 6.4 (see appendix for proofs).

Table 2. Comparison of the Naive method (maximizing estimated rewards), the proposed No-Harm (u) and No-Harm (w) methods in terms of the true reward, welfare change, and true harm on IHDP and JOBS. The CATE-based policy learning and recommendation-based policy learning are employed (with cost functions $c(x) = 0, 0.05, 0.10$), respectively, where the expected reward and counterfactual harm upper bound are estimated using **augmented inverse probability weighting (AIPW)** estimators in Section 6.

IHDP: TRUE HARM $\leq$ 13		CATE-BASED POLICY LEARNING			RECOMMENDATION-BASED POLICY LEARNING		
COST	METHOD	REWARD	ΔWELFARE	TRUE HARM	REWARD	ΔWELFARE	TRUE HARM
$c = 0.00$	NAIVE	570.96 $\pm$ 3.28 $\uparrow$	157.78 $\pm$ 4.11 $\uparrow$	19.12 $\pm$ 2.29 $\uparrow$	549.14 $\pm$ 1.61 $\uparrow$	139.16 $\pm$ 1.43 $\uparrow$	64.36 $\pm$ 0.87 $\uparrow$
	NO-HARM (u)	**496.93 $\pm$ 11.39**	**83.80 $\pm$ 10.42**	**10.34 $\pm$ 2.54**	**100.90 $\pm$ 15.11**	**43.82 $\pm$ 10.37**	**9.60 $\pm$ 2.52**
	NO-HARM (w)	459.80 $\pm$ 6.86	48.56 $\pm$ 6.82	5.98 $\pm$ 1.95	73.76 $\pm$ 15.62	31.82 $\pm$ 7.16	5.42 $\pm$ 2.26
$c = 0.05$	NAIVE	551.62 $\pm$ 4.15 $\uparrow$	154.40 $\pm$ 4.39 $\uparrow$	16.76 $\pm$ 2.21 $\uparrow$	515.30 $\pm$ 2.29 $\uparrow$	139.48 $\pm$ 1.38 $\uparrow$	64.16 $\pm$ 0.70 $\uparrow$
	NO-HARM (u)	**491.33 $\pm$ 13.44**	**84.34 $\pm$ 13.91**	**9.32 $\pm$ 2.83**	**101.42 $\pm$ 12.16**	**47.82 $\pm$ 9.10**	**9.88 $\pm$ 2.61**
	NO-HARM (w)	456.17 $\pm$ 6.88	50.50 $\pm$ 6.40	6.02 $\pm$ 2.01	67.59 $\pm$ 13.56	31.46 $\pm$ 7.74	5.98 $\pm$ 2.94
$c = 0.10$	NAIVE	534.27 $\pm$ 4.21 $\uparrow$	148.74 $\pm$ 4.10 $\uparrow$	14.86 $\pm$ 2.21 $\uparrow$	480.98 $\pm$ 2.63 $\uparrow$	139.50 $\pm$ 1.96 $\uparrow$	63.90 $\pm$ 0.96 $\uparrow$
	NO-HARM (u)	**482.14 $\pm$ 12.73**	**81.68 $\pm$ 15.00**	**8.60 $\pm$ 3.46**	**92.42 $\pm$ 15.17**	**47.34 $\pm$ 8.52**	**8.90 $\pm$ 2.84**
	NO-HARM (w)	452.29 $\pm$ 5.76	49.00 $\pm$ 7.03	5.42 $\pm$ 1.92	63.33 $\pm$ 12.54	31.82 $\pm$ 8.05	5.58 $\pm$ 2.17
JOBS: TRUE HARM $\leq$ 50		CATE-BASED POLICY LEARNING			RECOMMENDATION-BASED POLICY LEARNING		
COST	METHOD	REWARD	ΔWELFARE	TRUE HARM	REWARD	ΔWELFARE	TRUE HARM
$c = 0.00$	NAIVE	1798.60 $\pm$ 7.63 $\uparrow$	583.96 $\pm$ 10.54 $\uparrow$	113.73 $\pm$ 4.47 $\uparrow$	1965.33 $\pm$ 1.44 $\uparrow$	758.50 $\pm$ 1.52 $\uparrow$	251.30 $\pm$ 0.69 $\uparrow$
	NO-HARM (u)	**1453.00 $\pm$ 21.96**	**237.36 $\pm$ 29.81**	**43.23 $\pm$ 8.06**	**528.00 $\pm$ 22.16**	**195.73 $\pm$ 13.80**	**41.40 $\pm$ 4.85**
	NO-HARM (w)	1325.00 $\pm$ 48.62	113.74 $\pm$ 60.39	16.80 $\pm$ 8.41	197.46 $\pm$ 138.66	66.26 $\pm$ 52.88	17.16 $\pm$ 12.60
$c = 0.05$	NAIVE	1701.13 $\pm$ 10.41 $\uparrow$	566.23 $\pm$ 11.23 $\uparrow$	93.93 $\pm$ 4.68 $\uparrow$	1760.50 $\pm$ 11.30 $\uparrow$	705.26 $\pm$ 8.62 $\uparrow$	238.23 $\pm$ 3.68 $\uparrow$
	NO-HARM (u)	**1408.72 $\pm$ 27.01**	**242.66 $\pm$ 44.18**	**41.13 $\pm$ 9.31**	**504.18 $\pm$ 25.78**	**195.80 $\pm$ 18.89**	**42.86 $\pm$ 5.32**
	NO-HARM (w)	1325.56 $\pm$ 32.28	118.83 $\pm$ 55.50	19.76 $\pm$ 9.13	220.94 $\pm$ 113.36	77.30 $\pm$ 43.96	18.93 $\pm$ 8.92
$c = 0.10$	NAIVE	1612.20 $\pm$ 9.07 $\uparrow$	527.06 $\pm$ 52.29 $\uparrow$	72.66 $\pm$ 7.65 $\uparrow$	1529.96 $\pm$ 49.49 $\uparrow$	630.86 $\pm$ 34.30 $\uparrow$	212.93 $\pm$ 11.51 $\uparrow$
	NO-HARM (u)	**1362.20 $\pm$ 22.95**	**232.63 $\pm$ 51.70**	**36.26 $\pm$ 8.04**	**475.10 $\pm$ 20.52**	**193.83 $\pm$ 16.48**	**44.30 $\pm$ 5.84**
	NO-HARM (w)	1257.19 $\pm$ 39.17	67.83 $\pm$ 59.52	11.63 $\pm$ 8.91	214.76 $\pm$ 179.24	85.33 $\pm$ 82.95	22.93 $\pm$ 23.32

7. Experiments

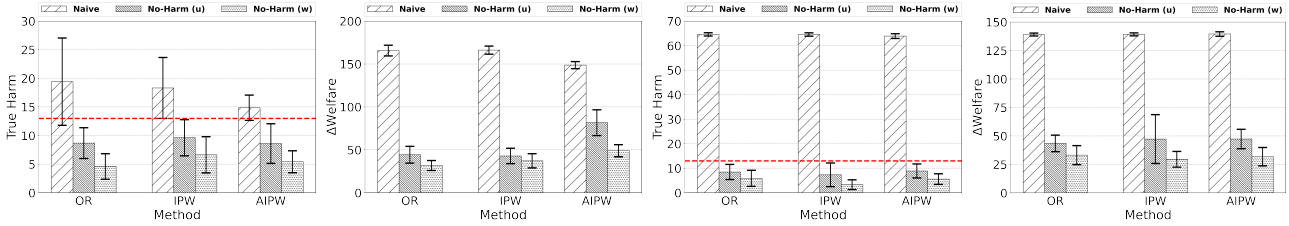
7.1. Experimental Setup

Dataset and Prepossessing. Following previous studies (Shalit et al., 2017; Louizos et al., 2017; Yoon et al., 2018; Yao et al., 2018), we conduct extensive experiments on one semi-synthetic dataset, IHDP, and one real-world dataset, JOBS. The IHDP dataset (Hill, 2011) is based on the Infant Health and Development Program (IHDP), and examines the effects of specialist home visits on future cognitive test scores. The dataset comprises 672 units (123 treated, 549 control) and 25 covariates measuring aspects of children and their mothers. The JOBS dataset (LaLonde, 1986) is based on the National Supported Work program, and examines the effects of job training on income and employment status after training. The dataset comprises 2,570 units (237 treated, 2,333 control) and 17 covariates from non-randomized observational studies.

Different from estimating causal effects, even for data collected from randomized controlled trials, we are unable to identify whether individuals are in the "harmful treatment" strata, i.e., $Y(0) = 1$ and $Y(1) = 0$. Thus, we simulate potential outcomes based on the covariates as follows: $Y_i(0) \sim \text{Bern}(\sigma(w_0 x_i + \epsilon_{0,i}))$, and $Y_i(1) \sim \text{Bern}(\sigma(w_1 x_i + \epsilon_{1,i}))$, where $\sigma(\cdot)$ is the sigmoid function, $w_0 \sim \mathcal{N}_{[-1,1]}(0, 1)$ follows a truncated normal distribution, $w_1 \sim \text{Unif}(-1, 1)$ follows a uniform distribution, $\epsilon_{0,i} \sim \mathcal{N}(\alpha_0, 1)$, and $\epsilon_{1,i} \sim \mathcal{N}(\alpha_1, 1)$. We set the noise parameters $\alpha_0 = 1$ and $\alpha_1 = 3$ for IHDP and $\alpha_0 = 0$ and $\alpha_1 = 2$ for JOBS.

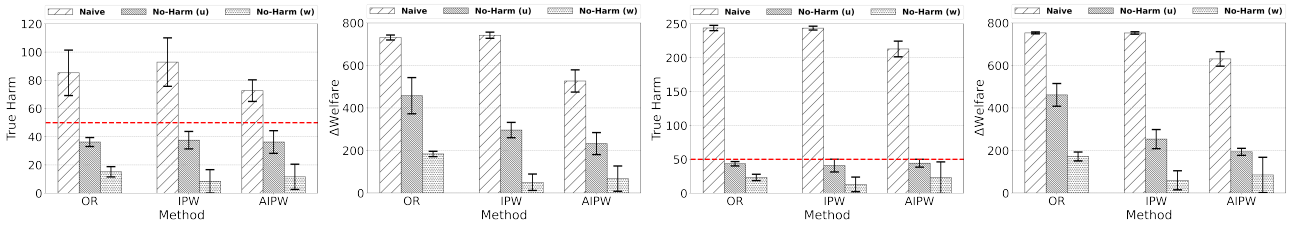
Experimental Details. The goal of our policy learning is to maximize the reward and the resulting change in welfare *while satisfying the counterfactual no-harm criterion*. Given that the simulated potential outcomes demonstrate 65 and 252 units in the "harmful treatment" strata on IHDP and JOBS, respectively, we define the counterfactual no-harm criterion as harming less than 20% of them by the learned policy, i.e., 13 units for IHDP and 50 units for JOBS. Formally, the reward for the learned policy $\pi(x)$ is $\sum_{i=1}^{n}(Y_i(1) - c)\pi(x_i) + Y_i(0)(1 - \pi(x_i))$ for CATE-based and $\sum_{i=1}^{n}(Y_i(1) - c)\pi(x_i)$ for recommendation-based policy learning, respectively. Following (Kitagawa & Tetenov, 2018), the change in welfare is defined as $\Delta W(\pi) = \sum_{i=1}^{n}[Y_i(1) \cdot \pi(x_i) + Y_i(0) \cdot (1 - \pi(x_i))] - \sum_{i=1}^{n} Y_i(0) = \sum_{i=1}^{n}[(Y_i(1) - Y_i(0)) \cdot \pi(x_i)]$. The true harm is $\sum_{i=1}^{n} \mathbb{1}\{Y_i(0) = 1, Y_i(1) = 0\} \cdot \pi(x_i)$.

We learn policies satisfying the counterfactual no-harm criterion based on the estimation of the upper bound $w_{\text{FNA}}(x)$ in Proposition 5.1 and the estimation of the upper bound $u_{\text{FNA}}(x)$ in Theorem 5.2, named "No-Harm (w)" and "No-Harm (u)" respectively, and compare them to the baseline method that directly maximizes the estimated reward. We tune the costs $c(x) = 0, 0.025, 0.05, 0.075, 0.10$ and use OR, IPW, AIPW estimators (see appendix for details).



(a) Real harm in CATE case. (b) ΔWelfare in CATE case. (c) Real harm in recommendation. (d) ΔWelfare in recommendation.

Figure 1. Comparison of Naive, No-Harm (u), and No-Harm (w) methods by OR, IPW, and AIPW estimators on IHDP.



(a) Real harm in CATE case. (b) ΔWelfare in CATE case. (c) Real harm in recommendation. (d) ΔWelfare in recommendation.

Figure 2. Comparison of Naive, No-Harm (u), and No-Harm (w) methods by OR, IPW, and AIPW estimators on JOBS.

7.2. Experimental Results

CATE and Recommendation-Based Policy Learning. We average over 100 realizations of policy learning in IHDP and JOBS, and the results are shown in Table 2. First, the Naive approach provides the largest reward and welfare change, however, it comes at the expense of individual harm and fails to satisfy the counterfactual no-harm criterion in all scenarios. Importantly, compared with the CATE-based policy learning, the recommendation-based policy learning has significantly greater true harm, which empirically validates the conclusions in Section 4.3. Meanwhile, it has a larger reward-to-welfare-change ratio, where the reward is interpreted as the total revenue gained by the ad agency through successful ad placements (expressed as purchases), the welfare change is the number of ad placements with a positive causal effect, and the true harm is related to the customer churn due to (aggressive) ad placements. Second, the proposed No-Harm (u) and No-Harm (w) satisfy the counterfactual no-harm criterion in all scenarios, with the former can result in pursuing higher reward and welfare changes while satisfying no-harm criterion due to the tighter upper bound $u_{\text{FNA}}(x) \leq w_{\text{FNA}}(x)$.

Effects of Varying Cost. We further study the effects of varying costs $c(x)$ (see appendix for more results). From Table 2, an increase in cost would decrease the total number of individuals with respect to reward, welfare change, and true harm, which empirically demonstrates the trade-off between group welfare and individual no-harm, validating the findings in Corollary 5.6. Nevertheless, even with $c(x) = 0.10$, the Naive approach still fails to satisfy the counterfactual no-harm criterion in all scenarios. Furthermore, compared with the Naive method, the proposed policy learning approaches have a more conservative treatment assignment due to the fact that they are constrained by the counterfactual harm upper bound. Thus, increasing costs may not have a significant impact on welfare changes.

Effects of OR, IPW, and AIPW Estimators. In addition to using the AIPW estimator to estimate the policy reward and counterfactual harm upper bound in Section 6, we further explore the use of outcome regression (OR), inverse probability weighting (IPW) as alternative estimators for policy learning, and the results on IHDP and JOBS are shown in Figures 1 and 2, respectively (see appendix for more results). Similar findings hold that our methods satisfy the counterfactual no-harm criterion in all scenarios, while the Naive method exhibits significantly violation of individual harm in the recommendation scenario. Moreover, the AIPW estimator on IHDP and the OR estimator on JOBS show the highest welfare increase for our proposal, respectively, which is interpreted as the effect from estimation error.

8. Conclusion

This paper formally discusses the counterfactual no-harm criterion for policy learning, with its theoretical upper bounds and estimation methods, and proves the consistency and asymptotic normality. We further propose a policy learning approach that satisfies the counterfactual no-harm criterion. One possible limitation of this study is that the optimal "no-harm" policy may harm other related outcomes, and the proposed method cannot lead to a strict counterfactual no-harm policy (with FNA equals to zero), instead we attempt to make counterfactual harm to be "controllable". Another limitation is how to specify the allowed harm more reasonably in practice, which we leave for future research.

Acknowledgements

This work has been supported in part by the National Key R&D Program of China (Grant No. 2020YFE0204200). Yue Liu was supported by the National Natural Science Foundation of China (Grant No. 12201629). Peng Wu was supported by the Disciplinary Funding of Beijing Technology and Business University. We thank Xiaojie Mao for insightful suggestions, and the anonymous reviewers for helpful comments.

References

Asimov, I. *I, robot*, volume 1. Spectra, 2004.

Athey, S. and Wager, S. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 47:1148–1178, 2019.

Ben-Michael, E., Imai, K., and Jiang, Z. Policy learning with asymmetric utilities. *arXiv preprint arXiv:2206.10479*, 2022.

Bertsimas, D., Kallus, N., Weinstein, A. M., and Zhuo, Y. D. Personalized Diabetes Management Using Electronic Medical Records. *Diabetes Care*, 40(2):210–217, 12 2016.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:1–68, 2018.

Chiappa, S. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019.

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

Ding, P., Geng, Z., Yan, W., and Zhou, X. H. Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *Journal of the American Statistical Association*, 106(496):1578–1591, 2011.

Duflo, E., Hanna, R., and Ryan, S. P. Incentives work: Getting teachers to come to school. *American Economic Review*, 102(4):1241–78, 2012.

Efron, B. and Feldman, D. Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, 86(413):9–17, 1991.

Floridi, L. Establishing the rules for building trustworthy ai. *Nature Machine Intelligence*, 1(6):261–262, 2019.

Frangakis, C. E. and Rubin, D. B. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86(2):365–379, 1999.

Frangakis, C. E. and Rubin, D. B. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

Gallop, R., Small, D. S., Lin, J. Y., Elliott, M. R., Joffe, M., and Have, T. Mediation analysis with principal stratification. *Statistics in Medicine*, 2009.

Guelman, L., Guillén, M., and Pérez-Marín, A. M. A decision support framework to implement optimal personalized marketing interventions. *Decision Support Systems*, 72:24–32, 2015.

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

Holland, P. W. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, 1986.

Imai, K. Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". *Statisitcs and Probability Letters*, 2008.

Jaskowski, M. and Jaroszewicz, S. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, volume 46, pp. 79–95, 2012.

Jiang, Z., Ding, P., and Geng, Z. Principal causal effect identification and surrogate end point evaluation by multiple trials. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(4):829–848, 2016.

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

Kallus, N. Treatment effect risk: Bounds and inference. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 213, New York, NY, USA, 2022a. Association for Computing Machinery.

Kallus, N. What's the harm? sharp bounds on the fraction negatively affected by treatment. *arXiv preprint arXiv:2205.10327*, 2022b.

Kaur, D., Uslu, S., Rittichier, K. J., and Durresi, A. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, 55(2):1–38, 2022.

Kitagawa, T. and Tetenov, A. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2), 2018.

Kosorok, M. R. *Introduction to empirical processes and semiparametric inference.* Springer, 2008.

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

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

LaLonde, R. J. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pp. 604–620, 1986.

Li, A. and Pearl, J. Unit selection based on counterfactual logic. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

Li, A. and Pearl, J. Unit selection with causal diagram. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5765–5772, 2022.

Lin, A. C. The unifying role of harm in environmental law, 2006 wis. *L. Rev*, 897:904, 2006.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Ma, X., Zhao, L., Huang, G., Wang, Z., Hu, Z., Zhu, X., and Gai, K. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1137–1140, 2018.

Mill, J. S. On liberty. In *A selection of his works*, pp. 1–147. Springer, 1966.

Morgan, S. L. and Winship, C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, second edition, 2015.

Murphy, S. A. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Neyman, J. S. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:465–472, 1990.

Pearl, J. *Causality*. Cambridge university press, 2009.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Richens, J. G., Beard, R., and Thompson, D. H. Counterfactual harm. *arXiv preprint arXiv:2204.12993*, 2022.

Robins, J., Rotnitzky, A., and Zhao, L. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994.

Robins, J., Rotnitzky, A., and Zhao, L. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90:106–121, 1995.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, 66:688–701, 1974.

Schwab, P., Linhardt, L., and Karlen, W. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

Semenova, V. and Chernozhukov, V. Debiased machine learning of conditional average treatment effects and and other causal functions. *The Econometrics Journal*, 24: 264–289, 2021.

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.

Shapiro, A. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30:169–186, 1991.

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Sokol, D. K. "first do no harm" revisited. *Bmj*, 347, 2013.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Thiebes, S., Lins, S., and Sunyaev, A. Trustworthy artificial intelligence. *Electronic Markets*, 31(2):447–464, 2021.

van der Vaart, A. W. and Wellner, J. A. *Weak convergence and empirical processes: with application to statistics*. Springer, 1996.

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

Wang, H., Chang, T.-W., Liu, T., Huang, J., Chen, Z., Yu, C., Li, R., and Chu, W. ESCM^2: Entire space counterfactual multi-task model for post-click conversion rate estimation. In *SIGIR*, 2022.

Wu, Y., Zhang, L., Wu, X., and Tong, H. PC-Fairness: A unified framework for measuring causality-based fairness. *Advances in Neural Information Processing Systems*, 32, 2019.

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, J. L. and Rubin, D. B. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368, 2003.

Zhang, W., Bao, W., Liu, X., Yang, K., Lin, Q., Wen, H., and Ramezani, R. Large-scale causal approaches to debiasing post-click conversion rate estimation with multi-task learning. In *WWW*, 2020.

A. Proofs in Section 5

Proposition 5.1 (Tight upper bounds). *(a) The tight upper bound of FNA(x), named $w_{\mathrm{FNA}}(x)$, is*

$$\min(\mathbb{P}(Y=1|T=0,X=x),\mathbb{P}(Y=0|T=1,X=x));$$

(b) Given a policy $\pi\in\Pi$, the tight upper bound of the FNA(π) is $\mathbb{E}[w_{\mathrm{FNA}}(X)\pi(X)]$.

Proof of Proposition 5.1. (a) It suffices to show that

$$\mathrm{FNA}(x)\leq\min(\mathbb{P}(Y=1|T=0,X=x),\mathbb{P}(Y=0|T=1,X=x)).$$

This follows immediately from the truth that

$$\begin{aligned}\mathrm{FNA}(x)=\mathbb{P}(Y(0)=1,Y(1)=0|X=x)&\leq\min(\mathbb{P}(Y(0)=1|X=x),\mathbb{P}(Y(1)=0|X=x))\\&=\min(\mathbb{P}(Y=1|T=0,X=x),\mathbb{P}(Y=0|T=1,X=x))\end{aligned}$$

(b) Given a policy $\pi\in\Pi$, the fraction harmed by the policy π is

$$\mathrm{FNA}(\pi)=\mathbb{E}[\mathbb{P}(Y(0)=1,Y(1)=0|X)\pi(X)]\leq\mathbb{E}[\min(\mathbb{P}(Y=1|T=0,X=x),\mathbb{P}(Y=0|T=1,X=x)\pi(X)].$$

□

Theorem 5.2 (Main result 1). *(a) If $Y(0)$ and $Y(1)$ are non-negatively correlated given $X=x$, the tight upper bound of the FNA(x), named $u_{\mathrm{FNA}}(x)$, is*

$$\mathbb{P}(Y=1|T=0,X=x)\mathbb{P}(Y=0|T=1,X=x);$$

(b) Given a policy $\pi\in\Pi$, the tight upper bound of the FNA(π) is $\mathbb{E}[u_{\mathrm{FNA}}(X)\pi(X)]$.

Proof of Theorem 5.2. (a) The non-negative conditional correlation condition is equivalent to

$$\begin{aligned}&\mathbb{P}(Y(0)=0,Y(1)=0\mid X=x)(0-\mathbb{P}(Y_{1,0}\mid X=x)-\mathbb{P}(Y_{1,1}\mid X=x))(0-\mathbb{P}(Y_{0,1}\mid X=x)-\mathbb{P}(Y_{1,1}\mid X=x))\\&+\mathbb{P}(Y(0)=0,Y(1)=1\mid X=x)(0-\mathbb{P}(Y_{1,0}\mid X=x)-\mathbb{P}(Y_{1,1}\mid X=x))(1-\mathbb{P}(Y_{0,1}\mid X=x)-\mathbb{P}(Y_{1,1}\mid X=x))\\&+\mathbb{P}(Y(0)=1,Y(1)=0\mid X=x)(1-\mathbb{P}(Y_{1,0}\mid X=x)-\mathbb{P}(Y_{1,1}\mid X=x))(0-\mathbb{P}(Y_{0,1}\mid X=x)-\mathbb{P}(Y_{1,1}\mid X=x))\\&+\mathbb{P}(Y(0)=1,Y(1)=1\mid X=x)(1-\mathbb{P}(Y_{1,0}\mid X=x)-\mathbb{P}(Y_{1,1}\mid X=x))(1-\mathbb{P}(Y_{0,1}\mid X=x)-\mathbb{P}(Y_{1,1}\mid X=x))\geq 0.\end{aligned}$$

For simplicity, we denote $\mathbb{P}(Y=1\mid T=0,X=x)=\mathbb{P}(Y_{1,0}\mid X=x)+\mathbb{P}(Y_{1,1}\mid X=x)$ as $\mu_0(x)$, and denote $\mathbb{P}(Y=1\mid T=1,X=x)=\mathbb{P}(Y_{0,1}\mid X=x)+\mathbb{P}(Y_{1,1}\mid X=x)$ as $\mu_1(x)$, then we have

$$\begin{aligned}&(1-\mu_1(x))\,(1-\mu_0(x))\,(\mu_0(x)-\mathbb{P}(Y_{1,0}\mid X=x))+\mu_1(x)\mu_0(x)\,(1-\mu_1(x)-\mathbb{P}(Y_{1,0}\mid X=x))\\&-(1-\mu_1(x))\,\mu_0(x)\,\{\mu_1(x)-(\mu_0(x)-\mathbb{P}(Y_{1,0}\mid X=x))\}-\mu_1(x)\,(1-\mu_0(x))\,\mathbb{P}(Y_{1,0}\mid X=x)\geq 0,\end{aligned}$$

which is equivalent to

$$\begin{aligned}&(1-\mu_1(x))\,(1-\mu_0(x))\,\mu_1(x)+\mu_1(x)\mu_0(x)\,(1-\mu_0(x))+(1-\mu_0(x))\,\mu_1(x)\,(\mu_1(x)-\mu_0(x))\\&\geq\mathbb{P}(Y_{0,1}\mid X=x)\,\{(1-\mu_1(x))\,(1-\mu_0(x))+\mu_1(x)\mu_0(x)+(1-\mu_1(x))\,\mu_0(x)+\mu_1(x)\,(1-\mu_0(x))\}\,.\end{aligned}$$

Note that

$$\begin{aligned}&\mathbb{P}(Y(0)=0,Y(1)=0\mid X=x)+\mathbb{P}(Y(0)=0,Y(1)=1\mid X=x)\\&+\mathbb{P}(Y(0)=1,Y(1)=0\mid X=x)+\mathbb{P}(Y(0)=1,Y(1)=1\mid X=x)=1,\end{aligned}$$

which leads to

$$\begin{aligned}\mathbb{P}(Y_{0,1}\mid X=x)&\leq\mu_1(x)\,(1-\mu_0(x))\\&=(\mathbb{P}(Y_{0,1}\mid X=x)+\mathbb{P}(Y_{1,1}\mid X=x))(\mathbb{P}(Y_{0,1}\mid X=x)+\mathbb{P}(Y_{0,0}\mid X=x))\\&=\mathbb{P}(Y(0)=1|X=x)\mathbb{P}(Y(1)=0|X=x),\end{aligned}$$

which implies $\text{FNA}(x) \leq \mathbb{P}(Y(0)=1|X=x)\mathbb{P}(Y(1)=0|X=x)$.

(b) Given a policy $\pi \in \Pi$, the fraction harmed by the policy π is

$$\text{FNA}(\pi) = \mathbb{E}[\mathbb{P}(Y(0)=1, Y(1)=0|X)\pi(X)] \leq \mathbb{E}[\mathbb{P}(Y(0)=1|X=x)\mathbb{P}(Y(1)=0|X=x)\pi(X)].$$

□

Theorem 5.5 (Relation between CATEs and upper bounds). For the upper bounds $w_{\text{FNA}}(x)$ in Proposition 5.1 and $u_{\text{FNA}}(x)$ in Theorem 5.2, for all $x \in \mathcal{X}$, we have

$$w_{\text{FNA}}(x) \leq \frac{1-\tau(x)}{2}, \quad \text{and} \quad u_{\text{FNA}}(x) \leq \frac{(1-\tau(x))^2}{4}.$$

Proof of Theorem 5.5. By $\min(a,b) \leq (a+b)/2$, and let $a = \mathbb{P}(Y=1|X=x,T=0)$ and $b = \mathbb{P}(Y=0|X=x,T=1)$, we have

$$w_{\text{FNA}}(x) = \min\{\mathbb{P}(Y=1|X=x,T=0), \mathbb{P}(Y=0|X=x,T=1)\} \leq \frac{1-\tau(x)}{2}.$$

We then show that the conditional average treatment effects (CATEs) $\tau(x)$ can be written as

$$\begin{aligned}\tau(x) &= \mathbb{P}(Y(1)-Y(0)|X=x) = \mathbb{P}(Y(1)=1|X=x) - \mathbb{P}(Y(0)=1|X=x) \\ &= \mathbb{P}(Y=1|X=x,T=1) - \mathbb{P}(Y=1|X=x,T=0),\end{aligned}$$

which leads to

$$\mathbb{P}(Y=1|X=x,T=0) + \mathbb{P}(Y=0|X=x,T=1) = 1-\tau(x).$$

By the inequality of arithmetic and geometric means that $ab \leq (a+b)^2/4$, and let $a = \mathbb{P}(Y=1|X=x,T=0)$ and $b = \mathbb{P}(Y=0|X=x,T=1)$, we have

$$u_{\text{FNA}}(x) = \mathbb{P}(Y=1|X=x,T=0)\mathbb{P}(Y=0|X=x,T=1) \leq \frac{(1-\tau(x))^2}{4}.$$

□

Corollary 5.6 (Relation to the cost). *For the upper bound $w_{\text{FNA}}(\pi)$ in Proposition 5.1 and $u_{\text{FNA}}(\pi)$ in Theorem 5.2, the optimal policies π^* in Section 4.2 satisfy*

$$w_{\text{FNA}}(\pi^*) \leq \mathbb{E}\Big[\frac{1-c(X)}{2}\pi^*(X)\Big], \quad \textit{and} \quad u_{\text{FNA}}(\pi^*) \leq \mathbb{E}\Big[\frac{(1-c(X))^2}{4}\pi^*(X)\Big].$$

Proof of Corollary 5.6. Recall the optimal policies π^* in Section 4.2 is

$$\pi^*(x;c) = \begin{cases} 1, & \mathbb{E}[Y(1)-Y(0) \mid X=x] = \tau(x) > c(x) \\ 0, & \mathbb{E}[Y(1)-Y(0) \mid X=x] = \tau(x) < c(x), \\ d, & \mathbb{E}[Y(1)-Y(0) \mid X=x] = \tau(x) = c(x) \end{cases}$$

where d is any value between 0 and 1, and π^* would always impose a treatment intervention $T=1$ for individuals whose $\tau(x)$ is greater than the cost $c(x)$. It follows directly that

$$\begin{aligned} w_{\text{FNA}}(\pi^*) &\leq \mathbb{E}\Big[\frac{1-\tau(X)}{2}\pi^*(X)\Big] \leq \mathbb{E}\Big[\frac{1-c(X)}{2}\pi^*(X)\Big], \text{ and} \\ u_{\text{FNA}}(\pi^*) &\leq \mathbb{E}\Big[\frac{(1-\tau(X))^2}{4}\pi^*(X)\Big] \leq \mathbb{E}\Big[\frac{(1-c(X))^2}{4}\pi^*(X)\Big]. \end{aligned}$$

□

B. Proofs in Section 6

B.1. Proofs of Lemma 6.1 and Theorem 6.2

Lemma 6.1. For all $\pi \in \Pi$,

$$R(\pi; c, \rho) = \mathbb{E}[\varphi_\pi(T, X, Y; e, \mu_0, \mu_1)], \quad u_{\text{FNA}}(\pi) = \mathbb{E}[\psi_\pi(T, X, Y; e, \mu_0, \mu_1)].$$

Proof of Lemma 6.1. Under strong ignorability assumption, the upper bound of $\text{FNA}(x)$ can be reformulated as

$$\begin{aligned} u_{\text{FNA}}(x) &= \mathbb{E}(Y(0)|X = x) \cdot [1 - \mathbb{E}(Y(1)|X = x)] \\ &= \mu_0(x) \cdot \{1 - \mu_1(x)\} \\ &= \mathbb{E}\left\{ \frac{(1-T)\{Y - \mu_0(X)\}}{1 - e(X)} + \mu_0(X) \Big| X = x \right\} \cdot \left[1 - \mathbb{E}\left\{ \frac{T\{Y - \mu_1(X)\}}{e(X)} + \mu_1(X) \Big| X = x \right\} \right], \end{aligned}$$

which implies that

$$\begin{aligned} u_{\text{FNA}}(\pi) &= \mathbb{E}[u_{\text{FNA}}(X) \cdot \pi(X)\}] \\ &= \mathbb{E}\left[\mathbb{E}\left(\frac{(1-T)\{Y - \mu_0(X)\}}{1 - e(X)} + \mu_0(X) \Big| X \right) \cdot \left\{ 1 - \mathbb{E}\left(\frac{T\{Y - \mu_1(X)\}}{e(X)} + \mu_1(X) \Big| X \right) \right\} \cdot \pi(X) \right] \\ &= \mathbb{E}\left[\mathbb{E}\left(\frac{(1-T)\{Y - \mu_0(X)\}}{1 - e(X)} + \mu_0(X) \Big| X \right) \cdot \pi(X) \right] \\ &\quad - \mathbb{E}\left[\mathbb{E}\left(\frac{(1-T)\{Y - \mu_0(X)\}}{1 - e(X)} + \mu_0(X) \Big| X \right) \cdot \mathbb{E}\left(\frac{T\{Y - \mu_1(X)\}}{e(X)} + \mu_1(X) \Big| X \right) \cdot \pi(X) \right] \\ &= \mathbb{E}\left[\left(\frac{(1-T)\{Y - \mu_0(X)\}}{1 - e(X)} + \mu_0(X) \right) \cdot \pi(X) \right] - \mathbb{E}\left[\mu_0(X) \cdot \mathbb{E}\left(\frac{T\{Y - \mu_1(X)\}}{e(X)} + \mu_1(X) \Big| X \right) \cdot \pi(X) \right] \\ &= \mathbb{E}\left[\left(\frac{(1-T)\{Y - \mu_0(X)\}}{1 - e(X)} + \mu_0(X) \right) \cdot \pi(X) \right] - \mathbb{E}\left[\mu_0(X) \cdot \left(\frac{T\{Y - \mu_1(X)\}}{e(X)} + \mu_1(X) \right) \cdot \pi(X) \right], \end{aligned}$$

which implies $u_{\text{FNA}}(\pi) = \mathbb{E}[\psi_\pi(T, X, Y; e, \mu_0, \mu_1)]$. Similarly, one can get $R(\pi; c, \rho) = \mathbb{E}[\varphi_\pi(T, X, Y; e, \mu_0, \mu_1)]$. This finishes the proof.

□

Theorem 6.2. *Suppose that* $||\hat{e}(x) - e(x)||_2 \cdot ||\hat{\mu}_t(x) - \mu_t(x)||_2 = o_{\mathbb{P}}(n^{-1/2})$ *for all* $x \in \mathcal{X}$ *and* $t \in \{0, 1\}$,

(a) $\hat{R}(\pi, c; \rho)$ *is consistent and asymptotically normal*

$$\sqrt{n}\{\hat{R}(\pi, c; \rho) - R(\pi, c; \rho)\} \to N(0, \sigma_1^2),$$

where $\sigma_1^2 = \mathbb{V}[\varphi_\pi(T, X, Y; e, \mu_0, \mu_1)]$.

(b) if $\mu_0(x) = \mu_0(x; \phi)$ *is a parametric model,* $\hat{u}_{\text{FNA}}(\pi)$ *is consistent and asymptotically normal*

$$\sqrt{n}\{\hat{u}_{\text{FNA}}(\pi) - u_{\text{FNA}}(\pi)\} \to N(0, \sigma_2^2),$$

where

$$\sigma_2^2 = \mathbb{V}\left[\psi_\pi(T, X, Y; e, \mu_0, \mu_1) - s(X)\mathbb{E}\Big\{ \frac{\partial \mu_0(X; \phi)}{\partial \phi} \mu_1(X)\pi(X) \Big\} \right],$$

and $s(X)$ *is the influence function of estimator of* ϕ.

Proof of Theorem 6.2. We first prove Theorem 6.2(b), and Theorem 6.2(a) can be derived analogously. Recall that

$$\psi_\pi(T, X, Y; \hat{e}, \hat{\mu}_0, \hat{\mu}_1) = \left(\frac{(1-T)\{Y - \hat{\mu}_0(X)\}}{1 - \hat{e}(X)} + \hat{\mu}_0(X) \right) \pi(X) - \left(\frac{T\{Y - \hat{\mu}_1(X)\}}{\hat{e}(X)} + \hat{\mu}_1(X) \right) \hat{\mu}_0(X)\pi(X)$$

and define

$$\tilde{\psi}_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1) = \left(\frac{(1-T)\{Y-\hat{\mu}_0(X)\}}{1-\hat{e}(X)} + \hat{\mu}_0(X)\right)\pi(X) - \left(\frac{T\{Y-\hat{\mu}_1(X)\}}{\hat{e}(X)} + \hat{\mu}_1(X)\right)\mu_0(X)\pi(X).$$

The estimator $\hat{u}_{\mathrm{FNA}}(\pi)$ can be decomposed as

$$\hat{u}_{\mathrm{FNA}}(\pi) - u_{\mathrm{FNA}}(\pi) = A_{1n} + A_{2n} + A_{3n},$$

where

$$\begin{aligned}
A_{1n} &= \frac{1}{n}\sum_{i=1}^{n}[\psi_\pi(T_i,X_i,Y_i;e,\mu_0,\mu_1) - u_{\mathrm{FNA}}(\pi)],\\
A_{2n} &= \frac{1}{n}\sum_{i=1}^{n}[\tilde{\psi}_\pi(T_i,X_i,Y_i;\hat{e},\hat{\mu}_0,\hat{\mu}_1) - \psi_\pi(T_i,X_i,Y_i;e,\mu_0,\mu_1)],\\
A_{3n} &= \frac{1}{n}\sum_{i=1}^{n}[\psi_\pi(T_i,X_i,Y_i;\hat{e},\hat{\mu}_0,\hat{\mu}_1) - \tilde{\psi}_\pi(T_i,X_i,Y_i;\hat{e},\hat{\mu}_0,\hat{\mu}_1)].
\end{aligned}$$

Note that A_{1n} is a sum of n independent variables with zero means. Next, we focus on analyzing A_{2n}, which can be be further expanded as

$$A_{2n} = A_{2n} - \mathbb{E}[A_{2n}] + \mathbb{E}[A_{2n}].$$

Define the Gateaux derivative of the generic function g in the direction $[\hat{e}-e, \hat{\mu}_0-\mu_0, \hat{\mu}_1-\mu_1]$ by $\partial_{[\hat{e}-e,\hat{\mu}_0-\mu_0,\hat{\mu}_1-\mu_1]}g$. By a Taylor expansion for $\mathbb{E}[A_{2n}]$ yields that

$$\begin{aligned}
\mathbb{E}[A_{2n}] &= \mathbb{E}[\tilde{\psi}_\pi(T_i,X_i,Y_i;\hat{e},\hat{\mu}_0,\hat{\mu}_1) - \psi_\pi(T_i,X_i,Y_i;e,\mu_0,\mu_1)]\\
&= \partial_{[\hat{e}-e,\hat{\mu}_0-\mu_0,\hat{\mu}_1-\mu_1]}\mathbb{E}[\psi_\pi(T_i,X_i,Y_i;e,\mu_0,\mu_1)]\\
&\quad + \frac{1}{2}\partial^2_{[\hat{e}-e,\hat{\mu}_0-\mu_0,\hat{\mu}_1-\mu_1]}\mathbb{E}[\psi_\pi(T_i,X_i,Y_i;e,\mu_0,\mu_1)] + \cdots
\end{aligned}$$

The first-order term

$$\begin{aligned}
&\partial_{[\hat{e}-e,\hat{\mu}_0-\mu_0,\hat{\mu}_1-\mu_1]}\mathbb{E}[\psi_\pi(T_i,X_i,Y_i;e,\mu_0,\mu_1)]\\
&= \mathbb{E}\Bigg[\left\{\frac{(1-T)\{Y-\mu_0(X)\}}{(1-e(X))^2}\{\hat{e}(X)-e(X)\} + \left(1-\frac{1-T}{1-e(X)}\right)\{\hat{\mu}_1(X)-\mu_1(X)\}\right\}\pi(X)\\
&\quad - \left\{-\frac{T\{Y-\mu_1(X)\}}{e(X)^2}\{\hat{e}(X)-e(X)\} + \left(1-\frac{T}{e(X)}\right)\{\hat{\mu}_1(X)-\mu_1(X)\}\right\}\mu_0(X)\pi(X)\Bigg]\\
&= 0,
\end{aligned}$$

where the last equation follows from the sample splitting. For the second-order term, we get

$$\begin{aligned}
&\frac{1}{2}\partial^2_{[\hat{e}-e,\hat{\mu}_0-\mu_0,\hat{\mu}_1-\mu_1]}\mathbb{E}[\psi_\pi(T_i,X_i,Y_i;e,\mu_0,\mu_1)]\\
&= \mathbb{E}\Bigg[\left\{\frac{(1-T)\{Y-\mu_0(X)\}}{(1-e(X))^3}\{\hat{e}(X)-e(X)\}^2 + \frac{1-T}{(1-e(X))^2}\{\hat{e}(X)-e(X)\}\{\hat{\mu}_1(X)-\mu_1(X)\}\right\}\pi(X)\\
&\quad - \left\{\frac{T\{Y-\mu_1(X)\}}{e(X)^3}\{\hat{e}(X)-e(X)\}^2 + \frac{T}{e(X)^2}\{\hat{e}(X)-e(X)\}\{\hat{\mu}_1(X)-\mu_1(X)\}\right\}\mu_0(X)\pi(X)\Bigg]\\
&= \mathbb{E}\left[\frac{1-T}{(1-e(X))^2}\{\hat{e}(X)-e(X)\}\{\hat{\mu}_1(X)-\mu_1(X)\}\pi(X) - \frac{T}{e(X)^2}\{\hat{e}(X)-e(X)\}\{\hat{\mu}_1(X)-\mu_1(X)\}\mu_0(X)\pi(X)\right]\\
&= \mathbb{E}\left[\frac{1}{1-e(X)}\{\hat{e}(X)-e(X)\}\{\hat{\mu}_1(X)-\mu_1(X)\}\pi(X) - \frac{1}{e(X)}\{\hat{e}(X)-e(X)\}\{\hat{\mu}_1(X)-\mu_1(X)\}\mu_0(X)\pi(X)\right]\\
&\leq C\cdot\mathbb{E}\Big[\{\hat{e}(X)-e(X)\}\{\hat{\mu}_1(X)-\mu_1(X)\} - \{\hat{e}(X)-e(X)\}\{\hat{\mu}_1(X)-\mu_1(X)\}\Big]\\
&\leq C\cdot||\hat{e}(X)-e(X)||_2\cdot\Big\{||\hat{\mu}_1(X)-\mu_1(X)||_2 + ||\hat{\mu}_0(X)-\mu_0(X)||_2\Big\}\\
&= o_{\mathbb{P}}(n^{-1/2}),
\end{aligned}$$

where C is a finite constant. All higher-order terms can be shown to be dominated by the second-order term. Therefore,

$$\mathbb{E}[A_{2n}] = o_{\mathbb{P}}(n^{-1/2}).$$

In addition, we get that $A_{2n} - \mathbb{E}[A_{2n}] = o_{\mathbb{P}}(n^{-1/2})$ by calculating $\text{Var}\{\sqrt{n}(A_{1n} - \mathbb{E}[A_{1n}])\} = o_{\mathbb{P}}(1)$.

$$\begin{aligned}
A_{3n} &= \frac{1}{n}\sum_{i=1}^{n}[\psi_\pi(T_i, X_i, Y_i; \hat{e}, \hat{\mu}_0, \hat{\mu}_1) - \tilde{\psi}_\pi(T_i, X_i, Y_i; \hat{e}, \hat{\mu}_0, \hat{\mu}_1)] \\
&= -\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_0(X_i) - \mu_0(X_i))\left(\frac{T_i\{Y_i(1) - \hat{\mu}_1(X_i)\}}{\hat{e}(X_i)} + \hat{\mu}_1(X_i)\right)\pi(X_i) \\
&= -\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_0(X_i) - \mu_0(X_i))Y_i(1)\pi(X_i) - \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_0(X_i) - \mu_0(X_i))\left(\frac{T_i}{\hat{e}(X_i)} - 1\right)\{Y_i(1) - \hat{\mu}_1(X_i)\}\pi(X_i), \\
&\triangleq A_{3n1} + A_{3n2}.
\end{aligned}$$

It can be shown that $A_{3n2} = o_{\mathbb{P}}(n^{-1/2})$ by calculating its expectation and variance.

$$\begin{aligned}
\sqrt{n}A_{3n1} &= -n^{-1/2}\sum_{i=1}^{n}(\hat{\mu}_0(X_i) - \mu_0(X_i))Y_i(1)\pi(X_i) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mu_0(X_i;\phi)}{\partial\phi}(\hat{\phi} - \phi)Y_i(1)\pi(X_i) + o_{\mathbb{P}}(1) \\
&= -\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mu_0(X_i;\phi)}{\partial\phi}Y_i(1)\pi(X_i)\right](\hat{\phi} - \phi) + o_{\mathbb{P}}(1) \\
&= -\mathbb{E}[\frac{\partial\mu_0(X;\phi)}{\partial\phi}\mu_1(X)\pi(X)]n^{-1/2}\sum_{i=1}^{n}s(X_i) + o_{\mathbb{P}}(1).
\end{aligned}$$

Combing the results of A_{1n}, A_{2n}, and A_{3n} yields that

$$\begin{aligned}
&\sqrt{n}(\hat{u}_{\text{FNA}}(\pi) - u_{\text{FNA}}(\pi)) = \sqrt{n}(A_{1n} + A_{2n} + A_{3n}) \\
&= n^{-1/2}\sum_{i=1}^{n}\left[\psi_\pi(T_i, X_i, Y_i; e, \mu_0, \mu_1) - u_{\text{FNA}}(\pi) - \mathbb{E}[\frac{\partial\mu_0(X;\phi)}{\partial\phi}\mu_1(X)\pi(X)]s(X_i)\right] + o_{\mathbb{P}}(1),
\end{aligned}$$

which implies the conclusion of Theorem 6.2(b).

Next, the estimator $\hat{R}(\pi; c, \rho)$ can be written as

$$\hat{R}(\pi; c, \rho) - R(\pi; c, \rho) = B_{1n} + B_{2n},$$

where

$$\begin{aligned}
B_{1n} &= \frac{1}{n}\sum_{i=1}^{n}[\varphi_\pi(T_i, X_i, Y_i; e, \mu_0, \mu_1) - R(\pi; c, \rho)], \\
B_{2n} &= \frac{1}{n}\sum_{i=1}^{n}[\varphi_\pi(T_i, X_i, Y_i; \hat{e}, \hat{\mu}_0, \hat{\mu}_1) - \varphi_\pi(T_i, X_i, Y_i; e, \mu_0, \mu_1)].
\end{aligned}$$

It can shown that $B_{2n} = o_{\mathbb{P}}(n^{-1/2})$ by a similar arguments of A_{2n}. Then by the central limit theorem,

$$\sqrt{n}B_{1n} \xrightarrow{d} N\left(0, \sigma_2^2\right).$$

This completes the proof.

□

B.2. Proof of Theorem 6.3

Lemma B.1. *(Shapiro, 1991) Let Θ be a compact subset of $\mathbb{R}^k$. Let $C(\Theta)$ denote the set of continuous real-valued functions on Θ, with $\mathcal{L} = C(\Theta) \times \ldots \times C(\Theta)$ the r-dimensional Cartesian product. Let $\psi(\theta) = (\psi_0, \ldots, \psi_r) \in \mathcal{L}$ be a vector of convex functions. Consider the quantity α^* defined as the solution to the following convex optimization program:*

$$\alpha^* = \min_{\theta \in \Theta} \psi_0(\theta)$$
$$\textit{subject to } \psi_j(\theta) \leq 0, j = 1, \ldots, r$$

Assume that Slater's condition holds, so that there is some $\theta \in \Theta$ for which the inequalities are satisfied and non-affine inequalities are strictly satisfied, i.e. $\psi_j(\theta) < 0$ if ψ_j is non-affine. Now consider a sequence of approximating programs, for $n = 1, 2, \ldots$:

$$\widehat{\alpha}_n = \min_{\theta \in \Theta} \quad \widehat{\psi}_{0n}(\theta)$$
$$\textit{subject to } \widehat{\psi}_{jn}(\theta) \leq 0, j = 1, \ldots, r$$

with $\widehat{\psi}_n(\theta) := \left(\widehat{\psi}_{0n}, \ldots, \widehat{\psi}_{rn}\right) \in \mathcal{L}$. Assume that $f(n)\left(\widehat{\psi}_n - \psi\right)$ converges in distribution to a random element $W \in \mathcal{L}$ for some real-valued function $f(n)$. Then:

$$f(n)\left(\widehat{\alpha}_n - \alpha_0\right) \rightsquigarrow L$$

for a particular random variable L. It follows that $\widehat{\alpha}_n - \alpha_0 = O_{\mathbb{P}}(1/f(n))$.

□

Theorem 6.3 *Suppose that for all $\pi \in \Pi$, $\pi(x) = \pi(x; \theta)$ is a continuously differentiable and convex function with respect to θ, where $\theta \in \Theta$ is a compact set, then under the assumptions in Theorem 6.2(b), we have*

(a) $R(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho) = O_{\mathbb{P}}(1/\sqrt{n})$;

(b) $\hat{R}(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho) = O_{\mathbb{P}}(1/\sqrt{n})$.

Proof of Theorem 6.3. We first show the Theorem 6.3(b). Under Assumptions in Theorem 6.2(b), we have that

$$\sqrt{n}\begin{pmatrix} \hat{R}(\pi; c, \rho) - R(\pi; c, \rho) \\ \hat{u}_{\text{FNA}}(\pi) - u_{\text{FNA}}(\pi) \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} \psi_\pi(T_i, X_i, Y_i; e, \mu_0, \mu_1) - u_{\text{FNA}}(\pi) - \mathbb{E}[\frac{\partial \mu_0(X;\phi)}{\partial \phi} \mu_1(X)\pi(X)]s(X_i) \\ \varphi_\pi(T_i, X_i, Y_i; e, \mu_0, \mu_1) - R(\pi; c, \rho) \end{pmatrix} + o_{\mathbb{P}}(1)$$

By the central limit theorem,

$$\sqrt{n}\begin{pmatrix} \hat{R}(\pi; c, \rho) - R(\pi; c, \rho) \\ \hat{u}_{\text{FNA}}(\pi) - u_{\text{FNA}}(\pi) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right),$$

where

$$\Sigma = \mathbb{V}\left(\begin{pmatrix} \psi_\pi(T, X, Y; e, \mu_0, \mu_1) - \mathbb{E}[\frac{\partial \mu_0(X;\phi)}{\partial \phi} \mu_1(X)\pi(X)]s(X) \\ \varphi_\pi(T, X, Y; e, \mu_0, \mu_1)) \end{pmatrix}\right).$$

This implies that

$$\begin{pmatrix} \hat{R}(\pi; c, \rho) - R(\pi; c, \rho) \\ \hat{u}_{\text{FNA}}(\pi) - u_{\text{FNA}}(\pi) \end{pmatrix} = O_{\mathbb{P}}(n^{-1/2}).$$

Under Assumptions that for all $\pi \in \Pi$, $\pi(x) = \pi(x; \theta)$ is a continuously differentiable and convex function with respect to θ, where $\theta \in \Theta$ is a compact set. We then show the Slater's condition holds, so that there is some $\theta \in \Theta$ for which the non-affine inequalities are strictly satisfied. In fact, one can consider a policy that treat $T = 0$ for all $x \in \mathcal{X}$, so that for all $\epsilon > 0$, we have $0 = \text{FNA}(\pi) \leq \hat{u}_{\text{FNA}}(\pi) < \epsilon$ and the inequality is strictly satisfied. The convexity of $\hat{R}(\pi; c, \rho)$ and $\hat{u}_{\text{FNA}}(\pi)$ follows directly from the convexity of $\pi(x) = \pi(x; \theta)$ with respect to θ, and the linearity of $\hat{R}(\pi; c, \rho)$ and $\hat{u}_{\text{FNA}}(\pi)$ with respect to $\pi \in \Pi$. Now consider the following convex optimization problem

$$\hat{R}(\hat{\pi}^*; c, \rho) = \max_{\pi \in \Pi} \hat{R}(\pi; c, \rho) = \frac{1}{n} \sum_{i=1}^{n} \varphi_\pi(T_i, X_i, Y_i; \hat{e}, \hat{\mu}_0, \hat{\mu}_1)$$
$$\text{s.t.} \quad \hat{u}_{\text{FNA}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \psi_\pi(T_i, X_i, Y_i; \hat{e}, \hat{\mu}_0, \hat{\mu}_1) \leq \epsilon,$$

Then the conclusion of Theorem 6.3(b) follows from the direct application of Lemma B.1, and $f(n) = \sqrt{n}$.

Next, we prove Theorem 6.3(a). Note that

$$R(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho) = R(\hat{\pi}^*; c, \rho) - \hat{R}(\hat{\pi}^*; c, \rho) + \hat{R}(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho),$$

the first term of the right side is $O_p(1/\sqrt{n})$ by the conclusion of Theorem 6.2(a), second term of the right side also is $O_p(1/\sqrt{n})$ by Theorem 6.3(b). Thus, $R(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho) = O_p(1/\sqrt{n})$.

□

B.3. Proof of Theorem 6.4

Theorem 6.4 (Main result 3) *Suppose that* Π *is a* $\mathbb{P}$*-G-C class,* $\hat{\mu}_t(x)$ *and* $\hat{e}(x)$ *are uniformly consistent estimators of* $\mu_t(x)$ *and* $e(x)$ *for* $t = 0, 1$*, respectively, and* $a\pi \in \Pi$ *for any* $\pi \in \Pi$ *and* $0 < a < 1$*, then we have*

(a) $R(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho) = o_{\mathbb{P}}(1)$.

(b) $\hat{R}(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho) = o_{\mathbb{P}}(1)$.

Proof of Theorem 6.4. For clarity, we summarize the conditions in Theorem 6.4 as

(C1) $\hat{\mu}_0(x)$, $\hat{\mu}_1(x)$, and $\hat{e}(x)$ are uniformly consistent estimators of $\mu_0(x)$, $\mu_1(x)$, and $e(x)$, respectively,

(C2) Π is a $\mathbb{P}$-G-C class,

(C3) For any $\pi \in \Pi$ and $0 < a < 1$, $a\pi \in \Pi$,

and recall that

$$\hat{\pi}^* = \arg\max_{\pi \in \Pi} \hat{R}(\pi; c, \rho) = \arg\max_{\pi \in \Pi} \frac{1}{n} \sum_{i=1}^{n} \varphi_\pi(T_i, X_i, Y_i; \hat{e}, \hat{\mu}_0, \hat{\mu}_1)$$
$$\text{subject to } \frac{1}{n} \sum_{i=1}^{n} \psi_\pi(T_i, X_i, Y_i; \hat{e}, \hat{\mu}_0, \hat{\mu}_1) \leq \lambda,$$

$$\pi^* = \arg\max_{\pi \in \Pi} R(\pi; c, \rho) = \arg\max_{\pi \in \Pi} \mathbb{E}\psi_\pi(T, X, Y; e, \mu_0, \mu_1)$$
$$\text{subject to } \quad \mathbb{E}[\psi_\pi(T_i, X_i, Y_i; e, \mu_0, \mu_1)] \leq \lambda.$$

Note that $\frac{T(Y-\mu_1(X))}{e(X)} + \mu_1(X) - c(X)$, $\frac{(1-T)(Y-\mu_0(X))}{1-e(X)} + \mu_0(X)$, $(\frac{T(Y-\mu_1(X))}{e(X)} + \mu_1(X))\mu_0(X)$ are bounded random variables, we have both $\{\varphi_\pi : \pi \in \Pi\}$ and $\{\psi_\pi : \pi \in \Pi\}$ are $\mathbb{P}$-G-C class by condition (C2) and the Theorem 9.26 in Kosorok (2008).

We first show Theorem 6.4(b). For ease of presentation, we define the operator $\mathbb{P}_n$ as the sample average and let

$$\Pi_\lambda = \{\pi \in \Pi \mid \mathbb{E}\psi_\pi(T, X, Y; e, \mu_0, \mu_1) \leq \lambda\},$$
$$\Pi_{n,\lambda} = \{\pi \in \Pi \mid \mathbb{P}_n\psi_\pi(T, X, Y; \hat{e}, \hat{\mu}_0, \hat{\mu}_1) \leq \lambda\},$$

then the estimation error $R(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho)$ can be rewritten as

$$R(\hat{\pi}^*; c, \rho) - R(\pi^*; c, \rho) = D_{1n} + D_{2n} + D_{3n},$$

where

$$D_{1n} = \max_{\pi \in \Pi_\lambda} \mathbb{E}[\varphi_\pi(T, X, Y; e, \mu_0, \mu_1)] - \max_{\pi \in \Pi_\lambda} \mathbb{P}_n\varphi_\pi(T, X, Y; e, \mu_0, \mu_1),$$
$$D_{2n} = \max_{\pi \in \Pi_\lambda} \mathbb{P}_n\varphi_\pi(T, X, Y; e, \mu_0, \mu_1) - \max_{\pi \in \Pi_\lambda} \mathbb{P}_n\varphi_\pi(T, X, Y; \hat{e}, \hat{\mu}_0, \hat{\mu}_1),$$
$$D_{3n} = \max_{\pi \in \Pi_\lambda} \mathbb{P}_n\varphi_\pi(T, X, Y; \hat{e}, \hat{\mu}_0, \hat{\mu}_1) - \max_{\pi \in \Pi_{n,\lambda}} \mathbb{P}_n\varphi_\pi(T, X, Y; \hat{e}, \hat{\mu}_0, \hat{\mu}_1).$$

We will discuss D_{1n}, D_{2n}, and D_{3n} one by one.

$$\begin{aligned} D_{1n} =& \max_{\pi\in\Pi_\lambda} \mathbb{E}[\varphi_\pi(T,X,Y;e,\mu_0,\mu_1)] - \max_{\pi\in\Pi_\lambda} \mathbb{P}_n\varphi_\pi(T,X,Y;e,\mu_0,\mu_1) \\ \leq& \max_{\pi\in\Pi_\lambda} \left|\mathbb{E}[\varphi_\pi(T,X,Y;e,\mu_0,\mu_1)] - \mathbb{P}_n\varphi_\pi(T,X,Y;e,\mu_0,\mu_1)\right| \\ =& o_{\mathbb{P}}(1), \end{aligned}$$

where the last inequality holds from that $\{\varphi_\pi : \pi \in \Pi\}$ is a $\mathbb{P}$-G-C class. Similarly,

$$\begin{aligned} D_{2n} =& \max_{\pi\in\Pi_\lambda} \mathbb{P}_n\varphi_\pi(T,X,Y;e,\mu_0,\mu_1) - \max_{\pi\in\Pi_\lambda} \mathbb{P}_n\varphi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1) \\ \leq& \max_{\pi\in\Pi_\lambda} \left|\mathbb{P}_n\big[\varphi_\pi(T,X,Y;e,\mu_0,\mu_1) - \varphi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1)\big]\right| \\ =& o_{\mathbb{P}}(1), \end{aligned}$$

where the last inequality follows from condition (C1).

Next, we focus on analyzing D_{3n}. For any $\pi \in \Pi$, we consider the difference between $\mathbb{E}\psi_\pi(T,X,Y;e,\mu_0,\mu_1)$ and $\mathbb{P}_n\psi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1)$, which can be reformulated as

$$\begin{aligned} &\mathbb{E}\psi_\pi(T,X,Y;e,\mu_0,\mu_1) - \mathbb{P}_n\psi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1) \\ =& \mathbb{E}\psi_\pi(T,X,Y;e,\mu_0,\mu_1) - \mathbb{P}_n\psi_\pi(T,X,Y;e,\mu_0,\mu_1) + \mathbb{P}_n\psi_\pi(T,X,Y;e,\mu_0,\mu_1) - \mathbb{P}_n\psi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1), \end{aligned}$$

where $\mathbb{E}\psi_\pi(T,X,Y;e,\mu_0,\mu_1) - \mathbb{P}_n\psi_\pi(T,X,Y;e,\mu_0,\mu_1)$ converges to zero uniformly over $\pi \in \Pi$ by noting that $\{\psi_\pi : \pi \in \Pi\}$ is $\mathbb{P}$-G-C class, and $\mathbb{P}_n\psi_\pi(T,X,Y;e,\mu_0,\mu_1) - \mathbb{P}_n\psi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1)$ converges to zero uniformly over $\pi \in \Pi$ by condition (C1). Thus, $\forall\epsilon > 0$, $\exists\, N \in \mathbf{N}$, such that for all $n > N$,

$$\left|\mathbb{E}\psi_\pi(T,X,Y;e,\mu_0,\mu_1) - \mathbb{P}_n\psi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1)\right| < \epsilon,$$

which implies that, for any $\pi \in \Pi_\lambda$, i.e., $\mathbb{E}\psi_\pi(T,X,Y;e,\mu_0,\mu_1) \leq \lambda$, we have

$$\mathbb{P}_n\psi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1) < \lambda + \epsilon,$$

that is, $\frac{\lambda}{\lambda+\epsilon}\pi \in \Pi_{n,\lambda}$, i.e., $\frac{\lambda}{\lambda+\epsilon}\pi \subseteq \Pi_{n,\lambda}$. On the other hand, since $\varphi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1)$ is uniformly bounded with sufficiently large samples according to condition(C1), there exist a positive constant L such that for any π_1 and π_2,

$$|\varphi_{\pi_1}(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1) - \varphi_{\pi_2}(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1)| \leq L \sup_{x\in\mathcal{X}} |\pi_1(x) - \pi_2(x)|.$$

Thus, $\forall\epsilon > 0$, $\exists\, N \in \mathbf{N}$, such that for all $n > N$

$$\begin{aligned} D_{3n} =& \max_{\pi\in\Pi_\lambda} \mathbb{P}_n\varphi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1) - \max_{\pi\in\Pi_{n,\lambda}} \mathbb{P}_n\varphi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1). \\ \leq& \max_{\pi\in\Pi_\lambda} \mathbb{P}_n\varphi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1) - \max_{\pi\in\frac{\lambda}{\lambda+\epsilon}\Pi_\lambda} \mathbb{P}_n\varphi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1). \\ \leq& \frac{\epsilon}{\lambda+\epsilon}L, \end{aligned}$$

Similarly, for the same ϵ, $\exists\, N' \in \mathbf{N}$, such that for all $n > N'$,

$$\begin{aligned} D_{3n} =& \max_{\pi\in\Pi_\lambda} \mathbb{P}_n\varphi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1) - \max_{\pi\in\Pi_{n,\lambda}} \mathbb{P}_n\varphi_\pi(T,X,Y;\hat{e},\hat{\mu}_0,\hat{\mu}_1). \\ \geq& -\frac{\epsilon}{\lambda+\epsilon}L, \end{aligned}$$

which leads to $D_{3n} = o_{\mathbb{P}}(1)$. This completes the proof of Theorem 6.4(b).

Then, we prove Theorem 6.4(a).

$$R(\pi^*;c,\rho) - R(\hat{\pi}^*;c,\rho) = H_{1n} + H_{2n} + H_{3n}$$

where

$$H_{1n} = R(\pi^*; c, \rho) - \hat{R}(\pi^*; c, \rho),$$
$$H_{2n} = \hat{R}(\pi^*; c, \rho) - \hat{R}(\hat{\pi}^*; c, \rho),$$
$$H_{3n} = \hat{R}(\hat{\pi}^*; c, \rho) - R(\hat{\pi}^*; c, \rho).$$

It can be shown that $H_{1n} = o_p(1)$ by a similar argument of D_{1n} and D_{2n}.

$$\begin{aligned} H_{3n} &= \mathbb{P}_n \varphi_{\hat{\pi}}(T, X, Y; \hat{e}, \hat{\mu}_0, \hat{\mu}_1) - \mathbb{E}[\varphi_{\hat{\pi}}(T, X, Y; e, \mu_0, \mu_1)] \\ &= \mathbb{P}_n \varphi_{\hat{\pi}}(T, X, Y; \hat{e}, \hat{\mu}_0, \hat{\mu}_1) - \mathbb{P}_n \varphi_{\hat{\pi}}(T, X, Y; e, \mu_0, \mu_1) + \mathbb{P}_n \varphi_{\hat{\pi}}(T, X, Y; e, \mu_0, \mu_1) - \mathbb{E}[\varphi_{\hat{\pi}}(T, X, Y; e, \mu_0, \mu_1)]. \end{aligned}$$

The condition (C1) implies

$$\mathbb{P}_n \varphi_{\hat{\pi}}(T, X, Y; \hat{e}, \hat{\mu}_0, \hat{\mu}_1) - \mathbb{P}_n \varphi_{\hat{\pi}}(T, X, Y; e, \mu_0, \mu_1) = o_{\mathbb{P}}(1),$$

and the truth that $\{\varphi_\pi : \pi \in \Pi\}$ is a $\mathbb{P}$-G-C class gives that

$$\mathbb{P}_n \varphi_{\hat{\pi}}(T, X, Y; e, \mu_0, \mu_1) - \mathbb{E}[\varphi_{\hat{\pi}}(T, X, Y; e, \mu_0, \mu_1) = o_{\mathbb{P}}(1).$$

Thus, $H_{3n} = o_{\mathbb{P}}(1)$. In addition, by a similar argument of D_{3n}, for any $\pi \in \Pi$ and $\epsilon > 0$, $\exists N' \in \mathbf{N}$, for all $n \geq N'$, $\frac{\lambda}{\lambda+\epsilon}\pi \in \Pi_{\lambda,n}$, then

$$\begin{aligned} H_{2n} &= \hat{R}(\pi^*; c, \rho)) - \hat{R}(\hat{\pi}^*; c, \rho) \\ &= \hat{R}(\pi^*; c, \rho)) - \hat{R}(\frac{\lambda}{\lambda+\epsilon}\pi^*; c, \rho) + \hat{R}(\frac{\lambda}{\lambda+\epsilon}\pi^*; c, \rho) - \hat{R}(\hat{\pi}^*; c, \rho) \\ &\leq \frac{\epsilon}{\lambda+\epsilon} L. \end{aligned}$$

Likewise, for any $\epsilon > 0$, $\exists N \in \mathbf{N}$, for all $n \geq N$, $\frac{\lambda}{\lambda+\epsilon}\hat{\pi}^* \in \Pi_\lambda$, which implies that

$$R(\pi^*; c, \rho) - R(\hat{\pi}^*; c, \rho) \geq R(\frac{\lambda}{\lambda+\epsilon}\hat{\pi}^*; c, \rho) - R(\hat{\pi}^*; c, \rho) \geq -\frac{\epsilon}{\lambda} L.$$

This finishes the proof.

□

C. Estimation of Nuisance Parameters with Sample Splitting

Let K be a small positive integer, and (for simplicity) suppose that $m = n/K$ is also an integer. Let I_1, ..., I_K be a random partition of the index set $I = \{1, ..., n\}$ so that $\#I_k = m$ for $k = 1, ..., K$. Denote I_k^C as the complement of I_k.

Step 1. Nuisance parameter training for each sub-sample.

for $k = 1$ **to** K **do**

(1) Construct estimates $\tilde{e}(x)$, $\tilde{\mu}_1(x)$, and $\tilde{\mu}_0(x)$ using the sample with I_k^C.

(2) Obtain the predicted values of $\tilde{e}(X_i)$, $\tilde{\mu}_1(X_i)$, and $\tilde{\mu}_0(X_i)$ for $i \in I_k$.

end

Step 2. All the predicted values $\tilde{e}(X_i)$, $\tilde{\mu}_1(X_i)$, and $\tilde{\mu}_0(X_i)$ for $i \in I$ consist of the estimates of $e(X_i)$, $\mu_1(X_i)$, and $\mu_0(x_i)$, denoted as $\hat{e}(X_i)$, $\hat{\mu}_1(X_i)$, and $\hat{\mu}_0(X_i)$, respectively.

Remark. (Cross-fitting) the full sample is split into K parts, and the average causal effect is estimated for each subsample while the nuisance parameter training is implemented in the corresponding complement sample. This is the "cross-fitting" approach to machine-learning-aided causal inference advocated by (Chernozhukov et al., 2018), which is prevalent in many recent literature of causal inference (Wager & Athey, 2018; Athey et al., 2019; Semenova & Chernozhukov, 2021).

D. Further Semi-sythetic and Real-world Experiments

Let $e(x) := \mathbb{P}(T = 1|X = x)$, $\mu_t(x) := \mathbb{E}[Y|T = t, X = x]$ for $t = 0, 1$, the estimators of $R(\pi; c, \rho)$ and $u_{\mathrm{FNA}}(\pi)$ are

$$\hat{R}(\pi; c, \rho) = \frac{1}{n}\sum_{i=1}^{n} \varphi_\pi(Z_i; \hat{e}, \hat{\mu}_0, \hat{\mu}_1), \quad \text{and} \quad \hat{u}_{\mathrm{FNA}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \psi_\pi(Z_i; \hat{e}, \hat{\mu}_0, \hat{\mu}_1),$$

where φ_π and ψ_π can be any of the outcome regression (OR), inverse probability weighting (IPW), and augmented inverse probability weighting (AIPW) estimators.

Specifically, the OR estimators are given as

$$\begin{aligned}
\varphi_\pi^{OR}(Z; e, \mu_0, \mu_1) &= (\mu_1(X) - c(X))\,\pi(X) + \rho\mu_0(X)(1 - \pi(X)),\\
\psi_\pi^{OR}(Z; e, \mu_0, \mu_1) &= \mu_0(X)\pi(X) - \mu_1(X)\mu_0(X)\pi(X),
\end{aligned}$$

where $Z = (T, X, Y)$.

The IPW estimators are given as

$$\begin{aligned}
\varphi_\pi^{IPS}(Z; e, \mu_0, \mu_1) &= \left(\frac{TY}{e(X)} - c(X)\right)\pi(X) + \rho\left(\frac{(1-T)Y}{1-e(X)}\right)(1 - \pi(X)),\\
\psi_\pi^{IPS}(Z; e, \mu_0, \mu_1) &= \left(\frac{(1-T)Y}{1-e(X)}\right)\pi(X) - \left(\frac{TY}{e(X)}\right)\mu_0(X)\pi(X).
\end{aligned}$$

The DR estimators are given as

$$\begin{aligned}
\varphi_\pi^{DR}(Z; e, \mu_0, \mu_1) &= \left(\frac{T(Y - \mu_1(X))}{e(X)} + \mu_1(X) - c(X)\right)\pi(X) + \rho\left(\frac{(1-T)(Y - \mu_0(X))}{1-e(X)} + \mu_0(X)\right)(1 - \pi(X)),\\
\psi_\pi^{DR}(Z; e, \mu_0, \mu_1) &= \left(\frac{(1-T)(Y - \mu_0(X))}{1-e(X)} + \mu_0(X)\right)\pi(X) - \left(\frac{T(Y - \mu_1(X))}{e(X)} + \mu_1(X)\right)\mu_0(X)\pi(X).
\end{aligned}$$

In the following, we show more experimental results using OR, IPW, and AIPW estimators on the semi-synthetic dataset IHDP, and the real-world dataset JOBS, in Tables 3, 4, and 5, respectively.

Table 3. Comparison of the Naive method (maximizing estimated rewards), the proposed No-Harm (u) and No-Harm (w) methods in terms of the true reward, welfare change, and true harm on IHDP and JOBS. The CATE-based policy learning and recommendation-based policy learning are employed (with cost functions $c(x) = 0, 0.025, 0.05, 0.075, 0.10$), respectively, where the expected reward and counterfactual harm upper bound are estimated using **outcome regression (OR)** estimators.

IHDP: TRUE HARM $\leq$ 13		CATE-BASED POLICY LEARNING			RECOMMENDATION-BASED POLICY LEARNING		
COST	METHOD	REWARD	ΔWELFARE	TRUE HARM	REWARD	ΔWELFARE	TRUE HARM
$c = 0.00$	NAIVE	569.56 $\pm$ 6.66 $\uparrow$	155.84 $\pm$ 1.44 $\uparrow$	40.26 $\pm$ 8.14 $\uparrow$	550.52 $\pm$ 1.67 $\uparrow$	139.54 $\pm$ 1.04 $\uparrow$	64.62 $\pm$ 0.62 $\uparrow$
	NO-HARM (u)	**456.70 $\pm$ 7.48**	**45.98 $\pm$ 10.27**	**9.32 $\pm$ 2.91**	**144.74 $\pm$ 15.75**	**47.44 $\pm$ 7.73**	**9.24 $\pm$ 3.03**
	NO-HARM (w)	444.63 $\pm$ 8.60	30.88 $\pm$ 8.29	4.42 $\pm$ 1.92	107.24 $\pm$ 16.14	33.24 $\pm$ 7.61	5.06 $\pm$ 2.35
$c = 0.025$	NAIVE	558.57 $\pm$ 7.86 $\uparrow$	157.52 $\pm$ 8.55 $\uparrow$	36.74 $\pm$ 9.62 $\uparrow$	533.40 $\pm$ 1.38 $\uparrow$	139.70 $\pm$ 1.36 $\uparrow$	64.28 $\pm$ 0.83 $\uparrow$
	NO-HARM (u)	**452.72 $\pm$ 6.55**	**44.58 $\pm$ 10.58**	**8.82 $\pm$ 2.97**	**135.42 $\pm$ 14.01**	**45.08 $\pm$ 8.24**	**9.28 $\pm$ 2.42**
	NO-HARM (w)	442.99 $\pm$ 6.88	32.4 $\pm$ 7.43	4.64 $\pm$ 2.20	93.59 $\pm$ 13.51	30.78 $\pm$ 6.67	4.68 $\pm$ 2.37
$c = 0.05$	NAIVE	549.85 $\pm$ 10.90 $\uparrow$	162.80 $\pm$ 10.08 $\uparrow$	28.58 $\pm$ 11.21 $\uparrow$	516.94 $\pm$ 1.70 $\uparrow$	139.50 $\pm$ 1.25 $\uparrow$	64.46 $\pm$ 0.78 $\uparrow$
	NO-HARM (u)	**448.60 $\pm$ 5.77**	**45.38 $\pm$ 8.33**	**7.82 $\pm$ 2.81**	**131.05 $\pm$ 18.30**	**44.94 $\pm$ 7.87**	**9.22 $\pm$ 3.38**
	NO-HARM (w)	440.05 $\pm$ 6.45	32.06 $\pm$ 5.16	2.43 $\pm$ 2.09	92.16 $\pm$ 12.14	33.48 $\pm$ 6.94	4.30 $\pm$ 2.13
$c = 0.075$	NAIVE	546.12 $\pm$ 8.15 $\uparrow$	165.58 $\pm$ 7.89 $\uparrow$	22.96 $\pm$ 8.83 $\uparrow$	500.23 $\pm$ 1.34 $\uparrow$	139.38 $\pm$ 1.23 $\uparrow$	64.54 $\pm$ 0.75 $\uparrow$
	NO-HARM (u)	**445.90 $\pm$ 6.34**	**43.68 $\pm$ 9.18**	**8.94 $\pm$ 3.05**	**125.54 $\pm$ 14.45**	**44.06 $\pm$ 8.99**	**9.16 $\pm$ 2.84**
	NO-HARM (w)	438.25 $\pm$ 7.29	31.44 $\pm$ 7.75	4.66 $\pm$ 2.38	94.34 $\pm$ 19.68	35.10 $\pm$ 8.45	4.52 $\pm$ 1.98
$c = 0.10$	NAIVE	537.80 $\pm$ 8.66 $\uparrow$	165.72 $\pm$ 6.25 $\uparrow$	14.59 $\pm$ 7.63 $\uparrow$	483.20 $\pm$ 1.56 $\uparrow$	139.02 $\pm$ 1.28 $\uparrow$	64.60 $\pm$ 0.56 $\uparrow$
	NO-HARM (u)	**440.68 $\pm$ 7.25**	**44.32 $\pm$ 9.78**	**8.68 $\pm$ 2.69**	**121.68 $\pm$ 16.09**	**43.46 $\pm$ 7.28**	**8.48 $\pm$ 3.08**
	NO-HARM (w)	432.85 $\pm$ 6.86	31.80 $\pm$ 5.76	4.60 $\pm$ 2.23	92.28 $\pm$ 17.61	33.16 $\pm$ 8.39	5.90 $\pm$ 3.28

JOBS: TRUE HARM $\leq$ 50		CATE-BASED POLICY LEARNING			RECOMMENDATION-BASED POLICY LEARNING		
COST	METHOD	REWARD	ΔWELFARE	TRUE HARM	REWARD	ΔWELFARE	TRUE HARM
$c = 0.00$	NAIVE	1991.00 $\pm$ 14.28 $\uparrow$	786.26 $\pm$ 10.50 $\uparrow$	106.57 $\pm$ 3.81 $\uparrow$	1965.57 $\pm$ 1.26 $\uparrow$	758.46 $\pm$ 1.02 $\uparrow$	251.80 $\pm$ 0.40 $\uparrow$
	NO-HARM (u)	**1725.50 $\pm$ 66.29**	**518.53 $\pm$ 65.31**	**37.43 $\pm$ 2.89**	**1032.37 $\pm$ 30.33**	**489.97 $\pm$ 22.53**	**44.47 $\pm$ 3.29**
	NO-HARM (w)	1397.40 $\pm$ 21.07	188.33 $\pm$ 13.81	15.40 $\pm$ 3.39	464.90 $\pm$ 49.61	181.87 $\pm$ 23.24	21.77 $\pm$ 3.62
$c = 0.025$	NAIVE	1936.55 $\pm$ 12.95 $\uparrow$	776.43 $\pm$ 11.65 $\uparrow$	99.96 $\pm$ 4.77 $\uparrow$	1901.49 $\pm$ 1.21 $\uparrow$	758.36 $\pm$ 1.05 $\uparrow$	251.80 $\pm$ 0.47 $\uparrow$
	NO-HARM (u)	**1680.94 $\pm$ 89.24**	**526.00 $\pm$ 59.31**	**37.83 $\pm$ 3.67**	**997.85 $\pm$ 29.31**	**484.80 $\pm$ 24.57**	**44.16 $\pm$ 3.54**
	NO-HARM (w)	1380.61 $\pm$ 10.71	182.50 $\pm$ 21.86	16.06 $\pm$ 4.73	437.67 $\pm$ 47.01	178.23 $\pm$ 16.01	22.00 $\pm$ 5.07
$c = 0.05$	NAIVE	1886.97 $\pm$ 11.13 $\uparrow$	763.76 $\pm$ 8.31 $\uparrow$	92.90 $\pm$ 3.56 $\uparrow$	1837.49 $\pm$ 1.19 $\uparrow$	758.63 $\pm$ 1.05 $\uparrow$	251.73 $\pm$ 0.51 $\uparrow$
	NO-HARM (u)	**1652.53 $\pm$ 52.50**	**486.96 $\pm$ 75.47**	**37.23 $\pm$ 3.28**	**960.03 $\pm$ 33.10**	**480.10 $\pm$ 27.15**	**44.23 $\pm$ 3.76**
	NO-HARM (w)	1372.85 $\pm$ 19.88	183.83 $\pm$ 12.38	15.56 $\pm$ 3.57	449.35 $\pm$ 65.58	184.16 $\pm$ 23.92	23.13 $\pm$ 4.19
$c = 0.075$	NAIVE	1831.76 $\pm$ 8.83 $\uparrow$	750.93 $\pm$ 10.72 $\uparrow$	89.30 $\pm$ 3.33 $\uparrow$	1772.82 $\pm$ 1.08 $\uparrow$	758.50 $\pm$ 0.95 $\uparrow$	251.76 $\pm$ 0.49 $\uparrow$
	NO-HARM (u)	**1632.88 $\pm$ 61.48**	**503.23 $\pm$ 71.29**	**36.00 $\pm$ 4.27**	**935.82 $\pm$ 18.88**	**482.33 $\pm$ 17.79**	**44.03 $\pm$ 3.14**
	NO-HARM (w)	1355.03 $\pm$ 18.64	182.93 $\pm$ 22.55	15.53 $\pm$ 4.00	430.02 $\pm$ 57.74	180.26 $\pm$ 24.04	23.03 $\pm$ 3.65
$c = 0.10$	NAIVE	1780.10 $\pm$ 7.62 $\uparrow$	731.50 $\pm$ 11.70 $\uparrow$	85.30 $\pm$ 16.11 $\uparrow$	1693.56 $\pm$ 6.14 $\uparrow$	753.23 $\pm$ 4.55 $\uparrow$	243.90 $\pm$ 3.87 $\uparrow$
	NO-HARM (u)	**1596.99 $\pm$ 58.06**	**458.06 $\pm$ 85.07**	**36.26 $\pm$ 3.18**	**877.43 $\pm$ 82.77**	**461.66 $\pm$ 53.40**	**43.46 $\pm$ 3.23**
	NO-HARM (w)	1351.57 $\pm$ 18.55	183.90 $\pm$ 12.59	15.20 $\pm$ 3.63	412.10 $\pm$ 54.78	172.13 $\pm$ 20.86	23.33 $\pm$ 4.66

Table 4. Comparison of the Naive method (maximizing estimated rewards), the proposed No-Harm (u) and No-Harm (w) methods in terms of the true reward, welfare change, and true harm on IHDP and JOBS. The CATE-based policy learning and recommendation-based policy learning are employed (with cost functions $c(x) = 0, 0.025, 0.05, 0.075, 0.10$), respectively, where the expected reward and counterfactual harm upper bound are estimated using **inverse probability weighting (IPW)** estimators.

IHDP: TRUE HARM $\leq$ 13		CATE-BASED POLICY LEARNING			RECOMMENDATION-BASED POLICY LEARNING		
COST	METHOD	REWARD	ΔWELFARE	TRUE HARM	REWARD	ΔWELFARE	TRUE HARM
$c = 0.00$	NAIVE	573.56 $\pm$ 6.37 $\uparrow$	157.08 $\pm$ 7.09 $\uparrow$	36.78 $\pm$ 9.22 $\uparrow$	550.72 $\pm$ 1.45 $\uparrow$	139.48 $\pm$ 1.23 $\uparrow$	64.52 $\pm$ 0.64 $\uparrow$
	NO-HARM (u)	**457.26 $\pm$ 7.33**	**45.72 $\pm$ 7.86**	**10.38 $\pm$ 2.97**	**143.88 $\pm$ 14.96**	**44.82 $\pm$ 6.30**	**9.96 $\pm$ 2.69**
	NO-HARM (w)	449.53 $\pm$ 6.73	35.32 $\pm$ 7.45	6.60 $\pm$ 2.44	118.40 $\pm$ 17.20	38.32 $\pm$ 7.77	6.82 $\pm$ 2.96
$c = 0.025$	NAIVE	562.26 $\pm$ 6.94 $\uparrow$	162.52 $\pm$ 7.44 $\uparrow$	29.60 $\pm$ 9.22 $\uparrow$	533.56 $\pm$ 1.63 $\uparrow$	139.23 $\pm$ 1.25 $\uparrow$	64.48 $\pm$ 0.72 $\uparrow$
	NO-HARM (u)	**452.45 $\pm$ 8.15**	**44.06 $\pm$ 7.22**	**9.92 $\pm$ 3.16**	**136.11 $\pm$ 53.14**	**46.75 $\pm$ 18.50**	**7.89 $\pm$ 4.42**
	NO-HARM (w)	444.63 $\pm$ 7.59	36.80 $\pm$ 7.01	6.62 $\pm$ 1.96	93.09 $\pm$ 35.06	32.63 $\pm$ 11.64	3.80 $\pm$ 2.69
$c = 0.05$	NAIVE	551.81 $\pm$ 9.64 $\uparrow$	162.96 $\pm$ 7.84 $\uparrow$	26.72 $\pm$ 9.48 $\uparrow$	516.72 $\pm$ 1.49 $\uparrow$	139.30 $\pm$ 1.19 $\uparrow$	64.53 $\pm$ 0.59 $\uparrow$
	NO-HARM (u)	**449.54 $\pm$ 6.75**	**44.84 $\pm$ 8.13**	**9.62 $\pm$ 3.27**	**130.54 $\pm$ 58.23**	**46.43 $\pm$ 21.28**	**7.81 $\pm$ 5.44**
	NO-HARM (w)	444.11 $\pm$ 6.95	35.40 $\pm$ 7.24	6.36 $\pm$ 2.27	84.16 $\pm$ 30.55	29.96 $\pm$ 11.31	3.78 $\pm$ 2.48
$c = 0.075$	NAIVE	546.41 $\pm$ 7.22 $\uparrow$	167.30 $\pm$ 5.06 $\uparrow$	19.58 $\pm$ 5.00 $\uparrow$	500.13 $\pm$ 1.69 $\uparrow$	139.48 $\pm$ 1.11 $\uparrow$	64.51 $\pm$ 0.72 $\uparrow$
	NO-HARM (u)	**444.35 $\pm$ 7.30**	**47.62 $\pm$ 8.09**	**9.94 $\pm$ 2.70**	**123.41 $\pm$ 49.26**	**45.53 $\pm$ 18.15**	**7.67 $\pm$ 4.02**
	NO-HARM (w)	439.14 $\pm$ 7.38	36.48 $\pm$ 6.07	6.22 $\pm$ 2.19	81.82 $\pm$ 30.43	30.18 $\pm$ 11.39	3.83 $\pm$ 2.70
$c = 0.10$	NAIVE	537.24 $\pm$ 7.21 $\uparrow$	166.22 $\pm$ 4.73 $\uparrow$	18.32 $\pm$ 5.32 $\uparrow$	483.06 $\pm$ 1.57 $\uparrow$	139.18 $\pm$ 1.32 $\uparrow$	64.54 $\pm$ 0.69 $\uparrow$
	NO-HARM (u)	**439.38 $\pm$ 5.47**	**42.92 $\pm$ 8.98**	**9.62 $\pm$ 3.16**	**125.39 $\pm$ 59.57**	**47.29 $\pm$ 21.43**	**7.33 $\pm$ 4.81**
	NO-HARM (w)	435.13 $\pm$ 4.92	37.24 $\pm$ 8.26	6.64 $\pm$ 3.16	74.97 $\pm$ 14.05	29.52 $\pm$ 6.90	3.30 $\pm$ 1.99
JOBS: TRUE HARM $\leq$ 50		CATE-BASED POLICY LEARNING			RECOMMENDATION-BASED POLICY LEARNING		
COST	METHOD	REWARD	ΔWELFARE	TRUE HARM	REWARD	ΔWELFARE	TRUE HARM
$c = 0.00$	NAIVE	1984.20 $\pm$ 13.21 $\uparrow$	786.93 $\pm$ 13.10 $\uparrow$	123.80 $\pm$ 13.29 $\uparrow$	1966.66 $\pm$ 0.47 $\uparrow$	760.00 $\pm$ 0.82 $\uparrow$	251.00 $\pm$ 0.82 $\uparrow$
	NO-HARM (u)	**1511.40 $\pm$ 21.42**	**296.32 $\pm$ 24.15**	**42.56 $\pm$ 5.69**	**570.00 $\pm$ 22.73**	**245.33 $\pm$ 6.80**	**37.33 $\pm$ 9.39**
	NO-HARM (w)	1330.60 $\pm$ 77.69	100.74 $\pm$ 75.55	11.90 $\pm$ 7.87	345.33 $\pm$ 11.84	119.33 $\pm$ 7.36	20.33 $\pm$ 3.09
$c = 0.025$	NAIVE	1932.78 $\pm$ 14.33 $\uparrow$	781.02 $\pm$ 16.96 $\uparrow$	115.46 $\pm$ 7.47 $\uparrow$	1901.81 $\pm$ 1.10 $\uparrow$	759.76 $\pm$ 1.08 $\uparrow$	250.73 $\pm$ 0.89 $\uparrow$
	NO-HARM (u)	**1494.08 $\pm$ 19.43**	**301.46 $\pm$ 23.38**	**40.40 $\pm$ 5.78**	**584.38 $\pm$ 70.30**	**245.56 $\pm$ 29.73**	**41.26 $\pm$ 7.46**
	NO-HARM (w)	1295.49 $\pm$ 62.65	106.65 $\pm$ 71.70	12.13 $\pm$ 7.67	221.85 $\pm$ 126.16	84.4 $\pm$ 53.59	13.36 $\pm$ 8.45
$c = 0.05$	NAIVE	1887.69 $\pm$ 23.29 $\uparrow$	768.43 $\pm$ 13.10 $\uparrow$	27.01 $\pm$ 14.27 $\uparrow$	1837.38 $\pm$ 1.18 $\uparrow$	759.90 $\pm$ 1.79 $\uparrow$	250.26 $\pm$ 1.09 $\uparrow$
	NO-HARM (u)	**1472.92 $\pm$ 25.37**	**311.48 $\pm$ 43.17**	**41.43 $\pm$ 7.98**	**564.00 $\pm$ 80.85**	**243.33 $\pm$ 29.23**	**40.76 $\pm$ 8.08**
	NO-HARM (w)	1254.64 $\pm$ 39.94	68.36 $\pm$ 49.85	8.93 $\pm$ 7.71	162.04 $\pm$ 126.33	55.76 $\pm$ 47.32	10.40 $\pm$ 9.47
$c = 0.075$	NAIVE	1829.04 $\pm$ 12.64 $\uparrow$	755.32 $\pm$ 13.87 $\uparrow$	96.30 $\pm$ 5.42 $\uparrow$	1772.48 $\pm$ 1.55 $\uparrow$	760.50 $\pm$ 2.15 $\uparrow$	249.00 $\pm$ 2.11 $\uparrow$
	NO-HARM (u)	**1463.46 $\pm$ 41.26**	**298.37 $\pm$ 29.81**	**37.33 $\pm$ 5.90**	**524.29 $\pm$ 54.50**	**234.00 $\pm$ 19.90**	**38.53 $\pm$ 5.74**
	NO-HARM (w)	1254.91 $\pm$ 43.66	58.08 $\pm$ 46.51	7.36 $\pm$ 8.53	157.57 $\pm$ 103.43	62.00 $\pm$ 44.58	11.46 $\pm$ 8.46
$c = 0.10$	NAIVE	1788.05 $\pm$ 13.89 $\uparrow$	742.21 $\pm$ 14.64 $\uparrow$	92.90 $\pm$ 17.15 $\uparrow$	1695.79 $\pm$ 5.96 $\uparrow$	752.86 $\pm$ 5.67 $\uparrow$	243.70 $\pm$ 2.86 $\uparrow$
	NO-HARM (u)	**1436.82 $\pm$ 34.44**	**296.32 $\pm$ 35.98**	**37.60 $\pm$ 6.18**	**553.18 $\pm$ 108.61**	**253.83 $\pm$ 45.00**	**40.80 $\pm$ 9.44**
	NO-HARM (w)	1244.65 $\pm$ 19.54	50.88 $\pm$ 38.55	8.40 $\pm$ 8.29	168.94 $\pm$ 114.49	60.13 $\pm$ 44.88	13.03 $\pm$ 10.80

Table 5. Comparison of the Naive method (maximizing estimated rewards), the proposed No-Harm (u) and No-Harm (w) methods in terms of the true reward, welfare change, and true harm on IHDP and JOBS. The CATE-based policy learning and recommendation-based policy learning are employed (with cost functions $c(x) = 0, 0.025, 0.05, 0.075, 0.10$), respectively, where the expected reward and counterfactual harm upper bound are estimated using **augmented inverse probability weighting (AIPW)** estimators.

IHDP: TRUE HARM $\leq$ 13		CATE-BASED POLICY LEARNING			RECOMMENDATION-BASED POLICY LEARNING		
COST	METHOD	REWARD	ΔWELFARE	TRUE HARM	REWARD	ΔWELFARE	TRUE HARM
$c = 0.00$	NAIVE	570.96 ± 3.28 ↑	157.78 ± 4.11 ↑	19.12 ± 2.29 ↑	549.14 ± 1.61 ↑	139.16 ± 1.43 ↑	64.36 ± 0.87 ↑
	NO-HARM (u)	**496.93 ± 11.39**	**83.8 ± 10.42**	**10.34 ± 2.54**	**100.90 ± 15.11**	**43.82 ± 10.37**	**9.60 ± 2.52**
	NO-HARM (w)	459.80 ± 6.86	48.56 ± 6.82	5.98 ± 1.95	73.76 ± 15.62	31.82 ± 7.16	5.42 ± 2.26
$c = 0.025$	NAIVE	561.60 ± 4.19 ↑	157.2 ± 3.77 ↑	18.04 ± 2.25 ↑	532.09 ± 1.84 ↑	139.24 ± 1.47 ↑	64.42 ± 0.66 ↑
	NO-HARM (u)	**494.78 ± 12.52**	**85.04 ± 12.27**	**10.24 ± 2.47**	**102.31 ± 12.87**	**46.58 ± 8.51**	**9.66 ± 3.15**
	NO-HARM (w)	461.50 ± 6.02	48.76 ± 6.38	6.06 ± 1.82	69.39 ± 12.30	32.12 ± 7.24	6.16 ± 2.46
$c = 0.05$	NAIVE	551.62 ± 4.15 ↑	154.4 ± 4.39 ↑	16.76 ± 2.21 ↑	515.30 ± 2.29 ↑	139.48 ± 1.38 ↑	64.16 ± 0.70 ↑
	NO-HARM (u)	**491.33 ± 13.44**	**84.34 ± 13.91**	**9.32 ± 2.83**	**101.42 ± 12.16**	**47.82 ± 9.10**	**9.88 ± 2.61**
	NO-HARM (w)	456.17 ± 6.88	50.50 ± 6.40	6.02 ± 2.01	67.59 ± 13.56	31.46 ± 7.74	5.98 ± 2.94
$c = 0.075$	NAIVE	542.33 ± 3.96 ↑	152.50 ± 4.55 ↑	15.26 ± 2.34 ↑	497.64 ± 2.09 ↑	139.08 ± 1.31 ↑	64.06 ± 0.88 ↑
	NO-HARM (u)	**485.32 ± 14.59**	**85.36 ± 13.78**	**9.42 ± 3.15**	**97.68 ± 20.44**	**46.68 ± 8.32**	**9.28 ± 3.44**
	NO-HARM (w)	454.34 ± 5.26	49.32 ± 5.65	5.64 ± 1.88	68.00 ± 17.56	32.00 ± 7.33	6.18 ± 3.47
$c = 0.10$	NAIVE	534.27 ± 4.21 ↑	148.74 ± 4.10 ↑	14.86 ± 2.21 ↑	480.98 ± 2.63 ↑	139.50 ± 1.96 ↑	63.90 ± 0.96 ↑
	NO-HARM (u)	**482.14 ± 12.73**	**81.68 ± 15.00**	**8.60 ± 3.46**	**92.42 ± 15.17**	**47.34 ± 8.52**	**8.90 ± 2.84**
	NO-HARM (w)	452.29 ± 5.76	49.00 ± 7.03	5.42 ± 1.92	63.33 ± 12.54	31.82 ± 8.05	5.58 ± 2.17
JOBS: TRUE HARM $\leq$ 50		CATE-BASED POLICY LEARNING			RECOMMENDATION-BASED POLICY LEARNING		
COST	METHOD	REWARD	ΔWELFARE	TRUE HARM	REWARD	ΔWELFARE	TRUE HARM
$c = 0.00$	NAIVE	1798.60 ± 7.63 ↑	583.96 ± 10.54 ↑	113.73 ± 4.47 ↑	1965.33 ± 1.44 ↑	758.50 ± 1.52 ↑	251.30 ± 0.69 ↑
	NO-HARM (u)	**1453.00 ± 21.96**	**237.36 ± 29.81**	**43.23 ± 8.06**	**528.00 ± 22.16**	**195.73 ± 13.80**	**41.40 ± 4.85**
	NO-HARM (w)	1325.00 ± 48.62	113.74 ± 60.39	16.80 ± 8.41	197.46 ± 138.66	66.26 ± 52.88	17.16 ± 12.60
$c = 0.025$	NAIVE	1745.02 ± 8.34 ↑	577.09 ± 9.00 ↑	105.03 ± 4.95 ↑	1862.44 ± 10.42 ↑	731.80 ± 8.35 ↑	245.00 ± 3.14 ↑
	NO-HARM (u)	**1444.40 ± 51.75**	**233.02 ± 29.29**	**40.30 ± 4.78**	**532.88 ± 34.84**	**200.40 ± 15.42**	**42.93 ± 6.75**
	NO-HARM (w)	1310.60 ± 50.36	245.94 ± 116.93	14.38 ± 9.27	226.97 ± 117.47	72.10 ± 44.27	20.13 ± 9.48
$c = 0.05$	NAIVE	1701.13 ± 10.41 ↑	566.23 ± 11.23 ↑	93.93 ± 4.68 ↑	1760.50 ± 11.30 ↑	705.26 ± 8.62 ↑	238.23 ± 3.68 ↑
	NO-HARM (u)	**1408.72 ± 27.01**	**242.66 ± 44.18**	**41.13 ± 9.31**	**504.18 ± 25.78**	**195.80 ± 18.89**	**42.86 ± 5.32**
	NO-HARM (w)	1325.56 ± 32.28	118.83 ± 55.50	19.76 ± 9.13	220.94 ± 113.36	77.30 ± 43.96	18.93 ± 8.92
$c = 0.075$	NAIVE	1656.66 ± 11.96 ↑	548.50 ± 11.85 ↑	82.83 ± 6.17 ↑	1658.70 ± 35.11 ↑	678.36 ± 24.10 ↑	229.60 ± 8.60 ↑
	NO-HARM (u)	**1387.93 ± 37.70**	**222.50 ± 25.68**	**36.56 ± 5.64**	**488.11 ± 25.88**	**197.03 ± 17.94**	**42.30 ± 5.44**
	NO-HARM (w)	1306.39 ± 44.39	71.16 ± 60.56	11.80 ± 9.14	202.11 ± 115.71	70.16 ± 47.71	20.73 ± 11.82
$c = 0.10$	NAIVE	1612.20 ± 9.07 ↑	527.06 ± 52.29 ↑	72.66 ± 7.65 ↑	1529.96 ± 49.49 ↑	630.86 ± 34.30 ↑	212.93 ± 11.51 ↑
	NO-HARM (u)	**1362.20 ± 22.95**	**232.63 ± 51.70**	**36.26 ± 8.04**	**475.10 ± 20.52**	**193.83 ± 16.48**	**44.30 ± 5.84**
	NO-HARM (w)	1257.19 ± 39.17	67.83 ± 59.52	11.63 ± 8.91	214.76 ± 179.24	85.33 ± 82.95	22.93 ± 23.32