# Policy Learning under Unobserved Confounding: A Robust and Efficient Approach*

Zequn Jin† Gaoqian Xu‡ Xi Zheng§ Yahong Zhou¶

July 29, 2025

**Abstract**

This paper develops a robust and efficient method for policy learning from observational data in the presence of unobserved confounding, complementing existing instrumental variable (IV) based approaches. We employ the marginal sensitivity model (MSM) to relax the commonly used yet restrictive unconfoundedness assumption by introducing a sensitivity parameter that captures the extent of selection bias induced by unobserved confounders. Building on this framework, we consider two distributionally robust welfare criteria, defined as the worst-case welfare and policy improvement functions, evaluated over an uncertainty set of counterfactual distributions characterized by the MSM. Closed-form expressions for both welfare criteria are derived. Leveraging these identification results, we construct doubly robust scores and estimate the robust policies by maximizing the proposed criteria. Our approach accommodates flexible machine learning methods for estimating nuisance components, even when these converge at moderately slow rate. We establish asymptotic regret bounds for the resulting policies, providing a robust guarantee against the most adversarial confounding scenario. The proposed method is evaluated through extensive simulation studies and empirical applications to the JTPA study and Head Start program.

*JEL codes*: C14, C31, C54

*Keywords*: Partial Identification, Targeted Policy, Treatment Effects, Sensitivity Analysis

*We thank Toru Kitagawa, Bo Zhang, and participants of the 2025 Shanghai Econometrics Workshop for helpful feedback.
†School of Economics, Shanghai University of Finance and Economics. polaris@163.sufe.edu.cn
‡Department of Economics, University of Washington. gx8@uw.edu
§Foster School of Business, University of Washington. xzheng01@uw.edu
¶School of Economics, Shanghai University of Finance and Economics. yahong.zhou@mail.shufe.edu.cn

# 1 Introduction

Policy targeting, allocating interventions based on individual characteristics, has become increasingly influential in applied economics and econometrics. For example, a local government may need to decide which workers receive job training, a school district may identify students most in need of educational support, and a regulatory agency has to determine which workplaces warrant safety inspections.

Policymakers often rely on observational data to guide targeting decisions. However, unobserved confounding, unobserved factors that simultaneously affect both treatment and potential outcomes, can severely bias causal inference from such data. Ignoring these confounders can result in misleading estimates and suboptimal policy decisions. For example, Johnson et al. (2023) note that safety inspections tend to target workplaces with a history of accidents or informal complaints, events that may reflect unobserved traits, such as poor safety culture or weak management, which also affect injury risk. Given the potentially large benefits of well-targeted inspections,[1] designing targeting without accounting for unobserved confounding may misidentify high-risk workplaces, leading to the inefficient use of limited regulatory resources. Similar challenges arise in education, where observational data guide targeting decisions for specific academic programs. However, actual enrollment often hinges on parental choice, which is based on factors inaccessible to policymakers, such as a child's latent abilities, behavioral traits, or home environment. This self-selection process based on unobserved characteristics complicates the design of effective targeting policies.

Many existing policy learning (or targeting) methods using observational data assume no unobserved confounding, that is, treatment assignment is random conditional on the observed covariates; see (Kitagawa and Tetenov, 2018; Athey and Wager, 2021). This assumption, commonly referred to as unconfoundedness, is often unreasonable, as historical policies or decisions may have been based on additional, unobserved information. When this assumption breaks down, researchers often turn to instrumental variable (IV) methods, which enable causal inference and policy targeting; see (Cui and Tchetgen Tchetgen, 2021; Qiu et al., 2021; Sasaki and Ura, 2024). Yet in practice, identifying a valid and credible instrument is often challenging. Moreover, even with a valid binary instrument, heterogeneous treatment effects are typically only partially identified, thereby limiting the ability to effectively target interventions.

As a supplement to IV methods, sensitivity analysis is used to assess the impact of unobserved confounding on policy effects, as originally proposed by Rosenbaum and Rubin (1983). Building on this idea, Kallus and Zhou (2021) incorporate sensitivity analysis into policy learning. Leveraging the marginal sensitivity model (MSM) of Tan (2006), they relax the unconfoundedness assumption and formulate a robust optimization framework to learn policies that remain effective under unobserved confounding. Specifically, building on the framework proposed by Zhao et al. (2019), they construct the objective function as the largest IPW estimator of expected welfare, where the putative propensity scores are restricted by the MSM. While the method offers robustness to unobserved confounding, it is not without limitations. First, as shown by Dorn and Guo (2023), it may yield overly conservative estimates of the upper bounds on expected welfare. Second, the method requires propensity scores to be estimated at the parametric rate, which precludes the use of many popular machine learning methods, such as random forests and LASSO. Third, the policy optimization algorithm is tailored to differentiable parametric policy classes and relies on a computationally intensive iterative routine.

[1] Workplace injuries and illnesses place a significant economic burden on the United States, with annual social costs estimated at $250 billion (Leigh, 2011).

## 1.1 Main Contributions

This work proposes a method for policy learning that is robust to unobserved confounding, serving as a complement to existing IV-based approaches. Building on the MSM framework, we construct a distributionally robust welfare criterion that captures the worst-case policy value over a plausible set of counterfactual outcome distributions. Based on this criterion, we develop a two-stage policy learning procedure: the first stage estimates nuisance components using flexible nonparametric or modern ML methods; the second stage optimizes the estimated criterion over a constrained policy class. The resulting procedure is both computationally efficient and straightforward to implement.

In Section 3, building on Dorn and Guo (2023) and Dorn et al. (2025), we derive closed-form expressions for the worst-case (i.e., sharp lower bounds of) average welfare and welfare improvement[2] compatible with the MSM. Given the identification of these two robust criteria/objectives using moment conditions, we can construct the doubly robust estimation of them, even when the nuisance components are estimated at moderately slow rate, i.e., $o_P(n^{-1/4})$.

Section 4 presents algorithms for learning optimal confounding-robust policies, where the policy class may be constrained by practical considerations such as implementability, cost, and interpretability. Our method follows the mainstream two-stage paradigm of Athey and Wager (2021); Zhou et al. (2023). Unlike Kallus and Zhou (2021), our second-stage policy optimization supports standard machine learning methods and readily available software packages, such as logistic regression, policy trees, and neural networks, to learn the optimal policy by maximizing the estimated criterion. For example, given the doubly robust scores, our method can be implemented using the widely adopted `R`-package `policytree` (Sverdrup et al., 2020). Moreover, Section 4 provides asymptotic upper regret bounds as performance guarantees for our estimated confounding-robust policies.

We validate our method via simulations in Section 5, comparing it to Kallus and Zhou (2021) and the empirical welfare maximization (EWM) of Athey and Wager (2021). We then demonstrate its practical relevance in Section 6 by applying it to two classic empirical settings where unobserved confounding is a primary concern, focusing on the contrast with the EWM.

Our applications suggest that ignoring such confounding may lead to misguided policy recommendations. First, in the National Job Training Partnership Act (JTPA) study, our method refines the naive EWM policy of treating nearly all participants, instead identifying more selective policies that target subgroups with higher education but lower prior earnings as concerns for unobserved confounding grow. In our Head Start application, the EWM policy assigns a mere 1.23% of slots to healthier, higher-income children. In contrast, our robust method prioritizes disadvantaged children with lower family income and maternal education, aligning with the program's mission. These findings demonstrate that accounting for unobserved confounding can fundamentally alter policy targeting, highlighting the importance of robust methods in practice.

## 1.2 Related Literature

Our work contributes to the growing literature on policy learning. Many policy learning works focus on mean-optimal policies under unconfoundedness (Kitagawa and Tetenov, 2018; Athey and Wager, 2021; Zhou et al., 2023). Alternative objectives under the same assumption have also been explored, including quantile-optimal policies (Wang et al., 2018; Leqi and Kennedy, 2021) as well as those motivated by fairness (Kitagawa and Tetenov, 2021; Fang et al., 2023; Viviano and Bradic, 2024; Fan et al., 2025).

To address unobserved confounding, several studies have proposed IV-based approaches for policy targeting. To avoid partial identification of the welfare function, Sasaki and Ura (2024) assume

[2]These are commonly used objective functions in policy learning, which are formally defined in Section 2.2.

the availability of a continuous instrument with sufficiently large support and identify the average welfare via the marginal treatment effect (MTE). A similar approach has been applied to the design of optimal encouragement rules under endogeneity; see (Chen and Xie, 2022; Liu, 2022). In the statistics literature, other works investigate optimal treatment rules or encouragement interventions using binary instruments; see (Cui and Tchetgen Tchetgen, 2021; Qiu et al., 2021). However, these methods rely on quite stringent assumptions for a valid IV to point identify the conditional average treatment effect (CATE).

Binary instruments, though widely used, typically allow identification only of the local average treatment effect (LATE) for the subpopulation of compliers; see (Imbens and Angrist, 1994; Angrist et al., 1996; Athey and Imbens, 2017). This presents major challenges for policy learning: the resulting policy intervention is effective only for compliers, a subpopulation that cannot be identified in advance in a new sample, thus limiting external validity. Moreover, compliance status is instrument-dependent, leading to instability and lack of generalizability across different IV designs. To address this issue, Pu and Zhang (2021) and d'Adamo (2021), propose methods for binary IV settings that partially identify heterogeneous treatment effects, learning robust policies by optimizing policy criteria constructed from IV-identified bounds on the CATE.

A related line of work studies distributionally robust policy learning, where uncertainty sets, often defined via Wasserstein distance or Kullback–Leibler divergence, are used to estimate policies that maximize worst-case expected outcomes (Adjaho and Christensen, 2022; Si et al., 2023; Qi et al., 2023). However, these approaches are often overly conservative and lack interpretability, as they do not explicitly model the source of distributional shift. Lei et al. (2023) address external validity under sample selection bias but do not explicitly account for latent confounding in treatment assignment. Finally, our work also connects to the growing literature on sensitivity analysis, which investigates how violations of unconfoundedness affect causal identification and inference; see (Tan, 2006; Masten and Poirier, 2018; Zhao et al., 2019; Rambachan et al., 2022; Dorn and Guo, 2023; Dorn et al., 2025; Masten et al., 2025).

## 2 Problem Statement and Preliminaries

We consider observational data consisting of random samples $\{Z_i\}_{i=1}^n = \{X_i, Y_i, A_i\}_{i=1}^n$, where $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ represents the observed covariates, $A_i \in \{0,1\}$ denotes a binary intervention/treatment, and $Y_i \in \mathbb{R}$ is the real-valued observed outcome. Within the Neyman-Rubin potential outcomes framework, let $Y_i(0)$ and $Y_i(1)$ denote the potential outcomes under control ($A_i = 0$) and treatment ($A_i = 1$), respectively. The observed outcome satisfies $Y_i = Y_i(A_i)$. Throughout this work, we interpret $Y_i$ as a measure of welfare or utility, where higher values correspond to more desirable outcomes.

Our objective is to use the observational data $\{Z_i\}_{i=1}^n$ to guide personalized policy interventions in settings where unobserved confounding may be present.

**Example 1** (Workplace Inspections)**.** Let $A_i$ indicate whether establishment $i$ is assigned to inspection. Let $Y_i(1)$ and $Y_i(0)$ denote the (negative) number of injuries establishment $i$ would experience under inspection and no inspection, respectively. To improve workplace safety and health, the regulator allocates inspections based on establishment characteristics $X_i$, including historical serious injury counts, number of employees, and total working hours.

**Example 2** (Head Start Enrollment)**.** Let $A_i$ indicate whether child $i$ is enrolled in the Head Start program. Let $Y_i(1)$ and $Y_i(0)$ denote academic outcomes under enrollment and non-enrollment (e.g., test scores or school readiness). To improve early childhood development, the policymaker allocates slots based on characteristics $X_i$, including household income, parental education, and number of siblings.

## 2.1 The Marginal Sensitivity Model

Our analysis builds on the marginal sensitivity model (MSM) introduced by Tan (2006) to control the selection bias caused by unobserved confounders. This model relaxes the unconfoundedness assumption by allowing for unobserved confounders $U \in \mathbb{R}^k$, where $k \in \mathbb{N}^+$ is unknown. Here, we assume that $U$ is a vector but have no prior knowledge of its dimension or other characteristics. We denote $P_o$ as the true distribution of $O \equiv (X, Y(1), Y(0), A, U)$. The true propensity score is given by $e_o(x,u) = \mathbb{P}_{P_o}[A = 1 | X = x, U = u]$, which accounts for both observed covariates and unobserved confounders. In contrast, the nominal propensity score, defined as $e(x) = \mathbb{P}_{P_o}[A = 1 | X = x]$. In the MSM, unobserved confounders are assumed to have a bounded influence on the odds of treatment assignment, restricting the extent of selection bias introduced by unobserved confounders.

We now present the formal specification of the MSM, a widely used framework for sensitivity analysis in the causal inference literature (see, e.g., Zhao et al. (2019); Dorn and Guo (2023); Oprescu et al. (2023); Dorn et al. (2025)).

**Assumption 2.1** (Marginal Sensitivity Model)**.** Suppose there exists a vector of unobserved confounders $U \in \mathbb{R}^k$ such that

$$(Y(1), Y(0)) \perp\!\!\!\perp A \mid (X, U)\,.$$

The distribution of $O \equiv (X, Y(1), Y(0), A, U)$ satisfies the selection bias condition with $1 \leq \Lambda < \infty$ if the following inequality holds $P_o$-almost surely,

$$\frac{1}{\Lambda} \leq \frac{e_o(x,u)/\left(1 - e_o(x,u)\right)}{e(x)/\left(1 - e(x)\right)} \leq \Lambda. \tag{2.1}$$

**Remark 2.1.** In Example 1, the latent variable $U$ may capture administrative information observed by regulators but unavailable to researchers, such as informal complaints, internal risk assessments, or discretionary targeting based on prior interactions. Similarly, in Example 2, $U$ may reflect unobserved characteristics of the child and family, including aspects of socioeconomic background, parental preferences, and investments in early human capital.

In practice, the sensitivity parameter $\Lambda$ in Assumption 2.1 is selected by policymakers based on their prior beliefs regarding the extent of unobserved confounding and the resulting selection bias. Implicitly, it is assumed that policymakers choose $\Lambda$ such that Assumption 2.1 is satisfied. The special case of $\Lambda = 1$ corresponds to the unconfoundedness, also known as the selection-on-observables assumption. Higher values of $\Lambda$ impose fewer restrictions on the degree of unobserved confounding, allowing for greater potential selection bias. Calibration methods can be used to select the sensitivity parameter; see Hsu and Small (2013).

## 2.2 Policy Learning under the MSM

In this subsection, we provide a brief review of policy learning under the assumption of unconfoundedness ($\Lambda = 1$). We then explore how to learn a robust policy when this assumption is violated, within the framework of the marginal sensitivity model. Throughout the rest of this work, we fix the sensitivity parameter $\Lambda \geq 1$.

A possibly randomized policy $\pi$ maps covariates $x \in \mathcal{X}$ to treatment assignment probabilities, with $\pi(x)$ denoting the probability of receiving treatment $A = 1$. The policymaker can pre-specify a policy class $\Pi$, defined as a collection of Borel measurable functions mapping from $\mathcal{X}$ to $[0, 1]$. This class typically incorporates application-specific constraints, including budgetary limitations, structural assumptions, and fairness requirements; see Section 4.2 for concrete examples.

We begin by recalling the framework of policy learning under the unconfoundedness assumption. Given a predefined policy class $\Pi$, the mean-optimal policy learning aims to identify a policy that

maximizes the expected outcome:

$$\pi^\star \in \underset{\pi\in\Pi}{\operatorname{argmax}}\, \mathbb{E}\left[Y\left(\pi\left(X\right)\right)\right], \tag{2.2}$$

where $Y\left(\pi\left(X\right)\right) = \pi(X)Y(1) + (1-\pi(X))Y(0)$. For simplicity, let $W_1 : \pi \mapsto \mathbb{E}\left[Y\left(\pi(X)\right)\right]$ denote the expected welfare function. Moreover, the regret of deploying a policy $\pi \in \Pi$ is defined as

$$\mathrm{Reg}(\pi) = \sup_{\pi'\in\Pi} \mathbb{E}[Y(\pi'(X))] - \mathbb{E}[Y(\pi(X))]. \tag{2.3}$$

It is evident that a policy maximizing $W_1(\pi)$ simultaneously minimizes the regret given in Eq. (2.3).

A central objective in policy learning is policy evaluation, that is, identifying and estimating the criterion function. Once an estimate of $W_1(\cdot)$ is available, the optimal policy can be estimated by maximizing the estimated welfare over the policy class. Unconfoundedness (i.e., $\Lambda = 1$) is a widely adopted assumption for identifying the expected welfare function; see Kitagawa and Tetenov (2018); Athey and Wager (2021); Zhou et al. (2023). Under this assumption, various methods such as inverse probability weighting (IPW) can be employed to identify and estimate $W_1(\pi)$. In particular, the expected welfare function can be expressed as

$$W_1(\pi) = \mathbb{E}[Y(0)] + \mathbb{E}[\tau(X)\pi(X)] \tag{2.4}$$

$$= \mathbb{E}\left[\frac{YA\pi(X)}{e(X)} + \frac{Y(1-A)(1-\pi(X))}{1-e(X)}\right], \tag{2.5}$$

where $e(x)$ is the nominal propensity score, and $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$ is the conditional average treatment effect (CATE). When $\Lambda = 1$, the outcome regression, nominal propensity score, and CATE are all identifiable.

When unconfoundedness fails (i.e., $\Lambda > 1$), the expected welfare $W_1(\pi)$ becomes unidentifiable, and policies based on Eq. (2.4) or Eq. (2.5) lack reliable guarantees. As a result, $W_1(\pi)$ is not a suitable criterion for policy learning in the presence of unobserved confounding.

We adopt a distributionally robust optimization (DRO) framework for policy learning, replacing $W_1(\pi)$ with its worst-case counterpart over distributions consistent with MSM specified in Assumption 2.1. Within this framework, we construct two welfare criteria that are both robust to unobserved confounding and identifiable, and derive the policy by optimizing one of them.

To implement the DRO, we first characterize the distributional uncertainty set over the counterfactual distribution of $(Y(1), Y(0), X, U, A)$ subject to the constraints imposed by Assumption 2.1. Formally, let $\mathcal{P}(\Lambda)$ denote the set of all probability distributions $Q$ on $\mathbb{R}^2 \times \mathcal{X} \times \{0,1\} \times \mathbb{R}^k$ satisfying the following conditions:

(1) If $(X, Y(0), Y(1), A, U) \sim Q$, then $(Y(0), Y(1)) \perp\!\!\!\perp A \mid (X, U)$ under $Q$;

(2) If $Y = AY(1) + (1-A)Y(0)$, then the distribution of $(X, Y, A)$ under $Q$ is identical to the observed-data distribution $P$;

(3) The odds ratio between the true propensity score and the nominal propensity score lies in $[1/\Lambda, \Lambda]$, i.e.,

$$\frac{1}{\Lambda} \le \frac{\mathbb{P}(A=1|X,U)/\mathbb{P}(A=0|X,U)}{\mathbb{P}(A=1|X)/\mathbb{P}(A=0|X)} \le \Lambda, \quad Q\text{-a.s.}$$

We now present two complementary policy learning methods that are robust to unobserved confounding under the MSM framework. Our first approach is the max-min expected welfare method.

Specifically, the *worst-case welfare function* $W_\Lambda : \Pi \to \mathbb{R}$ is defined as

$$W_\Lambda(\pi) = \inf_{Q \in \mathcal{P}(\Lambda)} \mathbb{E}_Q \left[ Y(\pi(X)) \right].$$

The corresponding max-min welfare (MMW) policy is given by

$$\pi_{W,\Lambda} \equiv \pi_W(\cdot; \Lambda) \in \operatorname*{argmax}_{\pi \in \Pi} W_\Lambda(\pi). \tag{2.6}$$

Under the worst-case welfare function $W_\Lambda$, we define the confounding-robust welfare regret (CRW-regret) of a policy $\pi \in \Pi$, relative to the best possible policy in $\Pi$, as

$$\mathrm{Reg}_W(\pi) = \sup_{\pi' \in \Pi} W_\Lambda(\pi') - W_\Lambda(\pi). \tag{2.7}$$

The second approach builds on the conditional average treatment effect (CATE). When $\Lambda = 1$, unconfoundedness holds and, the policy learning problem reduces to maximizing the policy improvement function $\mathbb{E}[\tau(X)\pi(X)]$. To handle the more general case with $\Lambda \geq 1$, we extend this objective to the *worst-case policy improvement function* $\Delta_\Lambda : \Pi \to \mathbb{R}$, defined as

$$\Delta_\Lambda(\pi) = \inf_{Q \in \mathcal{P}(\Lambda)} \mathbb{E}_Q \left[ \pi(X)(Y(1) - Y(0)) \right].$$

The corresponding max-min improvement (MMI) policy is given by

$$\pi_{\Delta,\Lambda} \equiv \pi_\Delta(\cdot; \Lambda) \in \operatorname*{argmax}_{\pi \in \Pi} \Delta_\Lambda(\pi). \tag{2.8}$$

Under the criterion $\Delta_\Lambda$, we define the confounding-robust policy improvement regret (CRI-regret) of a policy $\pi \in \Pi$, relative to the best possible policy in $\Pi$, as

$$\mathrm{Reg}_\Delta(\pi) = \sup_{\pi' \in \Pi} \Delta_\Lambda(\pi') - \Delta_\Lambda(\pi). \tag{2.9}$$

The quantity $\Delta_\Lambda(\pi) = \inf_{Q \in \mathcal{P}(\Lambda)} \mathbb{E}_Q \left[ Y(\pi(X)) - Y(0) \right]$ represents the worst-case welfare gain of policy $\pi$ relative to the baseline policy $\pi_0(x) = 0$. As a result, this approach can be readily generalized to the policy improvement of $\pi$ against any given baseline policy $\pi_0$, with the optimal policy obtained by:

$$\max_{\pi \in \Pi} \inf_{Q \in \mathcal{P}(\Lambda)} \mathbb{E}_Q \left[ Y(\pi(X)) - Y\left(\pi_0(X)\right) \right]. \tag{2.10}$$

**Remark 2.2.** When $\Lambda = 1$, the MMW policy in Eq. (2.8) coincides with the MMI policy in Eq. (2.6). However, in the presence of unobserved confounding (i.e. $\Lambda > 1$), the two policies may differ. A detailed discussion of the differences between these policies is deferred to Section 3.

# 3 Identification

Within our framework, learning a confounding-robust policy requires identifying and estimating either the worst-case welfare function $W_\Lambda(\pi)$ or the worst-case policy improvement function $\Delta_\Lambda(\pi)$. In this section, we formally characterize these criterion functions by leveraging partial identification results for conditional means and CATE under the MSM. We then develop doubly robust/orthogonal moment functions for identifying $W_\Lambda(\pi)$ and $\Delta_\Lambda(\pi)$, which enable the construction of efficient estimators.

For notational simplicity, we define two quantile functions $q^{\pm}_{\Lambda}(x,a)$ as

$$q^{+}_{\Lambda}(x,a) = \inf\left\{q : F(q|x,a) \geq \frac{\Lambda}{1+\Lambda}\right\},$$
$$q^{-}_{\Lambda}(x,a) = \inf\left\{q : F(q|x,a) \geq \frac{1}{1+\Lambda}\right\},$$

where $F(\cdot|x,a)$ denotes the CDF of $Y$ given $X = x$ and $A = a$. Moreover, let $e_a(x) = \mathbb{P}(A = a|X = x)$ for $a \in \{0,1\}$, so that $e_1(x) = e(x)$ and $e_0(x) = 1 - e(x)$.

## 3.1 Identification of Robust Criteria

In this subsection, we identify and characterize the robust criterion functions $W_{\Lambda}(\pi)$ and $\Delta_{\Lambda}(\pi)$. Additionally, we derive the first-best policies optimized under these robust criteria. Let $\Pi_o$ denote the set of all measurable policies:

$$\Pi_o = \{\pi : \mathcal{X} \to [0,1] : \pi \text{ is Borel measurable}\}.$$

A first-best policy refers to a policy that maximizes the criterion function over the unrestricted policy class $\Pi_o$.

We begin by identifying the worst-case welfare $W_{\Lambda}(\pi)$. When $\Lambda > 1$, the true conditional mean function $\mu_o(x,a) = \mathbb{E}_{P_o}[Y(a)|X = x]$ is no longer point identified due to unobserved confounding. Instead, we characterize the sharp bounds for $\mu_o(x,a)$ using Proposition A.1 in Appendix A, which also implies a sharp lower bound for the true welfare function $\mathbb{E}_{P_o}[Y(\pi)]$. Specifically, the proposition provides closed-form expressions for the upper and lower bounds, denoted $\mu^{\pm}_{\Lambda}(x,a)$, which are given by

$$\mu^{\pm}_{\Lambda}(x,a) = \mathbb{E}\left[Y\mathbb{1}\{A = a\}\left[1 + \frac{1 - e_a(X)}{e_a(X)}\Lambda^{\pm \mathrm{sgn}\left(Y - q^{\pm}_{\Lambda}(X,a)\right)}\right]\middle| X = x\right],$$

where $\mathrm{sgn}(t) = 1$ if $t \geq 0$ and $-1$ otherwise.

**Theorem 3.1.** Under Assumption 2.1, for any policy $\pi \in \Pi$, we have

$$W_{\Lambda}(\pi) = \inf_{Q \in \mathcal{P}(\Lambda)} \mathbb{E}_Q\left[Y(\pi(X))\right] = \mathbb{E}\left[\mu^{-}_{\Lambda}(X,1)\pi(X) + \mu^{-}_{\Lambda}(X,0)\left(1 - \pi(X)\right)\right].$$

Moreover, a first-best MMW policy that solves Eq. (2.6) with $\Pi = \Pi_o$ is given by

$$\pi^{\star}_{W,\Lambda}(x) = \mathbb{1}\left\{\mu^{-}_{\Lambda}(x,1) - \mu^{-}_{\Lambda}(x,0) > 0\right\}.$$

The identification of $\Delta_{\Lambda}(\pi)$ is established in Theorem 3.2, which follows as a corollary of Proposition A.2. To formalize the result, define $\tau^{-}_{\Lambda}(x) = \mu^{-}_{\Lambda}(x,1) - \mu^{+}_{\Lambda}(x,0)$.

**Theorem 3.2.** Under Assumption 2.1, for any policy $\pi \in \Pi$, we have

$$\Delta_{\Lambda}(\pi) = \inf_{Q \in \mathcal{P}(\Lambda)} \mathbb{E}_Q\left[(Y(1) - Y(0))\pi(X)\right] = \mathbb{E}\left[\tau^{-}_{\Lambda}(X)\pi(X)\right].$$

Moreover, a first-best MMI policy that solves Eq. (2.8) with $\Pi = \Pi_o$ is given by

$$\pi^{\star}_{\Delta,\Lambda}(x) = \mathbb{1}\{\tau^{-}_{\Lambda}(x) > 0\} = \mathbb{1}\left\{\mu^{-}_{\Lambda}(x,1) - \mu^{+}_{\Lambda}(x,0) > 0\right\}.$$

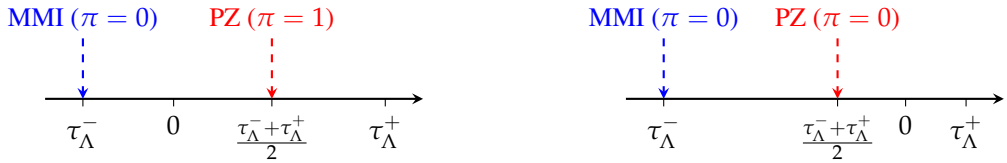
Theorem 3.1 characterizes the first-best MMW policy $\pi^{\star}_{W,\Lambda}(x)$, which assigns treatment by

comparing the lower bounds $\mu^-_\Lambda(x,1)$ and $\mu^-_\Lambda(x,0)$. In contrast, Theorem 3.2 shows that the first-best MMI policy $\pi^\star_{\Delta,\Lambda}(x)$ assigns treatment by comparing $\mu^-_\Lambda(x,1)$ with the upper bound $\mu^+_\Lambda(x,0)$. As a result, the MMI policy is more conservative: it treats only when the worst-case treated outcome exceeds the best-case control outcome. In the context of workplace inspections, the MMI policy $\pi^\star_{\Delta,\Lambda}(x)$ is particularly suitable when regulatory resources are severely limited. Conversely, when resources are abundant, the more expansive MMW policy may be preferable.

We conclude this subsection by comparing our first-best MMW and MMI policies with the Bayes decision rule (referred to as a policy in our framework) of Pu and Zhang (2021), hereafter referred to as PZ-policy.

**Remark 3.1.** The first best MMW policy can also be interpreted through a CATE-based approach, as it assigns treatment by comparing the worst-case potential outcomes. Since this rule is confined to a comparison of identified lower bounds, it is uninformative about the magnitude of the true treatment effect. For this reason, we turn our focus to the MMI policy, and a comparison between the first-best MMI policy and PZ-policy is more instructive.

While both approaches assign treatment based on CATE bounds, they differ in two aspects. First is the construction of the bounds: MMI policy uses the MSM framework, whereas the PZ-policy employs an IV approach. Second, and more importantly, the two policies differ in how they apply decision rules to these bounds. The first-best MMI policy is more conservative and follows a strict criterion: treatment is assigned only when the lower bound of the CATE exceeds zero. In contrast, the PZ-policy adopts a three-case decision rule: it treats if the lower bound is positive, withholds if the upper bound is negative, and, when the bounds straddle zero, treats if the upper bound has a larger absolute value than the lower bound. This contrast highlights why our MMI policy is more conservative: it avoids intervention in the presence of ambiguity, whereas the PZ-policy allows for treatment under uncertainty (see Fig. 1).




(a) When $\tau^-_\Lambda < 0 < \frac{1}{2}(\tau^+_\Lambda + \tau^-_\Lambda)$, the PZ policy recommends treatment ($\pi = 1$), while the MMI policy recommends no treatment ($\pi = 0$).

(b) When $\frac{1}{2}(\tau^+_\Lambda + \tau^-_\Lambda) < 0$, both the PZ and MMI policies recommends no treatment ($\pi = 0$).

Figure 1: Comparison of the MMI and PZ decision rules when applied to a given ambiguous CATE interval $[\tau^-_\Lambda, \tau^+_\Lambda]$.

## 3.2 Doubly/Locally Robust Scores

To effectively learn the confounding-robust policy, it is essential to estimate the entire worst-case welfare function with minimal estimation error. To facilitate the use of modern ML methods, we derive doubly robust scores for the worst-case criteria, $W_\Lambda(\pi)$ and $\Delta_\Lambda(\pi)$, such that the estimation of nuisance parameters has no first-order influence on the resulting policy evaluation. For notational convenience, we may write $\pi(a|x) = a\pi(x) + (1-a)(1-\pi(x))$, which compactly represents the probability that treatment $a \in \{0,1\}$ is assigned under policy $\pi$ given covariate $x$.

**Doubly Robust Score for $W_\Lambda(\pi)$**

We begin by presenting the doubly robust score for estimating $W_\Lambda(\pi)$. As shown in Theorem 3.1, this naturally leads to the following moment condition:

$$\mathbb{E}\left[\mu_\Lambda^-(X,1)\pi(1|X)+\mu_\Lambda^-(X,0)\pi(0|X)\right]-W_\Lambda(\pi)=0.$$

For $t\in\{0,1\}$, let

$$g_t\left(z;e,q_\Lambda^-\right)=y\mathbb{1}\{a=t\}\left[1+\frac{1-e_t(x)}{e_t(x)}\Lambda^{-\text{sgn}\left(y-q_\Lambda^-(x,t)\right)}\right].$$

Using explicit expressions for $\mu_\Lambda^\pm$ given in Proposition A.1, the moment condition above can be rewritten as

$$\mathbb{E}\left[\sum_{t\in\{0,1\}}g_t\left(Z;e,q_\Lambda^-\right)\pi(t|X)\right]-W_\Lambda(\pi)=0. \tag{3.1}$$

Here, $W_\Lambda(\pi)$ is the parameter of interest, while $e(\cdot)$ and $q_\Lambda^-(\cdot,\cdot)$ are two unknown nuisance parameters that can be estimated using ML/nonparametric methods.

A plug-in estimator for $W_\Lambda(\pi)$ can be constructed by substituting estimates into the score function $g_t(Z;\cdot)$ and averaging over the sample, yielding $n^{-1}\sum_{i=1}^n\sum_{t\in\{0,1\}}g_t\left(Z_i;\widehat{e},\widehat{q_\Lambda^-}\right)\pi(t|X_i)$. However, this plug-in estimator is sensitive to estimators of the nominal propensity score and the quantile functions, and may suffer from severe bias. This motivates the doubly robust estimator proposed in this section.

Following Chernozhukov et al. (2018, 2022), we construct the doubly robust score for $W_\Lambda(\pi)$ as follows:

$$\begin{aligned}\phi_t^-\left(z;e,q_\Lambda^-,\rho_{1,\Lambda}^-,\rho_{0,\Lambda}^-\right)&=y\mathbb{1}\{a=t\}\left[1+\frac{1-e_t(x)}{e_t(x)}\Lambda^{-\text{sgn}\left(y-q_\Lambda^-(x,t)\right)}\right]\\&+q_\Lambda^-(x,t)\mathbb{1}\{a=t\}\frac{1-e_t(x)}{e_t(x)}\left(\Lambda-\Lambda^{-1}\right)\left[\frac{1}{1+\Lambda}-\mathbb{1}\left\{y<q_\Lambda^-(x,t)\right\}\right]\\&-\frac{1}{e_t(x)}\left[\Lambda\rho_{1,\Lambda}^-(x,t)+\Lambda^{-1}\rho_{0,\Lambda}^-(x,t)\right]\left[\mathbb{1}\{a=t\}-e_t(x)\right],\end{aligned} \tag{3.2}$$

where

$$\begin{aligned}\rho_{1,\Lambda}^\pm(x,t)&=\mathbb{E}\left[Y\mathbb{1}\left\{Y<q_\Lambda^\pm(X,A)\right\}|X=x,A=t\right],\\\rho_{0,\Lambda}^\pm(x,t)&=\mathbb{E}\left[Y\mathbb{1}\left\{Y>q_\Lambda^\pm(X,A)\right\}|X=x,A=t\right].\end{aligned} \tag{3.3}$$

Let $\eta_{W,\Lambda}=\left(e,q_\Lambda^-,\rho_{1,\Lambda}^-,\rho_{0,\Lambda}^-\right)$ denote the tuple of nauisance functions, the doubly robust score for $W_\Lambda(\pi)$ is given by

$$\psi_W\left(z,\pi;\eta_{W,\Lambda}\right)=\sum_{t\in\{0,1\}}\phi_t^-\left(z;\eta_{W,\Lambda}\right)\pi(t|x).$$

The following Proposition 3.1 establishes the Neyman orthogonality for the moment condition. Moreover, the function $\psi_W(z,\pi;\eta_{W,\Lambda})-W_\Lambda(\pi)$ serves as the efficient influence function for estimating $W_\Lambda(\pi)$.

**Proposition 3.1.** Under Assumption 2.1, for any policy $\pi\in\Pi$, the score function $\psi_W$ satisfies:

(1) $\mathbb{E}\left[\psi_W\left(Z,\pi;\eta_{W,\Lambda}\right)\right]=W_\Lambda(\pi)$.

(2) For any $\widetilde{\eta}_{W,\Lambda}=(\widetilde{e},\widetilde{q_\Lambda^-},\widetilde{\rho_{1,\Lambda}^-},\widetilde{\rho_{0,\Lambda}^-})$ with $\widetilde{e}:\mathcal{X}\to(0,1)$, $\widetilde{q_\Lambda^-}:\mathcal{X}\times\{0,1\}\to\mathbb{R}$, and

$\widetilde{\rho}^{-}_{1,\Lambda}, \widetilde{\rho}^{-}_{0,\Lambda} : \mathcal{X} \times \{0,1\} \to \mathbb{R}$, the pathwise (or Gateaux) derivative satisfies

$$\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\left[\psi_W\left(Z, \pi; \eta_{W,\Lambda} + r\left(\widetilde{\eta}_{W,\Lambda} - \eta_{W,\Lambda}\right)\right)\right]_{r=0} = 0.$$

**Remark 3.2.** The framework of Kallus and Zhou (2021) relies on estimating the nominal propensity score at the parametric rate of $O_P(n^{-1/2})$ to achieve an $n^{-1/2}$ regret bound. However, this rate can only be attained under correct model specification, ruling out the use of machine learning and nonparametric methods. In contrast, our framework allows for the nuisance components to be estimated at a slower rate of $o_P(n^{-1/4})$, enabling the use of data-adaptive, model-agnostic estimation techniques.

**Doubly Robust Score for $\Delta_\Lambda(\pi)$**

We then construct the doubly robust score of $\Delta_\Lambda(\pi)$. Theorem 3.2 implies the moment condition as follows:

$$\mathbb{E}\left[\tau^{-}_{\Lambda}(X)\pi(X)\right] - \Delta_\Lambda(\pi) = 0.$$

Let $\eta_{\Delta,\Lambda} = \left(e, q^{\pm}_{\Lambda}, \rho^{\pm}_{1,\Lambda}, \rho^{\pm}_{0,\Lambda}\right)$ denote the tuple of nuisance parameters. The doubly robust score for $\Delta_\Lambda(\pi)$ can be constructed as

$$\psi_\Delta\left(z, \pi; \eta_{\Delta,\Lambda}\right) = \pi(x)\left[\phi^{-}_1\left(z; e, q^{-}_{\Lambda}, \rho^{-}_{1,\Lambda}, \rho^{-}_{0,\Lambda}\right) - \phi^{+}_0\left(z; e, q^{+}_{\Lambda}, \rho^{+}_{1,\Lambda}, \rho^{+}_{0,\Lambda}\right)\right],$$

where

$$\begin{aligned}\phi^{+}_t\left(z; e, q^{+}_{\Lambda}, \rho^{+}_{1,\Lambda}, \rho^{+}_{0,\Lambda}\right) &= y\mathbb{1}\{a=t\}\left[1 + \frac{1-e_t(x)}{e_t(x)}\Lambda^{\mathrm{sgn}\left(y - q^{+}_{\Lambda}(x,t)\right)}\right] \\ &\quad - q^{+}_{\Lambda}(x,t)\mathbb{1}\{a=t\}\frac{1-e_t(x)}{e_t(x)}\left(\Lambda - \Lambda^{-1}\right)\left[\frac{\Lambda}{1+\Lambda} - \mathbb{1}\left\{y < q^{+}_{\Lambda}(x,t)\right\}\right] \\ &\quad - \frac{1}{e_t(x)}\left[\Lambda^{-1}\rho^{+}_{1,\Lambda}(x,t) + \Lambda\rho^{+}_{0,\Lambda}(x,t)\right]\left[\mathbb{1}\{a=t\} - e_t(x)\right].\end{aligned} \tag{3.4}$$

We summarize several crucial properties of $\psi_\Delta$ in the following proposition.

**Proposition 3.2.** Under Assumption 2.1, for any policy $\pi \in \Pi$, the score function $\psi_\Delta$ satisfies:

(1) $\mathbb{E}\left[\psi_\Delta\left(Z, \pi; \eta_{\Delta,\Lambda}\right)\right] = \Delta_\Lambda(\pi)$.

(2) For any $\widetilde{\eta}_{\Delta,\Lambda} = \left(\widetilde{e}, \widetilde{q}^{\pm}_{\Lambda}, \widetilde{\rho}^{\pm}_{1,\Lambda}, \widetilde{\rho}^{\pm}_{0,\Lambda}\right)$ with $\widetilde{e} : \mathcal{X} \to (0,1)$, $\widetilde{q}^{\pm}_{\Lambda} : \mathcal{X} \times \{0,1\} \to \mathbb{R}$, and $\widetilde{\rho}^{\pm}_{1,\Lambda}, \widetilde{\rho}^{\pm}_{0,\Lambda} : \mathcal{X} \times \{0,1\} \to \mathbb{R}$, the pathwise (or Gateaux) derivative satisfies

$$\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\left[\psi_\Delta\left(Z, \pi; \eta + r\left(\widetilde{\eta}_{\Delta,\Lambda} - \eta_{\Delta,\Lambda}\right)\right)\right]_{r=0} = 0.$$

**Remark 3.3.** The efficient influence functions for $W_\Lambda(\pi)$ and $\Delta_\Lambda(\pi)$ are given by $\psi_W(Z, \pi; \eta_{W,\Lambda}) - W_\Lambda(\pi)$ and $\psi_\Delta(Z, \pi; \eta_{\Delta,\Lambda}) - \Delta_\Lambda(\pi)$, respectively; see Newey (1994).

# 4 Algorithm and Performance Guarantees

In this section, we focus on the algorithms for estimating the MMW and MMI policies, as defined in Section 2.2. Let $\widehat{\pi}_{W,\Lambda}$ and $\widehat{\pi}_{\Delta,\Lambda}$ denote the estimated MMW and MMI policies, respectively.

We also provide theoretical guarantees for these estimated policies in the form of upper bounds on their associated regrets. Specifically, we derive asymptotic upper bounds on the CRW-regret $\mathrm{Reg}_W(\widehat{\pi}_{W,\Lambda})$ and the CRI-regret $\mathrm{Reg}_\Delta(\widehat{\pi}_{\Delta,\Lambda})$, as defined in Eq. (2.7) and Eq. (2.9).

## 4.1 Algorithm

We estimate the MMW and MMI policies using a two-stage procedure: (1) estimation of the robust criterion, either $W_\Lambda(\pi)$ or $\Delta_\Lambda(\pi)$, using $K$-fold cross-fitting; and (2) policy optimization based on the estimated objective.

Given the doubly robust scores in Section 3.2, we first estimate $e(\cdot)$, $q^{\pm}_\Lambda(\cdot,\cdot)$ and $\rho^{\pm}_{t,\Lambda}(\cdot,\cdot)$ for $t \in \{0,1\}$. To this end, we divide the sample into $K$ evenly-sized folds $\cup_{k=1}^K \mathcal{I}_k$. For each fold $k \in [K]$, we estimate $e, q^{\pm}_\Lambda$ and $\rho^{\pm}_{t,\Lambda}$ using the rest $K-1$ folds, and denote the resulting estimates by

$$\widehat{\eta}^{-k}_{W,\Lambda} = \left(\widehat{e}^{-k}, q^{-,-k}_\Lambda, \rho^{-,-k}_{1,\Lambda}, \rho^{-,-k}_{0,\Lambda}\right) \quad \text{and} \quad \widehat{\eta}^{-k}_{\Delta,\Lambda} = \left(\widehat{e}^{-k}, q^{\pm,-k}_\Lambda, \rho^{\pm,-k}_{1,\Lambda}, \rho^{\pm,-k}_{0,\Lambda}\right).$$

Using these estimates, we construct the cross-fitted estimator for $W_\Lambda(\pi)$ and $\Delta_\Lambda(\pi)$ as

$$\widehat{W}_{\Lambda,n}(\pi) = \frac{1}{n}\sum_{k=1}^K \sum_{i\in\mathcal{I}_k} \psi_W\left(Z_i, \pi; \widehat{\eta}^{-k}_{W,\Lambda}\right) \text{ and } \widehat{\Delta}_{\Lambda,n}(\pi) = \frac{1}{n}\sum_{k=1}^K \sum_{i\in\mathcal{I}_k} \psi_\Delta\left(Z_i, \pi; \widehat{\eta}^{-k}_{\Delta,\Lambda}\right). \quad (4.1)$$

The final policy is selected from a pre-specified policy class $\Pi_n$ by maximizing the estimated objective, either $\widehat{W}_{\Lambda,n}(\pi)$ or $\widehat{\Delta}_{\Lambda,n}(\pi)$. The complete procedures for estimating the MMW and MMI policies are summarized in Algorithm 1 and Algorithm 2, respectively.

**Algorithm 1** Max-Min Welfare (MMW) Policy Learning

1: **Input:** Sample $\{Y_i, X_i, A_i\}_{i=1}^n$, sensitivity parameter $\Lambda \geq 1$, and a policy class $\Pi_n$
2: Choose $K \in \mathbb{N}^+$, and partition the sample into $K$ equally sized folds $\cup_{k=1}^K \mathcal{I}_k$
3: **for** each $k = 1, \ldots, K$ **do**
4: Fit estimators for $\eta_{W,\Lambda} = \left(e, q^-_\Lambda, \rho^-_{1,\Lambda}, \rho^-_{0,\Lambda}\right)$ using the rest $K-1$ folds, $\mathcal{I}_k^c \equiv [n]\backslash\mathcal{I}_k$, denoted by $\widehat{\eta}^{-k}_{W,\Lambda}$.
5: **end for**
6: Estimate $W_\Lambda(\pi)$ as

$$\widehat{W}_{\Lambda,n}(\pi) = \frac{1}{n}\sum_{k=1}^K \sum_{i\in\mathcal{I}_k} \psi_W\left(Z_i, \pi; \widehat{\eta}^{-k}_{W,\Lambda}\right).$$

7: Return $\widehat{\pi}_{W,\Lambda} = \mathrm{argmax}_{\pi\in\Pi_n} \widehat{W}_{\Lambda,n}(\pi)$.

**Algorithm 2** Max-Min Improvement (MMI) Policy Learning

1: **Input:** Sample $\{Y_i, X_i, A_i\}_{i=1}^n$, sensitivity parameter $\Lambda \geq 1$, and a policy class $\Pi_n$
2: Choose $K \in \mathbb{N}^+$, and partition the sample into $K$ equally sized folds $\cup_{k=1}^K \mathcal{I}_k$
3: **for** each $k = 1, \ldots, K$ **do**
4: Fit estimators for $\eta_{\Delta,\Lambda} = \left(e, q_\Lambda^\pm, \rho_{1,\Lambda}^\pm, \rho_{0,\Lambda}^\pm\right)$ using the rest $K-1$ folds, $\mathcal{I}_k^c \equiv [n] \backslash \mathcal{I}_k$, denoted by $\widehat{\eta}_{\Delta,\Lambda}^{-k}$.
5: **end for**
6: Estimate $\Delta_\Lambda(\pi)$ as

$$\widehat{\Delta}_{\Lambda,n}(\pi) = \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \psi_\Delta \left(Z_i, \pi; \widehat{\eta}_{\Delta,\Lambda}^{-k}\right). \tag{4.2}$$

7: Return $\widehat{\pi}_{\Delta,\Lambda} = \operatorname{argmax}_{\pi \in \Pi_n} \widehat{\Delta}_{\Lambda,n}(\pi)$.

## 4.2 Assumptions about the Policy Class

In policy design, it is crucial to account for multiple constraints, including budget, simplicity, interpretability, and functional form. Statistically speaking, to achieve regret bounds that decay at the rate of $n^{-1/2}$, one must control the complexity of the policy class. In the remainder of the paper, we allow the policy class $\Pi \equiv \Pi_n$ to vary with $n$. Consequently, we restrict $\Pi_n$ to be VC-subgraph class; see Vaart and Wellner (1996); Giné and Nickl (2021) for further details.

**Assumption 4.1.** The policy class $\Pi_n$ is VC-subrgaph with VC dimension satisfying $\mathrm{VC}(\Pi_n) \leq n^{\zeta_\Pi}$, where $0 < \zeta_\Pi < \frac{1}{2}$.

**Remark 4.1.** In the statistical learning literature, $\mathrm{VC}(\Pi_n)$ is often referred to as the pseudo-dimension, which generalizes the classical Vapnik–Chervonenkis (VC) dimension from binary-valued function classes to real-valued ones; see Anthony and Bartlett (2009); Bartlett et al. (2019). For binary-valued function classes, the VC dimension and pseudo-dimension are identical. In this work, we do not distinguish between the classical VC dimension and the pseudo-dimension, and refer to both simply as the VC dimension.

Various machine learning models can be used as policy classes. Below, we present several examples of such models along with their corresponding VC dimensions, covering both deterministic and randomized policy classes.

**Example 3** (Linear Policies)**.** The class of (deterministic) linear policies is defined as

$$\Pi_n = \left\{\mathbb{1}\{T(x)'\beta > 0\} : \beta \in \mathbb{R}^{d_n}\right\},$$

where $T(x) \in \mathbb{R}^{d_n}$ denotes a vector of transformed covariates constructed from the raw features $X$ using basis expansions such as polynomial terms, B-splines, or interaction terms. The VC dimension of linear policy class $\Pi_n$ is given by $d_n + 1$.

**Example 4** (Decision Trees)**.** A decision tree is a classifier $\pi : \mathcal{X} \to \{0, 1\}$ that recursively partitions the feature space $\mathcal{X}$ into disjoint rectangular regions, assigning a label to each partition. The class of depth-$L$ decision trees in $\mathbb{R}^d$ has VC dimension bounded on the order of $O\left(2^L \log d\right)$.

Next, we provide several examples of randomized policy classes.

**Example 5** (Neural Networks)**.** Deep neural networks have achieved remarkable success in solving complex classification and regression problems. Formally, a neural network defines a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Such networks can be employed to represent both deterministic

and randomized policy classes. Formally, neural networks model the relationship between inputs and outputs through layers of interconnected computational units (neurons), inspired by biological neural systems. The class of networks with $L$ layers, $p$ parameters and a piecewise linear activation function has a VC dimension bounded on the order of $O(Lp\log p)$; see Bartlett et al. (2019).

**Example 6** (Logistic Policies)**.** The class of logistic policies is defined as

$$\Pi_n = \left\{\sigma\left(T(x)'\beta\right) : \beta \in \mathbb{R}^{d_n}\right\},$$

where $\sigma$ is standard logistic function and $T(x)$ denotes a transformation of covariates $x$ as in Example 3. It is noted that $\sigma$ is strictly increasing, and $\Pi_n = \sigma \circ \{T(x)'\beta : \beta \in \mathbb{R}^{d_n}\}$. As a result, Theorem 2.6.18 in Vaart and Wellner (1996) implies $\mathrm{VC}(\Pi_n) = d_n + 1$.

## 4.3 Nuisance Estimators and Uniform Coupling

In this section, we present a key lemma demonstrating that the estimation error of the nuisance parameters becomes asymptotically negligible when the objective functions $W_\Lambda(\cdot)$ and $\Delta_\Lambda(\cdot)$ are estimated using the doubly robust score with cross-fitting.

**Assumption 4.2.** Suppose that $Y$ and $q^\pm_\Lambda(X,a)$, for $a \in \{0,1\}$ have finite second moments; that is, $\mathbb{E}|Y|^2 < \infty$ and $\mathbb{E}\left|q^\pm_\Lambda(X,a)\right|^2 < \infty$.

**Assumption 4.3.** There exist $\kappa \in (0, 1/2)$ such that the nominal propensity score satisfies $e(x) \in (\kappa, 1-\kappa)$ for all $x \in \mathcal{X}$.

Assumption 4.2 is straightforward to verify.Assumption 4.3 is the standard strict overlap condition in the literature, which is essential for establishing the asymptotical regret upper bounds. Moreover, we adopt an agnostic stance on how the nuisance components, $\widehat{e}$, $\widehat{q}_\Lambda$ and $\widehat{\rho}_{t,\Lambda}$, are estimated. Rather than specifying particular estimation procedures, we impose high-level assumptions on their convergence ratio, as stated below.

**Assumption 4.4.** Suppose we have uniformly consistent estimators of the nuisance parameters such that

$$\sup_x \left|\widehat{e}(x) - e(x)\right|, \quad \sup_{x,a} \left|\widehat{q}^\pm_\Lambda(x,a) - q^\pm_\Lambda(x,a)\right|, \quad \sup_{x,a,t} \left|\widehat{\rho}^\pm_{t,\Lambda}(x,a) - \rho^\pm_{t,\Lambda}(x,a)\right| \xrightarrow{P} 0.$$

Furthermore, there exist constants $\zeta_e, \zeta_q, \zeta_\rho \geq 1/2$ and a sequence $b(n) = o(1)$ such that

$$\mathbb{E}\left[\left|\widehat{e}(X_i) - e(X_i)\right|^2\right] \leq \frac{b(n)}{n^{\zeta_e}}, \qquad \mathbb{E}\left[\left|\widehat{q}^\pm_\Lambda(X_i,A_i) - q^\pm_\Lambda(X_i,A_i)\right|^2\right] \leq \frac{b(n)}{n^{\zeta_q}},$$
$$\mathbb{E}\left[\left|\widehat{\rho}^\pm_{t,\Lambda}(X_i,A_i) - \rho^\pm_{t,\Lambda}(X_i,A_i)\right|^2\right] \leq \frac{b(n)}{n^{\zeta_\rho}}.$$

**Remark 4.2.** The regression and quantile functions, $e(x)$ and $q^\pm(x,a)$ can be estimated using kernel and seive methods; see (Li and Racine, 2007; Chen, 2007; Belloni et al., 2019). Recent advances have widely adopted machine learning methods for estimating nuisance parameters, with desirable asymptotic properties well established in the literature; see, e.g., Lasso-based approaches (Belloni and Chernozhukov, 2011; Belloni et al., 2014, 2017) and deep neural network estimators (Schmidt-Hieber, 2020; Farrell et al., 2021; Kohler and Langer, 2021; Padilla et al., 2022). For estimating $\rho^\pm_{t,\Lambda}$, which corresponds to the conditional value at risk (CVaR) or expected shortfall, a variety of methods have been developed, including kernel, local linear, and deep neural network approaches; see (Fissler et al., 2023; Olma, 2021; Yu et al., 2025).

The use of doubly robust scores combined with cross-fitting ensures that errors from nuisance parameter estimation become asymptotically negligible for policy learning, provided $\mathrm{VC}(\Pi_n)$ does not grow too quickly with $n$. To formalize this, suppose the true nuisance functions $\eta_{W,\Lambda}$ and $\eta_{\Delta,\Lambda}$ defined in Section 3.2 are known, the oracle estimators for $W_\Lambda(\pi)$ and $\Delta_\Lambda(\pi)$ are give by

$$W_{\Lambda,n}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \psi_W\left(Z_i, \pi; \eta_{W,\Lambda}\right),$$
$$\Delta_{\Lambda,n}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \psi_\Delta\left(Z_i, \pi; \eta_{\Delta,\Lambda}\right).$$

We conclude this subsection by demonstrating that $\widehat{W}_{\Lambda,n}(\pi)$ is a valid approximation to $W_{\Lambda,n}(\pi)$, and similarly, $\widehat{\Delta}_{\Lambda,n}(\pi)$ approximates $\Delta_{\Lambda,n}(\pi)$, both with a convergence rate faster than $n^{-1/2}$.

**Lemma 4.1.** Under Assumption 4.1, Assumption 4.2, Assumption 4.3 and Assumption 4.4, we have

$$\mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|\widehat{W}_{\Lambda,n}(\pi) - W_{\Lambda,n}(\pi)\right|\right] = o(n^{-1/2}),$$
$$\mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|\widehat{\Delta}_{\Lambda,n}(\pi) - \Delta_{\Lambda,n}(\pi)\right|\right] = o(n^{-1/2}).$$

## 4.4 Asymptotic Regret Bounds

In this subsection, we will demonstrate that, under appropriate bounds on potential hidden confounding, the CRW-regret and CRI-regret of the learned MMW and MMI policies decay at a rate that is upper bounded by $\sqrt{\mathrm{VC}(\Pi_n)/n}$. Recall that the CRW-regret and CRI-regret of the learned optimal policies are defined as

$$\mathrm{Reg}_W\left(\widehat{\pi}_{W,n}\right) = \sup_{\pi\in\Pi_n} W_\Lambda(\pi) - W_\Lambda(\widehat{\pi}_{W,n}),$$
$$\mathrm{Reg}_\Delta\left(\widehat{\pi}_{\Delta,n}\right) = \sup_{\pi\in\Pi_n} \Delta_\Lambda(\pi) - \Delta_\Lambda(\widehat{\pi}_{\Delta,n}),$$

where $\widehat{\pi}_{W,n}$ and $\widehat{\pi}_{\Delta,n}$ are learned from Algorithm 1 and Algorithm 2.

**Theorem 4.1.** Under Assumption 2.1, Assumption 4.1, Assumption 4.2, Assumption 4.3 and Assumption 4.4, then

$$\mathbb{E}\left[\mathrm{Reg}_W\left(\widehat{\pi}_{W,n}\right)\right] = O\left(\sqrt{\mathrm{VC}(\Pi_n)/n}\right), \tag{4.3}$$
$$\mathbb{E}\left[\mathrm{Reg}_\Delta\left(\widehat{\pi}_{\Delta,n}\right)\right] = O\left(\sqrt{\mathrm{VC}(\Pi_n)/n}\right). \tag{4.4}$$

Theorem 4.1 improves upon the result in Kallus and Zhou (2021), where Proposition 6 shows that estimating the nominal propensity score leads to a regret bound that exceeds the oracle regret bound (i.e., the regret bound assuming a known nominal propensity score) by an additional term of the order

$$\frac{1}{n}\sum_{i=1}^{n}\left|1/\widehat{e}(X_i) - 1/e(X_i)\right|.$$

With the application of the doubly robust score and cross-fitting, the regrets of the learned MMW and MMI policies decay at the rate $\sqrt{\mathrm{VC}(\Pi_n)/n}$, and the estimation errors of the nuisance parameters have no asymptotic effect on the regret bound.

**Remark 4.3.** Let us analyze Eq. (4.3) in Theorem 4.1 in detail; the analysis of Eq. (4.4) proceeds in the same fashion. To be more precise, there is a universal constant $K > 0$ not depending on $\Lambda$ such that

$$\limsup_{n\to\infty} \frac{\mathrm{Reg}_W\left(\widehat{\pi}_{W,n}\right)}{\sqrt{\mathrm{VC}(\Pi_n)/n}} \leq K\sigma_W(\Lambda) \tag{4.5}$$

where $\sigma_W(\Lambda)^2 = \mathbb{E}\left|\left(\phi_0^- - \phi_1^-\right)(Z, \eta_{W,\Lambda})\right|^2$, and $\phi_t^-$ for $t \in \{0,1\}$ are defined in Eq. (3.2).

The asymptotic upper bound given in Eq. (4.5) depends on the universal constant $K$ and a variance complexity term $\sigma_W(\Lambda)$ that captures the second moments of $\phi_t^-(Z, \eta_{W,\Lambda})$ for $t \in \{0,1\}$. As shown in the proof of Theorem 4.1, the constant $K$ is universal in the sense that it does not depend on sample size $n$, the sensitivity parameter $\Lambda$, the nuisance function $\eta_{W,\Lambda}$, or even the underlying distribution of $P_o \in \mathcal{P}(\Lambda)$. The variance complexity term $\sigma_W(\Lambda)$ reflects the level of uncertainty induced by the unobserved confounding. When $\Lambda = 1$, then $\sigma_W^2(1)$ coincides with the semiparametric efficient variance for the average treatment effect under the selection-on-observables assumption; see Newey (1994); Robins et al. (1994).

# 5 Simulation Studies

In this section, we conduct simulation studies to evaluate the performance of the MMW and MMI policies, which are learned using Algorithm 1 and 2, respectively. To facilitate comparison with the results in Kallus and Zhou (2021), we use the logistic policies given in Example 6.

## 5.1 Simulation Design

The data-generating process (DGP) is specified as follows. Let $\log \Lambda^* = 1.5$, which implies that the true sensitivity parameter is $\Lambda^* = 4.482$. The observed covariates $X \in \mathbb{R}^2$, treatment assignment $A \in \{0,1\}$, and outcome $Y \in \mathbb{R}$ are generated according to:

$$X \sim \mathcal{N}(\mu_X, I_2), \quad U \mid X \sim \mathrm{Bern}\left(\frac{\Lambda^*}{1+\Lambda^*}e(X) + \frac{1}{1+\Lambda^*}(1-e(X))\right),$$
$$A \mid U, X \sim \mathrm{Bern}\left(\frac{e(X)}{e(X) + \Lambda^{*(1-2U)}(1-e(X))}\right),$$

where $\mu_X = [-1,1]'$ and $I_2 \in \mathbb{R}^{2\times 2}$ is the identity matrix. The nominal propensity score $e(X)$ is defined as:

$$e(X) = \mathbb{P}(A = 1 \mid X) = \sigma\left(\zeta(X)'\theta\right), \quad \text{with} \quad \sigma(z) = \frac{1}{1+e^{-z}},$$

where the parameter vector $\theta$ and the nonlinear feature map $\zeta(X)$ are given by

$$\theta = [0.2, 0.4, 0.1, -0.1, 0.5, -0.5], \quad \zeta(X) = \left[\max(x_1, 0),\ \frac{x_1 x_2^2}{10},\ \sin(x_2^2),\ x_1,\ x_2,\ 1\right]'.$$

It can be readily verified that $U$ perturbs the nominal odds ratio $e(X)/(1-e(X))$ by a multiplicative factor of $\Lambda^{*\pm 1}$, ensuring the DGP satisfies the MSM with $\Lambda^* = 4.482$. Finally, the potential outcome is generated as:

$$Y(A) = \beta_{\text{cons}} + \beta_A A + X'\beta_X + (A \cdot X)'\beta_{X,A} + \beta_U U + \epsilon, \quad \epsilon \sim \mathcal{N}(0,1),$$

where:

$$\beta_{\text{cons}} = -0.2, \quad \beta_A = -0.1, \quad \beta_X = [1, -1]', \quad \beta_{X,A} = [0.2, 0.4]', \quad \beta_U = 1.5.$$

We generate i.i.d. samples $\{X_i, A_i, U_i, Y_i(1), Y_i(0)\}_{i=1}^n$ according to the DGP described above, with $n = 2{,}000$. The observed data $\{X_i, A_i, Y_i\}_{i=1}^n$ are used to learn the policies. The conditional quantile functions $q_{\Lambda^*}^{\pm}$ are estimated using gradient boosted trees and the nominal propensity score $e$ and CVaR $\rho_{t,\Lambda^*}^{\pm}$ are estimated using random forests.

We evaluate the performance of estimated policy $\widehat{\pi}$ using three metrics: (i) the expected welfare, defined in Eq. (2.2); (ii) the worst-case welfare $W_\Lambda(\widehat{\pi})$, as given in Theorem 3.1; and (iii) the worst-case policy improvement $\Delta_\Lambda(\widehat{\pi})$ given in Theorem 3.2. All three metrics are estimated using sample averages computed on an independent out-of-sample dataset of 100,000 randomly drawn observations. To assess the stability of our method, we evaluate each metric over 100 repeated experiments. We report the mean from these repetitions and visualize the variability using shaded bands that represent the 95% confidence bands, calculated from the standard deviation across the 100 experiments.

## 5.2 Simulation Analysis

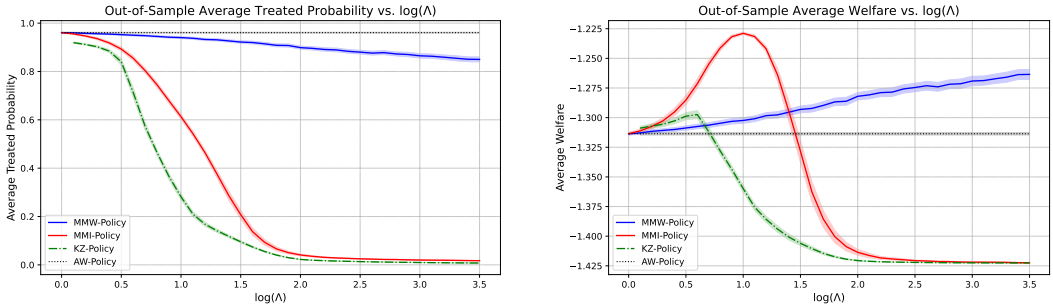
We report the performance of the MMW and MMI policies, as well as the robust IPW-based policy of Kallus and Zhou (2021) (KZ for short). As a benchmark, we include the policy proposed by Athey and Wager (2021) (AW), which assumes no unobserved confounding $(\Lambda = 1)$. To evaluate the MMW, MMI, and KZ policies, we conduct a sensitivity analysis by varying the sensitivity parameter over $\log \Lambda \in \{0.1, 0.2, \ldots, 3.5\}$. This assesses their robustness to miscalibration relative to the true level of unobserved confounding in the true DGP with $\log \Lambda^* = 1.5$. By design, the MMW and MMI policies reduce to the AW policy at the $\Lambda = 1$ setting. In contrast, the KZ policy remains different from AW even when $\Lambda = 1$, as its objective function does not include a bias-correction term for the estimated propensity scores.

Fig. 2a shows how each policy adjusts its treatment probability in response to the assumed level of unobserved confounding, and helps explain their relative performance in terms of expected welfare, presented in Fig. 2b. The AW policy, assuming no unobserved confounding($\Lambda = 1$), yields an aggressive strategy, assigning treatment to 95% individuals. In contrast, MMI, MMW and KZ are sensitivity-aware, adopting a conservative approach that is responsive to $\Lambda$. As $\log \Lambda$ increases, they systematically reduce treatment rates to hedge against selection bias. However, their conservative dynamics differ significantly. The KZ policy is overly conservative, exhibiting a precipitous drop in assignments due to its reliance on the loose lower bound on policy improvement from Zhao et al. (2019). In contrast, the MMI policy is precisely conservative; by precisely identifying heterogeneous treatment effects, it establishes a sharper bound on policy improvement, leading to a more calibrated and gradual reduction in treatment. Finally, the MMW policy's aggressiveness under high uncertainty is a direct consequence of its decision rule, illustrated in Section 3.2, which favors treatment whenever $\mu_\Lambda^-(x, 1)$ exceeds $\mu_\Lambda^-(x, 0)$.

Fig. 2b shows the resulting expected welfare. The MMI policy achieves its peak expected welfare at $\log \Lambda = 1$. It outperforms the MMW policy for assumed confounding levels below the true value of $\log \Lambda^* = 1.5$. In regimes of greater assumed uncertainty $\log \Lambda > 1.5$, however, the MMW policy's performance is superior, yielding significantly higher welfare. Their performance crossover occurs precisely at this true value, confirming their complementary nature. This divergence occurs because, in high-uncertainty regimes, the MMI policy becomes overly cautious, whereas the MMW policy adopts a more risk-tolerant strategy.

Finally, Fig. 3a and 3b validate that the observed behaviors are consequences of each policy's specific design. As expected, Fig. 3a shows the MMW policy achieving the highest worst-case

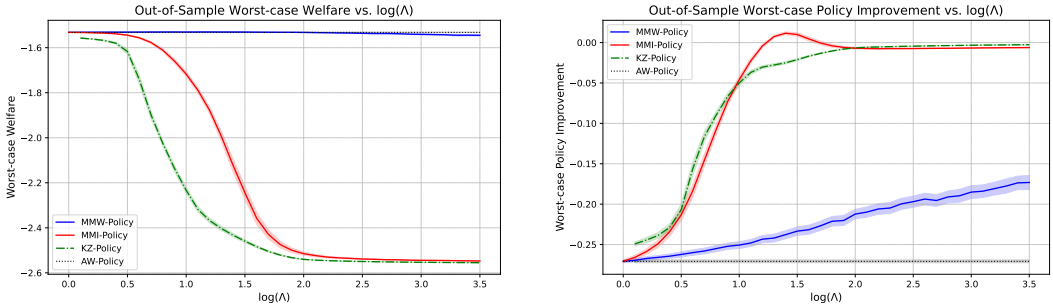
welfare, and Fig. 3b confirms the MMI policy's success in maximizing the worst-case policy improvement. This confirms that each policy effectively optimizes its intended objective, leading to the distinct and complementary performance profiles we identified.




(a) Average Treatment Probability

(b) Expected Welfare

Figure 2: Average treatment probability and expected welfare as a function of the sensitivity parameter $\log(\Lambda)$. The MMI and MMW policies are compared against the KZ and AW benchmarks. Shaded areas are 95% confidence bands.




(a) Worst-case Welfare

(b) Worst-case Policy Improvement

Figure 3: Performance on Intended Objectives. The figures plot (a) the MMW policy's objective of worst-case welfare and (b) the MMI policy's objective of worst-case policy improvement against $\log(\Lambda)$. Shaded areas denote 95% confidence bands.

# 6 Empirical Application

In this section, we apply our method to two empirical applications. Section 6.1 revisits the JTPA study and examines assignment to job training programs, while Section 6.2 explores enrollment decisions in the Head Start program. In both settings, unobserved confounders likely give rise to endogenous selection.

## 6.1 The JTPA Study

We apply the MMW and MMI policy learning methods to the National Job Training Partnership Act (JTPA) Study, a large-scale randomized controlled trial evaluating training services for disadvantaged adults (Bloom et al., 1997). In the experiment, eligibility for services was randomly assigned, and participants' earnings were tracked over the 30-month period after the assignment. This dataset was notably used by Kitagawa and Tetenov (2018) to estimate an intent-to-treat optimal policy for

9,223 applicants using their Empirical Welfare Maximization (EWM) approach, based on covariates such as education and prior earnings.

While eligibility was randomly assigned, compliance in the JTPA study was imperfect: approximately 23% of individuals did not adhere to their assigned eligibility status, as shown in Table 1. This imperfect compliance introduces self-selection into actual participation decisions, raising concerns about unobserved confounding. For instance, eligible individuals who opt into the program may exhibit stronger job search motivation or face fewer personal constraints, such as childcare responsibilities, health issues, or transportation barriers, than those who decline participation. These unobserved factors likely influence both program participation and post-program earnings, violating the unconfoundedness assumption required by conventional policy learning methods.

Table 1: Joint distribution of eligibility and participation, JTPA study

| | Eligibility (IV) | | |
|---|---|---|---|
| Participation ($A_i$) | 0 | 1 | Total |
| 0 | 3047 | 2118 | 5165 |
| 1 | 43 | 4015 | 4058 |
| Total | 3090 | 6133 | 9233 |

Data source: Kitagawa and Tetenov (2018) and Abadie et al. (2002).

We focus on actual program participation, as the policy-relevant objective is to target individuals who are most likely to benefit from job training and to allocate participation accordingly. Building on the perspective of d'Adamo (2021), who frame policy learning as deriving optimal treatment rules under potential unobserved confounding, we apply the MMW and MMI methods, both of which are explicitly designed to address such confounding. Unlike d'Adamo (2021), who use a binary instrumental variable (eligibility) to construct partial identification intervals for the CATE, we adopt a complementary strategy by employing with a pre-specified sensitivity parameter $\Lambda$.

We estimate the optimal MMW and MMI policies using Algorithms 1 and 2, with $K = 10$ folds for cross-fitting. The conditional quantile functions $q^{\pm}_{\Lambda^*}$ are estimated using gradient boosted trees, and the nominal propensity score $e$ and CVaR $\rho^{\pm}_{t,\Lambda^*}$ are estimated using random forests. To account for program costs, we subtract \$1,216 from the outcomes of treated individuals. This amount corresponds to the average cost of services per actual participant, as reported in Table 5 of Bloom et al. (1997). Following the approach of Kitagawa and Tetenov (2018) and d'Adamo (2021), we consider the class of quadrant treatment policies due to its simplicity and interpretability:

$$\Pi \equiv \{\mathbb{1}\{s_1(\text{edu} - t_1) > 0,\ s_2(\text{earnings} - t_2) > 0\} : s_1, s_2 \in \{-1, 1\},\ t_1, t_2 \in \mathbb{R}\}.$$

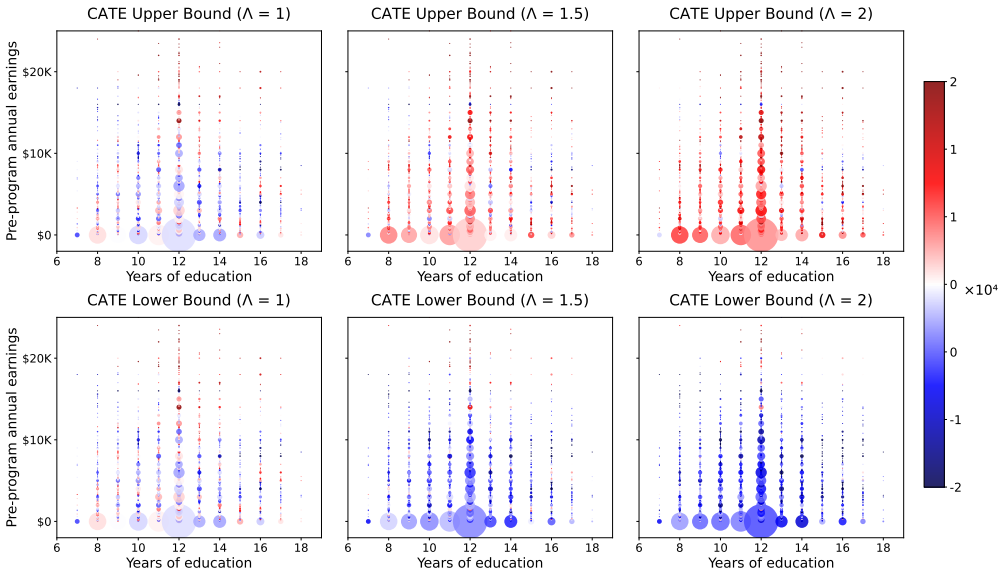



Figure 4: Upper and lower bounds of the CATE under uncertainty levels $\Lambda = 1, 1.5, 2$.

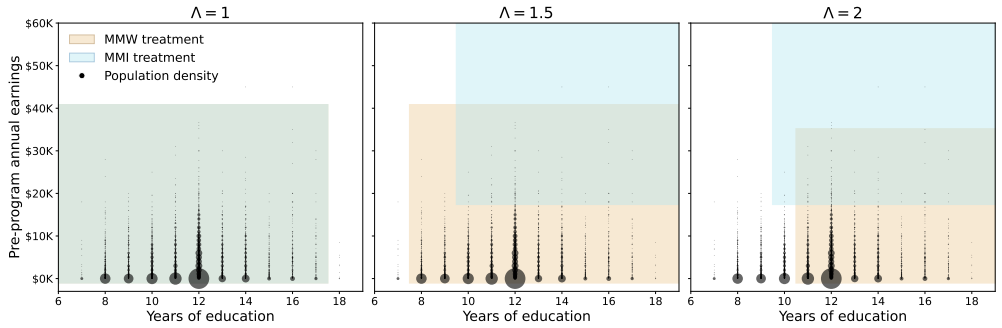



Figure 5: Optimal MMW and MMI quadrant policies under different uncertainty levels: $\Lambda = 1, 1.5$, 2. As $\Lambda$ increases, both policies become more conservative; however, MMI policies reduce treatment sharply, whereas MMW policies decrease treatment more gradually.

Figure 4 presents the estimated upper and lower bounds of the CATE under the MSM framework, evaluated across different values of the sensitivity parameter $\Lambda \in \{1, 1.5, 2\}$. When $\Lambda = 1$, the estimated upper and lower bounds coincide. Choosing a larger value of $\Lambda$ can yield more robust policy targeting by ensuring that the identified set contains the true causal effect. However, this robustness comes at the cost of producing less informative estimates and more conservative policy intervention. Therefore, we recommend that practitioners examine results across a range of $\Lambda$ values to assess the sensitivity of their conclusions to unobserved confounding.

Figure 5 illustrates the estimated optimal quadrant policies for MMW and MMI. When $\Lambda = 1$, both the MMW and MMI policies is exactly the AW policy, but with a treated fraction approaching one. As sensitivity parameter $\Lambda$ increases, the MMI policy becomes increasingly conservative, with the treatment rate dropping rapidly to near zero when $\Lambda \geq 1.5$. In contrast, the MMW policy shows a more gradual decline in the treatment ratio, which remains relatively high at 0.75 when $\Lambda = 2$, and drops sharply to around 0.05 for $\Lambda \geq 2.5$. The extremely low treatment rates under the MMI policy arise because the estimated lower bounds of the CATE fall below zero for nearly all individuals when

$\Lambda \geq 1.5$. Consequently, our subsequent analysis focuses on the MMW policy. When $\Lambda = 1$, the quadrant policy targets individuals with fewer than 17.5 years of education and pre-program earnings below \$40,800. In contrast, at $\Lambda = 2$, the MMW quadrant policy assigns treatment to those with at least 10.5 years of education and earnings below \$35,144. As the sensitivity parameter $\Lambda$ increases, the estimated MMW policy becomes more selective, prioritizing individuals with higher educational level and lower earnings.

## 6.2 The Head Start Program

We further demonstrate our confounding-robust policy learning method by applying it to enrollment decisions for the Head Start program. The objective is to develop a policy that improves academic outcomes by targeting admission to children most likely to benefit. This application presents a multi-valued treatment setting with three alternatives, Head Start, other preschools, or no preschool, necessitating the use of our MMW approach extended for multiple treatments as detailed in Appendix B.

Head Start is a federal matching grant initiative designed to foster the cognitive, social, and health development of children from low-income families, with the aim of preparing them to enter school on a more equal footing with their more advantaged peers. A substantial body of research has evaluated the impact of Head Start participation on outcomes such as academic achievement, physical health, and long-term socioeconomic status; see, e.g., (Currie and Thomas, 1995; Deming, 2009; Ludwig and Miller, 2007; Walters, 2015).

A key challenge in evaluating Head Start's impact is that enrollment is not randomly assigned. First, families self-select into Head Start or other preschool programs based on eligibility and perceived benefits. Those who value early education may also make unobserved investments in their children, such as fostering enriched home environments or providing supplemental learning. Second, limited program capacity in many areas gives administrators discretion over admissions, leading to selection based on both observed and unobserved characteristics, some of which also affect child outcomes. As a result, failing to account for such confounding can lead to targeting policies that misidentify the children most likely to benefit.

We draw on data from the National Longitudinal Survey of Youth (NLSY) and its child supplement, the National Longitudinal Survey of Youth 1979 Children and Young Adults (NLSCYA). The NLSY began in 1979 with a nationally representative sample of young women, and the NLSCYA has tracked their children since 1986. Our analysis sample is constructed by linking child-level outcomes from the NLSCYA with maternal background characteristics from the original NLSY cohort.

We use the child's standardized score on the Peabody Picture Vocabulary Test (PPVT), a widely used measure of early verbal ability, as a proxy for our policy objective of improving academic achievement. The pre-treatment covariates used in our analysis capture a range of child, maternal, and household characteristics. These include: (1) the percentile rank of average net family income between ages 0 and 4 (*Income Pctl*); (2) child's gender; (3) birth weight (*Birth Wt*); (4) weight at preschool entry (*Preschool Entry Wt*); (5) firstborn status; (6) mother's highest completed grade; (7) mother's AFQT percentile score (*Mother AFQT*); (8) number of the mother's biological siblings; (9–10) number of household members with less than 12 years and at least 16 years of education; and (11) race (White, Hispanic, or Black). After excluding observations with missing data, the final sample consists of 3,826 children: 755 attended Head Start, 2,020 enrolled in other preschool programs, and 1,051 did not attend any preschool.

While the full set of pre-treatment covariates is used to estimate doubly robust scores, we restrict the variables used for policy optimization to a small, interpretable subset: percentile rank of family income, mother's AFQT score, birth weight, and weight at preschool entry. These are selected based on fairness and ethical considerations. To ensure interpretability, we restrict the policy class

to depth-2 decision trees. Figure 6 shows the estimated policies under four values of the sensitivity parameter $\Lambda = 1, 2, 3, 4$.

The policy tree learned under $\Lambda = 1$ serves as our benchmark and exactly replicates the policy obtained by approaches such as (Athey and Wager, 2021; Zhou et al., 2023), which do not account for unobserved confounding. An analysis of this benchmark policy reveals several features that present challenges for direct implementation. First, the policy assigns Head Start to a very limited subgroup, only 1.23% of the population. Second, it prioritizes children from higher-income families (albeit with low preschool-entry weight), which deviates from the program's original mission of serving economically disadvantaged household. At the same time, the policy assigns the lowest-income children to either other preschools or no preschool at all. This motivates our subsequent sensitivity analysis to assess how the policy recommendations change when accounting for unobserved confounding.

As the sensitivity parameter $\Lambda$ increases, the optimal policy undergoes a clear evolution. Initially, at $\Lambda = 2$, the policy is still primarily driven by income, but becomes more stringent by reducing the Head Start allocation to just 0.78% of the sample. A more dramatic structural break occurs at $\Lambda = 3$, where the model abandons income-based logic in favor of health indicators, such as birth weight. The policy under $\Lambda = 4$ advocates for a targeted intervention for a clearly defined disadvantaged group while adopting a conservative, non-interventionist stance for the majority.

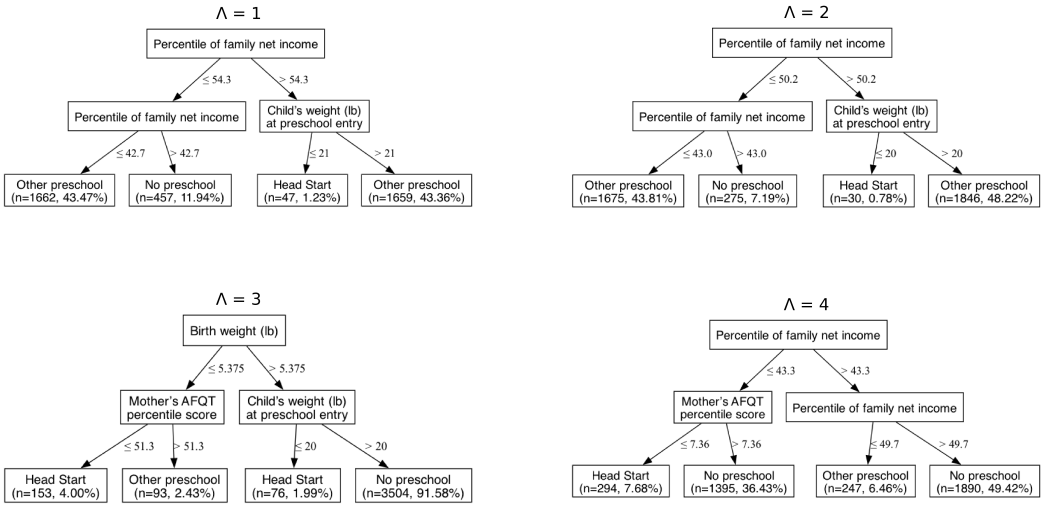



Figure 6: This figure presents a sensitivity analysis of the robust policy under varying assumptions about unobserved confounding. As $\Lambda$ increase, the primary decision criteria shift from economic factors to health indicators, and finally to maternal AFQT scores.

# 7 Concluding Remarks

Formulating effective policies from observational data is challenged by unobserved confounding, which can lead to biased and potentially harmful policy intervention. This paper proposes a robust and efficient framework for policy learning that explicitly accounts for such endogeneity. Our contributions are twofold: first, we derive sharp, closed-form identification results for robust welfare criteria under the Marginal Sensitivity Model; second, we develop doubly robust scores for their estimation. This innovation enables the use of flexible machine learning methods for nuisance estimation within our framework. Building on these components, we establish regret bounds for the

resulting confounding-robust policies.

Our work also offers insights into the practical challenges of applying policy learning methods. While powerful, ML-based approaches that optimize a single objective can yield policies that are difficult to trust and interpret, particularly when built on fragile assumptions. Our work underscores that sensitivity analysis is a critical step for examining a policy's foundational reliability.

Finally, several directions remain open for future research. First, while our framework primarily focuses on binary treatments, extending it to multivalued or continuous treatments is a promising direction that would broaden its applicability and introduce new methodological challenges. Second, the choice of the sensitivity parameter $\Lambda$ remains an important yet unresolved issue. Although we suggest that practitioners may specify $\Lambda$ based on domain knowledge, this approach is inherently ad hoc. While Hsu and Small (2013) offers some guidance, the development of systematic, data-driven methods for selecting $\Lambda$ remains a key avenue for future investigation.

# References


Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117.

Adjaho, C. and Christensen, T. (2022). Externally valid policy choice. *arXiv preprint arXiv:2205.05561*.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

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

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17.

Belloni, A. and Chernozhukov, V. (2011). $\ell$1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Fernández-Val, I. (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(1):4–29.

Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50.

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W., and Bos, J. M. (1997). The benefits and costs of jtpa title ii-a programs: Key findings from the national job training partnership act study. *Journal of Human Resources*, pages 549–576.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632.

Chen, Y.-C. and Xie, H. (2022). Personalized subsidy rules. *arXiv preprint arXiv:2202.13545*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535.

Cui, Y. and Tchetgen Tchetgen, E. (2021). A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *Journal of the American Statistical Association*, 116(533):162–173.

Currie, J. and Thomas, D. (1995). Does head start make a difference? *American Economic Review*, 85(3):341–364.

d'Adamo, R. (2021). Orthogonal policy learning under ambiguity. *arXiv preprint arXiv:2111.10904*.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from head start. *American Economic Journal: Applied Economics*, 1(3):111–134.

Dorn, J. and Guo, K. (2023). Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544):2645–2657.

Dorn, J., Guo, K., and Kallus, N. (2025). Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *Journal of the American Statistical Association*, 120(549):331–342.

Fan, Y., Qi, Y., and Xu, G. (2025). Policy learning with $\alpha$-expected welfare. *arXiv preprint arXiv:2505.00256*.

Fang, E. X., Wang, Z., and Wang, L. (2023). Fairness-oriented learning for optimal individualized treatment rules. *Journal of the American Statistical Association*, 118(543):1733–1746.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Fissler, T., Merz, M., and Wüthrich, M. V. (2023). Deep quantile and deep composite triplet regression. *Insurance: Mathematics and Economics*, 109:94–112.

Giné, E. and Nickl, R. (2021). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press.

Hsu, J. Y. and Small, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics*, 69(4):803–811.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Johnson, M. S., Levine, D. I., and Toffel, M. W. (2023). Improving regulatory effectiveness through better targeting: Evidence from osha. *American Economic Journal: Applied Economics*, 15(4):30–67.

Kallus, N. and Zhou, A. (2021). Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616. Supplementary materials and online appendix available at https://doi.org/10.3982/ECTA13288.

Kitagawa, T. and Tetenov, A. (2021). Equality-minded treatment choice. *Journal of Business & Economic Statistics*, 39(2):561–574.

Kohler, M. and Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249.

Lei, L., Sahoo, R., and Wager, S. (2023). Policy learning under biased sample selection. *arXiv preprint arXiv:2304.11735*.

Leigh, J. P. (2011). Economic burden of occupational injury and illness in the united states. *The Milbank Quarterly*, 89(4):728–772.

Leqi, L. and Kennedy, E. H. (2021). Median optimal treatment regimes. *arXiv preprint arXiv:2103.01802*.

Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Liu, Y. (2022). Policy learning under endogeneity using instrumental variables. *arXiv preprint arXiv:2206.09883*.

Ludwig, J. and Miller, D. L. (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, 122(1):159–208.

Masten, M. A. and Poirier, A. (2018). Identification of treatment effects under conditional partial independence. *Econometrica*, 86(1):317–351.

Masten, M. A., Poirier, A., and Ren, M. (2025). A general approach to relaxing unconfoundedness. *arXiv preprint arXiv:2501.15400*.

Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–82.

Olma, T. (2021). Nonparametric estimation of truncated conditional expectation functions. *arXiv preprint arXiv:2109.06150*.

Oprescu, M., Dorn, J., Ghoummaid, M., Jesson, A., Kallus, N., and Shalit, U. (2023). B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In *International Conference on Machine Learning*, pages 26599–26618. PMLR.

Padilla, O. H. M., Tansey, W., and Chen, Y. (2022). Quantile regression with relu networks: Estimators and minimax rates. *Journal of Machine Learning Research*, 23(247):1–42.

Pu, H. and Zhang, B. (2021). Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):318–345.

Qi, Z., Pang, J.-S., and Liu, Y. (2023). On robustness of individualized decision rules. *Journal of the American Statistical Association*, 118(543):2143–2157.

Qiu, H., Carone, M., Sadikova, E., Petukhova, M., Kessler, R. C., and Luedtke, A. (2021). Optimal individualized decision rules using instrumental variable methods. *Journal of the American Statistical Association*, 116(533):174–191.

Rambachan, A., Coston, A., and Kennedy, E. (2022). Counterfactual risk assessments under unmeasured confounding. *arXiv preprint arXiv:2212.09844*.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218.

Sasaki, Y. and Ura, T. (2024). Welfare analysis via marginal treatment effects. *Econometric Theory*, pages 1–24.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875.

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. (2023). Distributionally robust batch contextual bandits. *Management Science*, 69(10):5772–5793.

Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S., and Wager, S. (2020). policytree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software*, 5(50):2232.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Vaart, A. v. d. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Viviano, D. and Bradic, J. (2024). Fair policy targeting. *Journal of the American Statistical Association*, 119(545):730–743.

Walters, C. R. (2015). Inputs in the production of early childhood human capital: Evidence from head start. *American Economic Journal: Applied Economics*, 7(4):76–102.

Wang, L., Zhou, Y., Song, R., and Sherwood, B. (2018). Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254.

Yu, M., Wang, Y., Xie, S., Tan, K. M., and Zhou, W.-X. (2025). Estimation and inference for nonparametric expected shortfall regression over rkhs. *Journal of the American Statistical Association*, pages 1–12.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761.

Zhou, Z., Athey, S., and Wager, S. (2023). Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183.

# A Identification Results under MSM

In this section, we present partial identification results for the conditional mean and CATE under the MSM, drawing primarily from (Dorn and Guo, 2023; Oprescu et al., 2023; Dorn et al., 2025). Recall that $\mu_o(x,a) = \mathbb{E}_{P_o}[Y(a)|X=x]$ denote the conditional mean of the potential outcome, and the true CATE is defined as $\tau_o(x) = \mu_o(x,1) - \mu_o(x,0)$.

To formally characterize partial identification, we describe the identified sets of these functions under distributional uncertainty, where the true counterfactual distribution may deviate within an uncertainty set $\mathcal{P}(\Lambda)$. For any distribution $Q \in \mathcal{P}(\Lambda)$, define $\mu_Q(x,a) = \mathbb{E}_Q[Y(a)|X=x]$ and $\tau_Q(x) = \mu_Q(x,1) - \mu_Q(x,0)$, which represent the conditional mean and CATE under $Q$. The identified sets for these functions are then

$$\Theta_{\mu,\Lambda}(x,a) \equiv \left\{\mu_Q(x,a) : Q \in \mathcal{P}(\Lambda)\right\},$$
$$\Theta_{\tau,\Lambda}(x) \equiv \left\{\tau_Q(x) : Q \in \mathcal{P}(\Lambda)\right\},$$

for any $x \in \mathcal{X}$ and $a \in \{0,1\}$. When $\Lambda = 1$, the uncertainty set $\mathcal{P}(\Lambda)$ collapses to a singleton, implying point identification: $\mu_o(x,a)$ and $\tau_o(x)$ are uniquely determined by $\Theta_{\mu,\Lambda}(x,a)$ and $\Theta_{\tau,\Lambda}(x)$.

For general $\Lambda$, let $\mu^{\pm}_{\Lambda}(x,a)$ denote the endpoints of $\Theta_{\mu,\Lambda}(x,a)$, i.e.,

$$\mu^{-}_{\Lambda}(x,a) = \inf_{Q\in\mathcal{P}(\Lambda)} \mu_Q(x,a) \quad \text{and} \quad \mu^{+}_{\Lambda}(x,a) = \sup_{Q\in\mathcal{P}(\Lambda)} \mu_Q(x,a).$$

The following Propositions A.1 and A.2 provide sharp upper and lower bounds for both $\Theta_{\mu,\Lambda}(x,a)$ and $\Theta_{\tau,\Lambda}(x)$.

**Proposition A.1.** Under Assumption 2.1, there are distributions $P^{\pm}_{\mu,a} \in \mathcal{P}(\Lambda)$ such that $\mu_{P^{+}_{\mu,a}}(x,a) = \mu^{+}_{\Lambda}(x,a)$ and $\mu_{P^{-}_{\mu,a}}(x,a) = \mu^{-}_{\Lambda}(x,a)$, almost surely. In particular, these sharp bounds can be expressed as:

$$\mu^{\pm}_{\Lambda}(x,a) = \mathbb{E}\left[Y\mathbb{1}\{A=a\}\left[1 + \frac{1-e_a(X)}{e_a(X)}\Lambda^{\pm\mathrm{sgn}\left(Y - q^{\pm}_{\Lambda}(X,a)\right)}\right]\middle| X=x\right], \tag{A.1}$$

where $\mathrm{sgn}(t) = 1$ if $t \geq 0$ and $-1$ otherwise.

We next characterize the upper and lower bounds for $\Theta_{\tau,\Lambda}(x)$. Specifically, the endpoints of $\Theta_{\tau,\Lambda}(x)$ are defined as

$$\tau^{-}_{\Lambda}(x) = \inf_{Q\in\mathcal{P}(\Lambda)} \tau_Q(x) \quad \text{and} \quad \tau^{+}_{\Lambda}(x) = \sup_{Q\in\mathcal{P}(\Lambda)} \tau_Q(x).$$

The following result, Proposition A.2, builds on Proposition A.1.

**Proposition A.2.** Under Assumption 2.1, there are distributions $P^{\pm}_{\tau} \in \mathcal{P}(\Lambda)$ such that $\mu_{P^{+}_{\tau}}(x) = \tau^{+}_{\Lambda}(x)$ and $\tau_{P^{-}_{\tau}}(x) = \tau^{-}_{\Lambda}(x)$, almost surely. Moreover, the sharp bounds for the CATE satisfy:

$$\tau^{\pm}_{\Lambda}(x) = \mu^{\pm}_{\Lambda}(x,1) - \mu^{\mp}_{\Lambda}(x,0). \tag{A.2}$$

We illustrate the intuition behind Proposition A.1. Suppose we had oracle access to the true propensity score $e_o(X,U)$. Under Assumption 2.1, the true conditional mean $\mu_o(x,a)$ would then be point-identified via IPW:

$$\mu_o(x,a) = \mathbb{E}_{P_o}\left[\frac{Y\mathbb{1}\{A=a\}}{ae_o(X,U) + (1-a)\left(1-e_o(X,U)\right)}\middle| X=x\right],$$

However, when $\Lambda > 1$, the true propensity score $e_o(X, U)$ is no longer identifiable. Instead, Assumption 2.1 restricts $e_o(X, U)$ to an uncertainty set, which in turn induces a partially identified set for $\mu_o(x, a)$. Specifically, the uncertainty set consists of all random variables $E$ that satisfy both the sensitivity restriction and the usual balancing constraints:

$$\Lambda^{-1} \leq \frac{E/(1-E)}{e(X)/(1-e(X))} \leq \Lambda, \tag{A.3}$$

$$\mathbb{E}[A/E|X] = \mathbb{E}[(1-A)/(1-E)|X] = 1. \tag{A.4}$$

Formally, we define

$$\mathcal{E}(\Lambda) = \{E : \text{Eq. (A.3) and Eq. (A.4) hold a.s.}\}.$$

Given this uncertainty set, the identified set $\Theta_{\mu,\Lambda}(x, a)$ can be equivalently expressed as:

$$\Theta_{\mu,\Lambda}(x, a) = \left\{ \mathbb{E}\left[ \frac{Y \mathbb{1}\{A = a\}}{aE + (1-a)(1-E)} \middle| X = x \right] : E \in \mathcal{E}(\Lambda) \right\}.$$

For completeness, the formal proofs of Propositions A.1 and A.2 are provided in Appendix C for interested readers.

**Remark A.1.** Our Propositions A.1 and A.2 extend the pointwise identification results of Dorn and Guo (2023); Oprescu et al. (2023) to a uniform setting, ensuring that the identified bounds hold simultaneously for almost all $x$ and $a$. Such uniform bounds are crucial for policy evaluation and learning under the MSM. First, computing the robust criteria $W_\Lambda(\pi)$ and $\Delta_\Lambda(\pi)$ involves integration over the covariate and treatment spaces, which requires uniformly valid bounds on $\Theta_\mu(x, a)$ and $\Theta_\tau(x)$. Second, the closed-form sharp bounds enable the construction of orthogonal moments for the efficient estimation of $W_\Lambda(\pi)$ and $\Delta_\Lambda(\pi)$. As shown by Kallus and Zhou (2021), the policy improvement guarantee can be severely biased by the slow convergence of propensity score estimation; by leveraging our identification results, this bias can be mitigated via orthogonal moments, yielding $\sqrt{n}$-consistent estimates of $W_\Lambda(\cdot)$ and $\Delta_\Lambda(\cdot)$.

**Remark A.2.** Kallus and Zhou (2021), building on Zhao et al. (2019), constructs an uncertainty set for the putative propensity scores by imposing unconditional balancing constraints, specifically $\mathbb{E}[A/E] = \mathbb{E}[(1-A)/(1-E)] = 1$. In contrast, our uncertainty set $\mathcal{E}(\Lambda)$ incorporates the conditional balancing constraints defined in Eq. (A.4), which are strictly stronger than those employed in Kallus and Zhou (2021). According to Dorn and Guo (2023), such conditional balancing constraints are necessary for obtaining sharp bounds on the CATE, as established in Proposition A.2. Consequently, the policy learned in Kallus and Zhou (2021) is more conservative than our MMI policy.

# B Extending MMW Policy to Multi-Valued Treatments

This section extends our max-min welfare (MMW) policy learning approach to the multi-valued treatment setting, providing theoretical justification for its application in Section 6.2. Our identification here is constructive, directly informing a procedure for estimating the lower bound of the worst-case welfare function. This, in turn, allows for policy optimization using a method directly analogous to that in Algorithm 1. As these procedures and the corresponding doubly robust estimation closely follow the binary treatment case, we omit a detailed discussion here.

The identification results, including sharp upper and lower bounds for conditional mean outcomes and a lower bound for the worst-case welfare function, are established in Appendix B.1. Detailed proofs are presented in Appendix D.

We first extend the MSM framework to the setting where the treatment variable $A$ takes more than two values. Let $\mathcal{A}$ denote the set of $d$ possible treatment levels: $\mathcal{A} = \{a_1, \ldots, a_d\}$. Define the true and nominal propensity scores, respectively, as $e_{o,a}(x,u) = \mathbb{P}_{P_o}[A = a | X = x, U = u]$ and $e_a(x) = \mathbb{P}_{P_o}[A = a | X = x]$ for $a \in \mathcal{A}$.

**Assumption B.1.** Suppose there exists a vector of unobserved confounders $U \in \mathbb{R}^k$ such that

$$(Y(a_1), Y(a_2), \ldots, Y(a_d)) \perp\!\!\!\perp A \mid (X, U)\,.$$

The distribution of $(X, Y(a_1), Y(a_2), \ldots, Y(a_d), A, U)$ satisfies the selection bias condition with $1 \leq \Lambda < \infty$ if the following inequality holds $P_o$-almost surely,

$$\frac{1}{\Lambda} \leq \frac{e_{o,a}(x,u)/\left(1 - e_{o,a}(x,u)\right)}{e_a(x)/\left(1 - e_a(x)\right)} \leq \Lambda \quad \text{for} \quad a \in \mathcal{A}. \tag{B.1}$$

Assumption B.1 naturally induces a distributional uncertainty set over the counterfactual distribution of $(X, Y(a_1), Y(a_2), \ldots, Y(a_d), A, U)$. Formally, let $\mathcal{P}_{\mathrm{M}}$ denote the set of all probability distributions $Q$ on $\mathcal{X} \times \mathbb{R}^d \times \mathcal{A} \times \mathbb{R}^k$ that satisfy:

(1) If $(X, Y(a_1), Y(a_2), \ldots, Y(a_d), A, U) \sim Q$, then $(Y(a_1), Y(a_2), \ldots, Y(a_d)) \perp\!\!\!\perp A \mid (X, U)$ under $Q$;

(2) If $Y = Y(A)$, then the distribution of $(X, Y, A)$ under $Q$ coincides with the observed-data distribution $P$;

(3) For all $a \in \mathcal{A}$, the odds ratio between the true propensity score and the nominal propensity score lies in $[1/\Lambda, \Lambda]$, i.e.,

$$\frac{1}{\Lambda} \leq \frac{\mathbb{P}(A = a | X, U)/\big(1 - \mathbb{P}(A = a | X, U)\big)}{\mathbb{P}(A = a | X)/\big(1 - \mathbb{P}(A = a | X)\big)} \leq \Lambda, \quad Q\text{-a.s.}$$

## B.1 Identifying the Worst-Case Welfare Function

For any $(x, a) \in \mathcal{X} \times \mathcal{A}$, let $\mu_Q(x,a) = \mathbb{E}_Q[Y(a) | X = x]$ denote the conditional mean under counterfactual distribution $Q$. The identified set for $\mu_o(x,a) = \mathbb{E}_{P_o}[Y(a) | X = x]$ is formally given by

$$\Theta_{\mathrm{M},\mu}(x,a) \equiv \left\{\mu_Q(x,a) : Q \in \mathcal{P}_{\mathrm{M}}\right\}.$$

Let $\mu^{\pm}(x,a)$ denote the endpoints of $\Theta_{\mathrm{M},\mu}(x,a)$, i.e.,

$$\mu^-(x,a) = \inf_{Q \in \mathcal{P}_{\mathrm{M}}} \mu_Q(x,a) \quad \text{and} \quad \mu^+(x,a) = \sup_{Q \in \mathcal{P}_{\mathrm{M}}} \mu_Q(x,a).$$

The following Proposition B.1 establishes the sharp upper and lower bounds for the identified set $\Theta_{\mathrm{M},\mu}(x,a)$.

**Proposition B.1.** Under Assumption B.1, there exists distributions $P^{\pm}_{\mu,a} \in \mathcal{P}_{\mathrm{M}}$ such that $\mu_{P^+_{\mu,a}}(x,a) = \mu^+(x,a)$ and $\mu_{P^-_{\mu,a}}(x,a) = \mu^-(x,a)$ for almost all $(x,a) \in \mathcal{X} \times \mathcal{A}$, respectively. In particular, the sharp bounds admit the closed-form representation:

$$\mu^{\pm}(x,a) = \mathbb{E}\left[Y \mathbb{1}\{A = a\}\left[1 + \frac{1 - e_a(X)}{e_a(X)} \Lambda^{\pm \mathrm{sgn}\left(Y - q^{\pm}(X,a)\right)}\right] \Big| X = x\right]. \tag{B.2}$$

In the multi-valued treatment setting, a policy is a mapping from the input space $\mathcal{X}$ to a decision $a \in \mathcal{A}$. A randomized policy can be represented as a function from $\mathcal{X}$ to the probability simplex

$\Delta(\mathcal{A})$ over the action space $\mathcal{A}$. In contrast, a deterministic policy is a function $\pi : \mathcal{X} \to \{0,1\}^d$, where the output indicates a deterministic choice among $d$ possible actions.

The worst-case welfare function $W(\pi)$ introduced in Section 2.2 can be naturally extended to the multi-valued treatment setting:

$$W(\pi) = \inf_{Q \in \mathcal{P}_M} \mathbb{E}_Q \left[ Y(\pi(X)) \right] = \inf_{Q \in \mathcal{P}_M} \mathbb{E}_Q \left[ \sum_{a \in \mathcal{A}} Y(a) \pi_a(X) \right],$$

Let $\Pi_M$ denote the set of multi-valued policies specified by the policy maker. The corresponding max-min welfare (MMW) policy is obtained by solving $\max_{\pi \in \Pi_M} W(\pi)$. The following theorem characterizes a lower bound for $W(\pi)$.

**Theorem B.1.** Under Assumption B.1, for any policy $\pi \in \Pi_M$,

$$W(\pi) \geq \mathbb{E} \left[ \sum_{a \in \mathcal{A}} \mu^-(X, a) \pi_a(X) \right].$$

**Remark B.1.** Whether the lower bound for $W(\pi)$ in Theorem B.1 is sharp remains an open question. While sharp lower bounds for $\mu_o(x,a)$ are available for any fixed $(x,a)$, identifying a single distribution that simultaneously attains these bounds for all $(x,a)$ may be challenging and lies beyond the scope of this study.

# C Proofs for Results in the Main Text

Since the sensitivity parameter $\Lambda$ is treated as fixed throughout, we omit the subscript $\Lambda$ for notational simplicity. For instance, we may write $\mathcal{P} \equiv \mathcal{P}(\Lambda)$, $\mu^+(x,a) \equiv \mu^+_\Lambda(x,a)$ and $\rho^-(x,a) \equiv \rho^-_\Lambda(x,a)$. This convention will be used throughout the remainder of the appendix.

## C.1 Proof of Proposition A.1

*Proof of Proposition A.1.* **Step 1. Preliminary Results.** We begin by showing that for any $(x,a) \in \mathcal{X} \times \{0,1\}$, the partially identified set $\Theta_\mu(x,a)$ is an interval whose endpoints solve the following optimization problems:

$$\begin{aligned} \mu^+(x,a) &= \sup_{E \in \mathcal{E}} \mathbb{E} \left[ \frac{Y \mathbb{1}\{A = a\}}{aE + (1-a)(1-E)} \middle| X = x \right], \\ \mu^-(x,a) &= \inf_{E \in \mathcal{E}} \mathbb{E} \left[ \frac{Y \mathbb{1}\{A = a\}}{aE + (1-a)(1-E)} \middle| X = x \right]. \end{aligned} \tag{C.1}$$

We focus on verifying the result for $\mu^+(x,a)$; the argument for $\mu^-(x,a)$ is analogous and omitted for brevity.

By Proposition 1B of Dorn and Guo (2023), for any random variable $E$ defined on the same probability space as $(X, Y, A)$ and satisfying $\mathbb{E}\left[\mathbb{1}\{A = a\} / \left(aE + (1-a)(1-E)\right) | X\right] = 1$, we can construct a distribution $Q \in \mathcal{P}$ such that

$$\mathbb{E}_Q \left[ Y(a) | X = x \right] = \mathbb{E}_Q \left[ \frac{Y \mathbb{1}\{A = a\}}{\mathbb{P}_Q(A = a | X, U)} \middle| X = x \right] = \mathbb{E} \left[ \frac{Y \mathbb{1}\{A = a\}}{aE + (1-a)(1-E)} \middle| X = x \right]. \tag{C.2}$$

Thus,

$$\mu^+(x,a) = \sup_{Q\in\mathcal{P}} \mathbb{E}_Q\left[Y(a)|X=x\right] \geq \mathbb{E}\left[\frac{Y\mathbb{1}\{A=a\}}{aE+(1-a)(1-E)}\Big|X=x\right].$$

Since this inequality holds for any $E$ satisfying the balancing condition, it holds in particular for the supremum over all such $E$.

Conversely, for any $Q\in\mathcal{P}$,

$$\begin{aligned}\mathbb{E}_Q\left[Y(a)|X=x\right] =&\mathbb{E}_Q\left[\frac{Y\mathbb{1}\{A=a\}}{\mathbb{P}_Q(A=a|X,U)}\Big|X=x\right]\\ =&\mathbb{E}_Q\left[Y\mathbb{1}\{A=a\}\mathbb{E}\left[\frac{1}{\mathbb{P}_Q(A=a|X,U)}\Big|X,Y,A=a\right]\Big|X=x\right].\end{aligned}$$

Define $e_Q(X,Y,A) = 1/\mathbb{E}\left[1/\mathbb{P}_Q(A=a|X,U)|X,Y,A\right]$, and introduce a random variable $E$ on the same probability space $P$ such that $AE+(1-A)(1-E) = e_Q(X,Y,A)$. It is straightforward to verify that $E\in\mathcal{E}$ and

$$\mathbb{E}_Q\left[Y(a)|X=x\right] = \mathbb{E}\left[\frac{Y\mathbb{1}\{A=a\}}{e_Q(X,Y,a)}\Big|X=x\right] = \mathbb{E}\left[\frac{Y\mathbb{1}\{A=a\}}{aE+(1-a)(1-E)}\Big|X=x\right].$$

Thus,

$$\mathbb{E}_Q\left[Y(a)|X=x\right] \leq \sup_{E\in\mathcal{E}} \mathbb{E}\left[\frac{Y\mathbb{1}\{A=a\}}{aE+(1-a)(1-E)}\Big|X=x\right].$$

Since $Q$ is arbitrary, the inequality continues to hold after taking the supremum over $Q\in\mathcal{P}$.

Finally, the proof that the partially identified set $\Theta_\mu(x,a)$ forms an interval follows the same argument as in Dorn and Guo (2023).

**Step 2. The Closed-form Expression.** For $a\in\{0,1\}$, define $E_a^\pm$ by

$$\frac{1}{aE_a^\pm+(1-a)(1-E_a^\pm)} = 1+\frac{1-e_a(X)}{e_a(X)}\Lambda^{\pm\mathrm{sgn}\left(Y-q^\pm(X,a)\right)}. \tag{C.3}$$

It is straightforward to verify that $E_a^\pm\in\mathcal{E}$, and thus $E_a^\pm$ is feasible for Eq. (C.1). For any $(x,a)\in\mathcal{X}\times 0,1$ and $E\in\mathcal{E}$, we have

$$\begin{aligned}&\mathbb{E}\left[\frac{Y\mathbb{1}\{A=a\}}{aE+(1-a)(1-E)}\Big|X=x\right]\\ =&q^+(x,a)\mathbb{E}\left[\frac{\mathbb{1}\{A=a\}}{aE+(1-a)(1-E)}\Big|X=x\right]+\mathbb{E}\left[\frac{\left(Y-q^+(X,a)\right)\mathbb{1}\{A=a\}}{aE+(1-a)(1-E)}\Big|X=x\right]\\ \leq&q^+(x,a)\mathbb{E}\left[\frac{\mathbb{1}\{A=a\}}{aE_a^+(1-a)(1-E_a^+)}\Big|X=x\right]+\mathbb{E}\left[\frac{\left(Y-q^+(X,a)\right)\mathbb{1}\{A=a\}}{aE_a^+(1-a)(1-E_a^+)}\Big|X=x\right]\\ =&\mathbb{E}\left[\frac{Y\mathbb{1}\{A=a\}}{aE_a^+(1-a)(1-E_a^+)}\Big|X=x\right],\end{aligned}$$

where the inequality follows because $1/\left(aE_a^+(1-a)(1-E_a^+)\right)$ attains the maximum (minimum) allowable value when $\left(Y-q^+(X,a)\right)\mathbb{1}\{A=a\}$ is positive (negative). By a similar argument, we can show that for any $(x,a)\in\mathcal{X}\times\{0,1\}$ and $E\in\mathcal{E}$,

$$\mathbb{E}\left[\frac{Y\mathbb{1}\{A=a\}}{aE+(1-a)(1-E)}\Big|X=x\right] \geq \mathbb{E}\left[\frac{Y\mathbb{1}\{A=a\}}{aE_a^-+(1-a)(1-E_a^-)}\Big|X=x\right].$$

Since $E$ is arbitrary, this completes the proof.

*Step 3. The Existence of Counterfactual Distribution.* Eqs. (C.1) and (C.3) together imply that random variables $E_a^{\pm} \in \mathcal{E}$ satisfy

$$\mu^+(x,a) = \mathbb{E}\left[\frac{Y\mathbb{1}\{A=a\}}{aE_a^+ + (1-a)(1-E_a^+)}\middle| X=x\right],$$
$$\mu^-(x,a) = \mathbb{E}\left[\frac{Y\mathbb{1}\{A=a\}}{aE_a^- + (1-a)(1-E_a^-)}\middle| X=x\right],$$

in $(x,a) \in \mathcal{X} \times \{0,1\}$ almost surely. By Proposition 1B in Dorn and Guo (2023) and Eq. (C.2), we conclude that there exist distributions $P_{\mu,a}^{\pm} \in \mathcal{P}$ such that $\mu_{P_{\mu,a}^+}(x,a) = \mu^+(x,a)$ and $\mu_{P_{\mu,a}^-}(x,a) = \mu^-(x,a)$ in $(x,a) \in \mathcal{X} \times \{0,1\}$ almost surely. □

## C.2 Proof of Proposition A.2

*Proof of Proposition A.2.* We complete the proof by constructing a data-compatible distribution in $\mathcal{P}$ that simultaneously attains $\mu^+(x,1)$ and $\mu^-(x,0)$. Similar arguments can be applied to derive the lower bound, which we omit for brevity.

Define $E_\tau^+ = AE_1^+ + (1-A)E_0^-$. It is straightforward to verify that $E_\tau^+ \in \mathcal{E}$. Applying Proposition A.1 and similar arguments as C.4.2 in Dorn and Guo (2023), it is straightforward to construct the distribution $P_\tau^+ \in \mathcal{P}$ such that

$$\mathbb{E}_{P_\tau^+}[Y(1)-Y(0)|X=x] = \mathbb{E}\left[\frac{YA}{E_\tau^+}\middle|X=x\right] - \mathbb{E}\left[\frac{Y(1-A)}{E_\tau^+}\middle|X=x\right] = \mu^+(x,1) - \mu^-(x,0)$$

in $x \in \mathcal{X}$ almost surely. Since

$$\begin{aligned}\tau^+(x) &= \sup_{Q\in\mathcal{P}} \mathbb{E}_Q[Y(1)-Y(0)|X=x] \\ &\leq \sup_{Q\in\mathcal{P}} \mathbb{E}_Q[Y(1)|X=x] - \inf_{Q\in\mathcal{P}} \mathbb{E}_Q[Y(0)|X=x] \\ &= \mu^+(x,1) - \mu^-(x,0),\end{aligned}$$

in $x \in \mathcal{X}$ almost surely. We conclude that $\tau^+(x) = \mu^+(x,1) - \mu^-(x,0)$ in $x \in \mathcal{X}$ almost surely. □

## C.3 Proof of Theorem 3.1

*Proof of Theorem 3.1.* Notice that

$$\begin{aligned}\inf_{Q\in\mathcal{P}} \mathbb{E}_Q[Y(\pi(X))] &= \inf_{Q\in\mathcal{P}} \left\{\mathbb{E}\Big[\pi(X)\mu_Q(X,1)\Big] + \mathbb{E}\Big[(1-\pi(X))\mu_Q(X,0)\Big]\right\} \\ &\geq \inf_{Q\in\mathcal{P}} \mathbb{E}\Big[\pi(X)\mu_Q(X,1)\Big] + \inf_{Q\in\mathcal{P}} \mathbb{E}\Big[(1-\pi(X))\mu_Q(X,0)\Big].\end{aligned}$$

Then we will show that inf and expectation operators are exchangeable. On one hand, for any distribution $Q \in \mathcal{P}$,

$$\mathbb{E}\Big[\pi(X)\mu_Q(X,1)\Big] \geq \mathbb{E}\left[\pi(X)\inf_{Q\in\mathcal{P}} \mu_Q(X,1)\right].$$

Since $Q$ is arbitrary, the inequality continues to hold after taking the infimum over $\mathcal{P}$ on both sides. For the other side, Proposition A.1 has shown that there exists a distribution $P_{\mu,1}^- \in \mathcal{P}$ satisfying

$\mu^-(x,1) = \mu_{P^-_{\mu,1}}(x,1)$ for any $x \in \mathcal{X}$. Thus,

$$\inf_{Q\in\mathcal{P}} \mathbb{E}\Big[\pi(X)\mu_Q(X,1)\Big] \le \mathbb{E}\Big[\pi(X)\mu_{P^-_{\mu,1}}(X,1)\Big] = \mathbb{E}\Big[\pi(X)\mu^-(X,1)\Big] = \mathbb{E}\left[\pi(X)\inf_{Q\in\mathcal{P}}\mu_Q(X,1)\right].$$

Combining the results above, we can conclude that

$$\inf_{Q\in\mathcal{P}} \mathbb{E}\Big[\pi(X)\mu_Q(X,1)\Big] = \mathbb{E}\left[\pi(X)\inf_{Q\in\mathcal{P}}\mu_Q(X,1)\right].$$

Similar arguments can be applied to derive that

$$\inf_{Q\in\mathcal{P}} \mathbb{E}\Big[(1-\pi(X))\mu_Q(X,0)\Big] = \mathbb{E}\left[(1-\pi(X))\inf_{Q\in\mathcal{P}}\mu_Q(X,0)\right].$$

Define $E^- = AE_1^- + (1-A)E_0^-$ satisfying $E^- \in \mathcal{E}$. Applying Proposition A.1 and similar arguments as C.4.3 in Dorn and Guo (2023), we can construct the distribution $P^-$ fulfilling

$$\mathbb{E}_{P^-}[Y(1)|X=x] = \mathbb{E}\left[\frac{YA}{E^-}\Big|X=x\right] = \mu^-(x,1)$$

and

$$\mathbb{E}_{P^-}[Y(0)|X=x] = \mathbb{E}\left[\frac{Y(1-A)}{E^-}\Big|X=x\right] = \mu^-(x,0).$$

which imply that

$$\mathbb{E}_{P^-}[Y(\pi(X))] = \mathbb{E}\left[\pi(X)\mu^-(X,1) + (1-\pi(X))\mu^-(X,0)\right].$$

Thus, we conclude that

$$\begin{aligned}
&\inf_{Q\in\mathcal{P}} \mathbb{E}_Q[Y(\pi(X))] \le \mathbb{E}_{P^-}[Y(\pi(X))] \\
=&\mathbb{E}\left[\pi(X)\mu^-(X,1) + (1-\pi(X))\mu^-(X,0)\right] \\
=&\mathbb{E}\left[\pi(X)\inf_{Q\in\mathcal{P}}\mu_Q(X,1)\right] + \mathbb{E}\left[(1-\pi(X))\inf_{Q\in\mathcal{P}}\mu_Q(X,0)\right].
\end{aligned}$$

Results above together indicate that

$$\begin{aligned}
\inf_{Q\in\mathcal{P}} \mathbb{E}_Q[Y(\pi(X))] =&\mathbb{E}\left[\pi(X)\inf_{Q\in\mathcal{P}}\mu_Q(X,1)\right] + \mathbb{E}\left[(1-\pi(X))\inf_{Q\in\mathcal{P}}\mu_Q(X,0)\right] \\
=&\mathbb{E}\Big[\pi(X)\left(\mu^-(X,1) - \mu^-(X,0)\right)\Big] + \mathbb{E}[\mu^-(X,0)],
\end{aligned}$$

which completes the proof. □

## C.4 Proof of Theorem 3.2

*Proof of Theorem 3.2.* Applying similar arguments as the proof of Theorem 3.1, we have that

$$\inf_{Q\in\mathcal{P}} \mathbb{E}\left[\tau_Q(X)\pi(X)\right] = \mathbb{E}\left[\inf_{Q\in\mathcal{P}}\tau_Q(X)\pi(X)\right] = \mathbb{E}\left[\tau^-(X)\pi(X)\right]. \tag{C.4}$$

According to Eq. (C.4), the first-best policy is given by $\pi^\star_\Delta(x) = \mathbb{1}\{\tau^-(x) > 0\}$. □

## C.5 Proof of Proposition 3.1

Recall that

$$\psi_W(z,\pi;\eta_W) = \sum_{t\in\{0,1\}} \phi_t^-(z;\eta_W)\,\pi(t|x),$$

where $\eta_W = \left(e, q^-, \rho_1^-, \rho_0^-\right)$, $\phi_t^-(z;\eta_W)$ are defined in Eq. (3.2), and the functions $\rho_1^{\pm}(x,t)$ and $\rho_0^{\pm}(x,t)$ are defined in Eq. (3.3).

*Proof of Proposition 3.1.* **Step 1.** The proof of part (1) follows directly from Theorem 3.1, together with the following derivation:

$$\mathbb{E}\left[\mathbb{1}\{A=t\}\left(\frac{1}{1+\Lambda} - \mathbb{1}\left\{Y < q^-(X,t)\right\}\right)\Big|X\right] = \mathbb{E}[\mathbb{1}\{A=t\} - e_t(X)|X] = 0. \tag{C.5}$$

**Step 2.** To show part (2), given the linearity of pathwise derivative, it suffices to show the pathwise derivative of $r \mapsto \mathbb{E}\left[\psi_W(Z,\pi;\eta_{W,\Lambda} + r\bar{\eta})\right] = 0$ for all perturbation directions $\bar{\eta}$ varying only in the components corresponding to $e, q^-, \rho_1^-$ and $\rho_0^-$, respectively. We only show that

$$\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\left[\psi_W\left(Z,\pi;q^- + r(\widetilde{q}^- - q^-)\right)\right]_{r=0} = 0, \tag{C.6}$$

for any $\widetilde{q}^-$ belonging to some small neighborhood of $q^-$. The derivations of the pathwise derivatives with respect to the perturbation directions in $e$, $\rho_1^-$ and $\rho_0^-$ are analogous and thus omitted for brevity. Notice that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\left[\psi_W\left(Z,\pi;q^- + rh\right)\right] \\
=&\sum_{t\in\{0,1\}} \frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\left[\phi_t^-\left(Z;q^- + r(\widetilde{q}^- - q^-)\right)\pi(t|X)\right].
\end{aligned}$$

To derive Eq. (C.6), it suffices to show that

$$\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\left[\phi_1^-\left(Z;q^- + r\left(\widetilde{q}^- - q^-\right)\right)\pi(t|X)\right]\Big|_{r=0} = 0,$$

for all $t \in \{0,1\}$. By Eq. (C.5), we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\left[\phi_1^-\left(Z;e,q^- + r\left(\widetilde{q}^- - q^-\right),\rho_1^-,\rho_0^-\right)\pi(X)\right]_{r=0} \\
=&\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\left[YA\left(1 + \frac{1-e(X)}{e(X)}\Lambda^{-\mathrm{sgn}\left(Y - q^-(X,1) - r\left(\widetilde{q}^-(X,1) - q^-(X,1)\right)\right)}\right)\right]\Big|_{r=0} \\
+&\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\left[q^-(X,1)A\frac{1-e(X)}{e(X)}\left(\Lambda - \Lambda^{-1}\right)\left(\frac{1}{1+\Lambda} - \mathbb{1}\left\{Y < q^-(X,1) + r\left(\widetilde{q}^-(X,1) - q^-(X,1)\right)\right\}\right)\right]_{r=0}
\end{aligned}$$

Let $\mathrm{II}_{A.1}$ and $\mathrm{II}_{A.2}$ denote the two summands in the expression above. It is not difficult to verify that

$$\begin{aligned}
\mathrm{II}_{A.1} &= \frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\left[YA\frac{1-e(X)}{e(X)}\left(\Lambda - \Lambda^{-1}\right)\mathbb{1}\left\{Y < q^-(X,1) + r\left(\widetilde{q}^-(X,1) - q^-(X,1)\right)\right\}\right]_{r=0} \\
&= \mathbb{E}\left[q^-(X,1)\frac{1-e(X)}{e(X)}\left(\Lambda - \Lambda^{-1}\right)f_Y\left(q^-(X,1)|X,1\right)\left(\widetilde{q}^-(X,1) - q^-(X,1)\right)\right] \\
&= \mathrm{II}_{A.2},
\end{aligned}$$

which implies that the pathwise derivative with respect to $q^-$ is zero. This concludes the proof of Part (2).

□

## C.6 Proof of Proposition 3.2

*Proof of Proposition 3.2.* The proof is identical to that of Proposition 3.1 and is omitted here. □

## C.7 Proof of Lemma 4.1

*Proof of Lemma 4.1.* We focus on deriving the result for $\widehat{W}_n(\pi) - W_n(\pi)$; the argument for $\widehat{\Delta}_n(\pi) - \Delta_n(\pi)$ is similar and omitted for brevity. Notice that

$$\begin{aligned}\widehat{W}_n(\pi) - W_n(\pi) =& \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\pi(X_i)\left(\phi_1^-\left(Z_i;\widehat{\eta}_W^{-k}\right) - \phi_1^-\left(Z_i;\eta_W\right)\right)\\ &- \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}(1-\pi(X_i))\left(\phi_0^-\left(Z_i;\widehat{\eta}_W^{-k}\right) - \phi_0^-\left(Z_i;\eta_W\right)\right).\end{aligned}$$

We bound the first term at the right-hand-side. The analysis of the second term is similar and omitted. To simplify notation, we suppress the superscript $-k$ of the nuisance estimators, e.g., $\widehat{\eta}_W = \widehat{\eta}_W^{-k}$. By definition,

$$\begin{aligned}&\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\pi(X_i)\left(\phi_1^-\left(Z_i;\widehat{\eta}_W\right) - \phi_1^-\left(Z_i;\eta_W\right)\right)\\ =&\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\pi(X_i)Y_iA_i\left(\frac{1-\widehat{e}(X_i)}{\widehat{e}(X_i)}\Lambda^{-\mathrm{sgn}\left(Y_i-\widehat{q}^-(X_i,1)\right)} - \frac{1-e(X_i)}{e(X_i)}\Lambda^{-\mathrm{sgn}\left(Y_i-q^-(X_i,1)\right)}\right)\\ +&\frac{\Lambda-\Lambda^{-1}}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\pi(X_i)A_i\left(\frac{1}{1+\Lambda} - \mathbb{1}\left\{Y_i<\widehat{q}^-(X_i,1)\right\}\right)\left(\widehat{q}^-(X_i,1)\frac{1-\widehat{e}(X_i)}{\widehat{e}(X_i)} - q^-(X_i,1)\frac{1-e(X_i)}{e(X_i)}\right)\\ -&\frac{\Lambda-\Lambda^{-1}}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\pi(X_i)A_iq^-(X_i,1)\frac{1-e(X_i)}{e(X_i)}\left(\mathbb{1}\left\{Y_i<\widehat{q}^-(X_i,1)\right\} - \mathbb{1}\left\{Y_i<q^-(X_i,1)\right\}\right)\\ -&\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\pi(X_i)A_i\left(\Lambda\widehat{\rho}_1^-(X_i,1)+\Lambda^{-1}\widehat{\rho}_0^-(X_i,1)\right)\left(\frac{1}{\widehat{e}(X_i)} - \frac{1}{e(X_i)}\right)\\ -&\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\pi(X_i)\frac{A_i-e(X_i)}{e(X_i)}\left[\Lambda\left(\widehat{\rho}_1^-(X_i,1)-\rho_1^-(X_i,1)\right)+\Lambda^{-1}\left(\widehat{\rho}_0^-(X_i,1)-\rho_0^-(X_i,1)\right)\right].\end{aligned}$$

Denote these five summands by $D_j(\pi)$ for $1\le j\le 5$.

*First term.* To bound the first term, it is useful to separate the contributions of each of the $K$ folds:

$$\begin{aligned}
D_{1,k}(\pi) =& \frac{1}{n}\sum_{i\in\mathcal{I}_k}\pi(X_i)Y_iA_i\left(\frac{1-\widehat{e}(X_i)}{\widehat{e}(X_i)}\Lambda^{-\mathrm{sgn}\left(Y_i-\widehat{q}^-(X_i,1)\right)}-\frac{1-e(X_i)}{e(X_i)}\Lambda^{-\mathrm{sgn}\left(Y_i-q^-(X_i,1)\right)}\right)\\
=&\frac{1}{n}\sum_{i\in\mathcal{I}_k}\pi(X_i)Y_iA_i\Lambda^{-\mathrm{sgn}\left(Y_i-q^-(X_i,1)\right)}\left(\frac{1}{\widehat{e}(X_i)}-\frac{1}{e(X_i)}\right)\\
+&\frac{\Lambda-\Lambda^{-1}}{n}\sum_{i\in\mathcal{I}_k}\pi(X_i)Y_iA_i\frac{1-e(X_i)}{e(X_i)}\left(\mathbb{1}\left\{Y_i<\widehat{q}^-(X_i,1)\right\}-\mathbb{1}\left\{Y_i<q^-(X_i,1)\right\}\right)\\
+&\frac{\Lambda-\Lambda^{-1}}{n}\sum_{i\in\mathcal{I}_k}\pi(X_i)Y_iA_i\left(\frac{1}{\widehat{e}(X_i)}-\frac{1}{e(X_i)}\right)\left(\mathbb{1}\left\{Y_i<\widehat{q}^-(X_i,1)\right\}-\mathbb{1}\left\{Y_i<q^-(X_i,1)\right\}\right)\\
=&D_{1,k}^{(1)}(\pi)+D_{1,k}^{(2)}(\pi)+D_{1,k}^{(3)}(\pi).
\end{aligned}$$

We can use the Cauchy-Schwartz inequality to verify that

$$\begin{aligned}
\sup_{\pi\in\Pi_n}\left|D_{1,k}^{(3)}(\pi)\right| \le& \left(\Lambda-\Lambda^{-1}\right)\sqrt{\frac{1}{n}\sum_{i\in\mathcal{I}_k}Y_i^2\left(\frac{1}{\widehat{e}(X_i)}-\frac{1}{e(X_i)}\right)^2}\\
&\times\sqrt{\frac{1}{n}\sum_{i\in\mathcal{I}_k}\left(\mathbb{1}\left\{Y_i<\widehat{q}^-(X_i,1)\right\}-\mathbb{1}\left\{Y_i<q^-(X_i,1)\right\}\right)^2}\\
=&O_P\left(\frac{b(n)}{n^{(\zeta_e+\zeta_q)/2}}\right)=O_P\left(n^{-1/2}\right),
\end{aligned}$$

where the first equality holds by Markov's inequality, Assumption 4.2 and Assumption 4.4. Then, uniformly over $\pi\in\Pi_n$,

$$D_{1,k}(\pi)=D_{1,k}^{(1)}(\pi)+D_{1,k}^{(2)}(\pi)+O_P\left(n^{-1/2}\right).$$

*Second term.* For bounding the second term, we still separate out the contributions of the $K$ different folds. After applying similar arguments as the proceeding one, we obtain that uniformly over $\pi\in\Pi_n$,

$$D_{2,k}(\pi)=\widetilde{D}_{2,k}(\pi)+O_P\left(n^{-1/2}\right), \tag{C.7}$$

where

$$\begin{aligned}
\widetilde{D}_{2,k}(\pi)=&\frac{\Lambda-\Lambda^{-1}}{n}\sum_{i\in\mathcal{I}_k}\pi(X_i)A_i\left(\frac{1}{1+\Lambda}-\mathbb{1}\left\{Y_i<q^-(X_i,1)\right\}\right)\\
&\times\left(\widehat{q}^-(X_i,1)\frac{1-\widehat{e}(X_i)}{\widehat{e}(X_i)}-q^-(X_i,1)\frac{1-e(X_i)}{e(X_i)}\right).
\end{aligned}$$

Since $\widehat{q}^-$ and $\widehat{e}$ were computed using data $\{Z_i\}_{i\in\mathcal{I}_k^c}$, one has

$$\mathbb{E}\left[\widetilde{D}_{2,k}(\pi)\middle|\widehat{q}^-,\widehat{e}\right]=0.$$

By Assumption 4.4, we have that uniformly over $i$,

$$\left|\widehat{q}^-(X_i,1)\frac{1-\widehat{e}(X_i)}{\widehat{e}(X_i)}-q^-(X_i,1)\frac{1-e(X_i)}{e(X_i)}\right|\le 1$$

with probability approaching to 1. So the individual summands in $\widetilde{D}_{2,k}(\pi)$ are all $\nu$-sub Gaussian with probability approaching to 1. Define

$$\widehat{V}_{2,k} = \frac{\Lambda}{(1+\Lambda)^2}\mathbb{E}\left[e(X_i)\left(\widehat{q}^-(X_i,1)\frac{1-\widehat{e}(X_i)}{\widehat{e}(X_i)} - q^-(X_i,1)\frac{1-e(X_i)}{e(X_i)}\right)^2 \middle| \widehat{q}^-, \widehat{e}\right].$$

We can apply Corollary 3 in Athey and Wager (2021) to establish that

$$\begin{aligned}\mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|\widetilde{D}_{2,k}(\pi)\right| \middle| \widehat{q}^-, \widehat{e}\right] =& \mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|\widetilde{D}_{2,k}(\pi) - \mathbb{E}\left[\widetilde{D}_{2,k}(\pi)\middle|\widehat{q}^-, \widehat{e}\right]\right| \middle| \widehat{q}^-, \widehat{e}\right] \\ =& O\left(\sqrt{\mathrm{VC}(\Pi_n)\frac{\widehat{V}_{2,k}}{n}}\right).\end{aligned} \tag{C.8}$$

Assumption 4.4, an application of Jensen's inequality and Eqs. (C.7) and (C.8) result that

$$\mathbb{E}\left[\sup_{\pi\in\Pi_n}|D_2(\pi)|\right] = O\left(\sqrt{\mathrm{VC}(\Pi_n)\frac{b(n)}{n^{1+\min\{\zeta_e,\zeta_q\}}}}\right).$$

*Third term.* Denote

$$D_{3,k}(\pi) = -\frac{\Lambda-\Lambda^{-1}}{n}\sum_{i\in\mathcal{I}_k}\pi(X_i)A_i q^-(X_i,1)\frac{1-e(X_i)}{e(X_i)}\left(\mathbb{1}\left\{Y_i < \widehat{q}^-(X_i,1)\right\} - \mathbb{1}\left\{Y_i < q^-(X_i,1)\right\}\right).$$

as the $k$-th fold in $D_3(\pi)$. Then

$$\begin{aligned}&D_{1,k}^{(2)}(\pi) + D_{3,k}(\pi) \\ =&\frac{\Lambda-\Lambda^{-1}}{n}\sum_{i\in\mathcal{I}_k}\pi(X_i)A_i\frac{1-e(X_i)}{e(X_i)}\left(Y_i - q^-(X_i,1)\right)\left(\mathbb{1}\left\{Y_i < \widehat{q}^-(X_i,1)\right\} - \mathbb{1}\left\{Y_i < q^-(X_i,1)\right\}\right).\end{aligned}$$

Define

$$\widehat{V}_{3,k} = \mathbb{E}\left[A_i\left(Y_i - q^-(X_i,1)\right)^2\left(\mathbb{1}\left\{Y_i < \widehat{q}^-(X_i,1)\right\} - \mathbb{1}\left\{Y_i < q^-(X_i,1)\right\}\right)^2 \middle| \widehat{q}^-\right],$$

and it can be bounded as follows:

$$\mathbb{E}\left[\widehat{V}_{3,k}\right] = \mathbb{E}\left[e(X_i)\mathbb{E}\left[\left(Y_i - q^-(X_i,1)\right)^2\left(\mathbb{1}\left\{Y_i < \widehat{q}^-(X_i,1)\right\} - \mathbb{1}\left\{Y_i < q^-(X_i,1)\right\}\right)^2 \middle| X_i, A_i = 1, \widehat{q}^-\right]\right].$$

Notice that

$$\begin{aligned}&\mathbb{E}\left[\left(Y_i - q^-(X_i,1)\right)^2\left(\mathbb{1}\left\{Y_i < \widehat{q}^-(X_i,1)\right\} - \mathbb{1}\left\{Y_i < q^-(X_i,1)\right\}\right)^2 \middle| X_i, A_i = 1, \widehat{q}^-\right] \\ =&\int_{\widehat{q}^-(X_i,1)}^{q^-(X_i,1)}\left(y - q^-(X_i,1)\right)^2 dF_{Y_i}\left(y\middle|X_i, A_i = 1, \widehat{q}^-\right) + \int_{q^-(X_i,1)}^{\widehat{q}^-(X_i,1)}\left(y - q^-(X_i,1)\right)^2 dF_{Y_i}\left(y\middle|X_i, A_i = 1, \widehat{q}^-\right) \\ =&O_P\left(\left(\widehat{q}^-(X_i,1) - q^-(X_i,1)\right)^3\right),\end{aligned}$$

where the last equality follows by the mean value theorem. Hence, by Assumption 4.4,

$$\mathbb{E}\left[\widehat{V}_{3,k}\right] = O\left(\frac{b(n)}{n^{3\zeta_q/2}}\right).$$

After applying a similar argument as the proof of $\widetilde{D}_{2,k}(\pi)$, we have that for any $k$,

$$\mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|D_{1,k}^{(2)}(\pi)+D_{3,k}(\pi)-\mathbb{E}\left[D_{1,k}^{(2)}(\pi)+D_{3,k}(\pi)\middle|\widehat{q}^{-}\right]\right|\middle|\widehat{q}^{-}\right] = O\left(\sqrt{\mathrm{VC}(\Pi_n)\frac{b(n)}{n^{1+3\zeta_q/2}}}\right). \tag{C.9}$$

By Assumption 4.3, $\pi(X_i)\big(1-e(X_i)\big)/e(X_i)$ is bounded uniformly over $i$ and $\pi$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[D_{1,k}^{(2)}(\pi)+D_{3,k}(\pi)\middle|\widehat{q}^{-}\right]\\
\leq&\frac{\Lambda-\Lambda^{-1}}{K}\frac{1-\kappa}{\kappa}\mathbb{E}\left[A_i\middle|Y_i-q^{-}(X_i,1)\middle|\left|\mathbb{1}\left\{Y_i<\widehat{q}^{-}(X_i,1)\right\}-\mathbb{1}\left\{Y_i<q^{-}(X_i,1)\right\}\right|\middle|\widehat{q}^{-}\right]\\
=&O\left(\frac{b(n)}{n^{\zeta_q}}\right)
\end{aligned} \tag{C.10}$$

uniformly over $\pi\in\Pi_n$. Applying Jensen's inequality, triangle inequality, and Eqs. (C.9) and (C.10) yield

$$\mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|D_1^{(2)}(\pi)+D_3(\pi)\right|\right] = O\left(\sqrt{\mathrm{VC}(\Pi_n)\frac{b(n)}{n^{1+3\zeta_q/2}}}+\frac{b(n)}{n^{\zeta_q}}\right).$$

*Fourth term.* Define the $k$-th fold of $D_4(\pi)$ as

$$D_{4,k}(\pi) = -\frac{1}{n}\sum_{i\in\mathcal{I}_k}\pi(X_i)A_i\left(\Lambda\widehat{\rho}_1^{-}(X_i,1)+\Lambda^{-1}\widehat{\rho}_0^{-}(X_i,1)\right)\left(\frac{1}{\widehat{e}(X_i)}-\frac{1}{e(X_i)}\right).$$

Applying similar arguments as the proceeding one, we can derive that

$$D_{4,k}(\pi) = \widetilde{D}_{4,k}(\pi)+O_P\left(n^{-1/2}\right)$$

uniformly over $\pi\in\Pi_n$, where

$$\widetilde{D}_{4,k}(\pi) = -\frac{1}{n}\sum_{i\in\mathcal{I}_k}\pi(X_i)A_i\left(\Lambda\rho_1^{-}(X_i,1)+\Lambda^{-1}\rho_0^{-}(X_i,1)\right)\left(\frac{1}{\widehat{e}(X_i)}-\frac{1}{e(X_i)}\right).$$

Then we consider

$$\begin{aligned}
&D_{1,k}^{(1)}(\pi)+\widetilde{D}_{4,k}(\pi)\\
=&\frac{1}{n}\sum_{i\in\mathcal{I}_k}\pi(X_i)A_i\left(Y_i\Lambda^{-\mathrm{sgn}\left(Y_i-q^{-}(X_i,1)\right)}-\Lambda\rho_1^{-}(X_i,1)-\Lambda^{-1}\rho_0^{-}(X_i,1)\right)\left(\frac{1}{\widehat{e}(X_i)}-\frac{1}{e(X_i)}\right).
\end{aligned}$$

It is not difficult to verify that

$$\mathbb{E}\left[D_{1,k}^{(1)}(\pi)+\widetilde{D}_{4,k}(\pi)\middle|\widehat{e}\right] = 0,$$

because

$$\mathbb{E}\left[Y_i\Lambda^{-\operatorname{sgn}\left(Y_i-q^-(X_i,1)\right)}-\Lambda\rho_1^-(X_i,1)-\Lambda^{-1}\rho_0^-(X_i,1)\middle|X_i,A_i=1,\widehat{e}\right]=0.$$

Similar arguments are used to show that

$$\mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|D_1^{(1)}(\pi)+D_4(\pi)\right|\right]=O\left(\sqrt{\mathrm{VC}(\Pi_n)\frac{b(n)}{n^{1+\zeta_e}}}\right).$$

*Fifth term.* Recall that $\mathbb{E}\left[A_i-e(X_i)\middle|X_i,\widehat{\rho_1^-},\widehat{\rho_0^-}\right]=0$. Applying a similar argument to bound the fourth term, we can similarly bound the fifth term as follows:

$$\mathbb{E}\left[\sup_{\pi\in\Pi_n}|D_5(\pi)|\right]=O\left(\sqrt{\mathrm{VC}(\Pi_n)\frac{b(n)}{n^{1+\zeta_\rho}}}\right).$$

*Wrapping Up Lemma 4.1.* Combining five terms above with Assumption 4.1 and Assumption 4.4 gives

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi\in\Pi_n}\left|\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\pi(X_i)\left(\phi_1^-\left(Z_i;\widehat{\eta}_W\right)-\phi_1^-\left(Z_i;\eta_W\right)\right)\right|\right]\\
=&O\left(\sqrt{\frac{\mathrm{VC}(\Pi_n)b\left(n\right)}{n^{1+\min\left\{\zeta_\rho,\zeta_q,\zeta_e\right\}}}}\right)+O\left(\frac{b(n)}{n^{\zeta_q}}\right)\\
=&o(n^{-1/2}).
\end{aligned}$$

□

## C.8 Proof of Theorem 4.1

*Proof of Theorem 4.1.* We prove Eq. (4.3); the proof of Eq. (4.4) follows by a similar argument and is omitted here for brevity.

Suppose $\pi_{W,n}\in\operatorname{argmax}_{\pi\in\Pi_n}W_n(\pi)$. Otherwise, the proof can be modified by incorporating an $\varepsilon$-approximate optimizer argument. By the definitions of $\pi_{W,n}$ and $\widehat{\pi}_{W,n}$ and Assumption 4.4, it follows that for any $\pi\in\Pi_n$:

$$\begin{aligned}
W(\pi)-W(\widehat{\pi}_{W,n})=&W(\pi)-W_n(\pi)+\underbrace{W_n(\pi)-W_n(\pi_{W,n})}_{\leq 0}+\underbrace{W_n(\pi_{W,n})-\widehat{W}_n(\pi_{W,n})}_{=o_P(n^{-1/2})}\\
&+\underbrace{\widehat{W}_n(\pi_{W,n})-\widehat{W}_n(\widehat{\pi}_{W,n})}_{\leq 0}+\underbrace{\widehat{W}_n(\widehat{\pi}_{W,n})-W_n(\widehat{\pi}_{W,n})}_{o_P(n^{-1/2})}+W_n(\widehat{\pi}_{W,n})-W(\widehat{\pi}_{W,n})\\
\leq&\ W(\pi)-W_n(\pi)+W_n(\widehat{\pi}_{W,n})-W(\widehat{\pi}_{W,n})+r_n\\
\leq&\ 2\sup_{\pi\in\Pi_n}|W(\pi)-W_n(\pi)|+r_n
\end{aligned}$$

where $r_n=o_P(n^{-1/2})$ and $\sqrt{n}\mathbb{E}|r_n|=o(1)$. For notational simplicity, we write

$$\psi_{W,\pi}:z\mapsto\psi_W(z,\pi;\eta_W).$$

Thus, for all $\pi \in \Pi_n$:

$$\begin{aligned} W(\pi) - W(\widehat{\pi}_{W,n}) &\leq 2 \sup_{\pi \in \Pi_n} |W_n(\pi) - W(\pi)| + r_n \\ &= 2 \sup_{\pi \in \Pi_n} |(\mathbb{E}_n - \mathbb{E})\, \psi_{W,\pi}| + r_n. \end{aligned} \tag{C.11}$$

Without loss of generality, suppose that there exists $\pi^*_{W,n} \in \Pi_n$ such that $W(\pi^*_{W,n}) = \max_{\pi \in \Pi_n} W(\pi)$. If no such $\pi^*_{W,n}$ exists, the proof can be adapted using an $\varepsilon$-approximate optimizer, where $\varepsilon \to 0$. Substituting $\pi^*_{W,n}$ into the preceding expression yields

$$0 \leq \mathrm{Reg}_W(\widehat{\pi}_{W,n}) = W(\pi^*_{W,n}) - W(\widehat{\pi}_{W,n}) \leq 2 \sup_{\pi \in \Pi_n} |(\mathbb{E}_n - \mathbb{E})\, \psi_{W,\pi}| + r_n.$$

Define $\mathcal{F}_n = \{\psi_{W,\pi} : \pi \in \Pi_n\}$. To complete the proof, it suffices to show the empirical process term decaying at rate $\sqrt{\mathrm{VC}(\Pi_n)/n}$, i.e.,

$$\sup_{\psi \in \mathcal{F}_n} |(\mathbb{E}_n - \mathbb{E})\, \psi| = \sup_{\pi \in \Pi_n} |(\mathbb{E}_n - \mathbb{E})\, \psi_{W,\pi}| = O_P\left(\sqrt{\mathrm{VC}(\Pi_n)/n}\right). \tag{C.12}$$

We show Eq. (C.12) using a similar argument in Theorem 2.14.1 in Vaart and Wellner (1996), while allowing the function class $\mathcal{F}_n$ to vary with $n$.

**Step 1. Envelope function for $\mathcal{F}_n$.** As a first step for upper bounding $\sup_{\psi \in \mathcal{F}_n} |(\mathbb{E}_n - \mathbb{E})\, \psi|$, we construct an envelope function for the function class $\mathcal{F}_n$. Under Assumption 4.3, we have the following bound:

$$\begin{aligned} \mathbb{E}|Y|^2 &= \sum_{a \in \{0,1\}} \int \mathbb{E}\left[Y^2 | X = x, A = a\right] e_a(x) \mathrm{d}F_{X|A}(x|a) \\ &\geq \kappa \int \mathbb{E}\left[Y^2 | X = x, A = a\right] \mathrm{d}F_{X|A}(x|a), \end{aligned}$$

where $F_{X|A}(x|a)$ denote the conditional distribution of $X$ given $A = a$. Now, for any fixed $\Lambda \in [1, \infty)$ and $j, a \in \{0, 1\}$, applying Jensen's inequality along with the inequality above yields:

$$\begin{aligned} \mathbb{E}\left|\rho^-_{j,\Lambda}(X, a)\right|^2 &\leq \mathbb{E}\left|\mathbb{E}\left[Y^2 | X, A = a\right]\right|^2 = \int \mathbb{E}\left[Y^2 | X = x, A = a\right] \mathrm{d}F_{X|A=a}(x) \\ &\leq \mathbb{E}|Y|^2/\kappa. \end{aligned}$$

As a result, the functions $e(x)$, $q^-(x, a)$, $\rho^-_1(x, a)$, and $\rho^-_0(x, a)$ are all square-integrable, and hence $\phi^-_t(z; \eta_W)$ defined in Eq. (3.2) for $t \in \{0, 1\}$ is also square-integrable by Minkowski inequality. The function $F(\cdot) \equiv \left|\phi^-_0(\cdot; \eta_W)\right| + \left|\phi^-_1(\cdot; \eta_W)\right|$ is an envelop function for $\mathcal{F}_n$ satisfying $|\psi| \leq F$ for all $\psi \in \mathcal{F}_n$ and $\|F\|_{\mathbb{P},2} < \infty$.

**Step 2. Uniformly entropy bound.** Second, we derive a uniform upper bound on the covering number of $\mathcal{F}_n$ for all $n$. Lemma A.1 in supplement of Kitagawa and Tetenov (2018) implies that $\mathrm{VC}(\mathcal{F}_n) \leq \mathrm{VC}(\Pi_n)$. By Theorem 2.6.7 in Vaart and Wellner (1996), there is a universal $K > 0$ such that for all $n$, the following inequality holds

$$\sup_Q \log N\left(\epsilon \|F\|_{Q,2}, \mathcal{F}_n, L^2(Q)\right) \leq K \mathrm{VC}(\mathcal{F}_n) \log \epsilon^{-1}, \quad \forall 0 < \epsilon < 1,$$

where the supremum is taken over all discrete probability measures $Q$ with $\|F\|_{Q,2} > 0$. As a

consequence, the uniform entropy can be upper bounded by:

$$\sup_Q \int_0^1 \sqrt{1+\log N\left(\epsilon\|F\|_{Q,2}, \mathcal{F}_n, L^2(Q)\right)}\mathrm{d}\epsilon \leq 1+\sqrt{K}(1+\sqrt{\pi}/2)\sqrt{\mathrm{VC}(\Pi_n)}.$$

It is important to note that the universal constant $K$ is independent of the function class $\mathcal{F}_n$; see Theorem 2.6.4 and 2.6.7 in Vaart and Wellner (1996) for details.

**Step 3. Bounding the supremum of the symmetrized process.** Given the sample $\{Z_i\}_{i=1}^n$, define the symmetrized empirical process $\{\mathbb{G}_n^o f : f \in \mathcal{F}_n\}$ as

$$\mathbb{G}_n^o : f \mapsto \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(Z_i),$$

where the $\varepsilon_i$ are independent Rademacher random variables such that $\varepsilon_i = \pm 1$ with probability 1/2 each. Moreover, let $\phi(x) = e^{x^2} - 1$ and the conditional Orlicz norm on $\mathcal{F}_n$ is defined as

$$\|\mathbb{G}_n^o f\|_{\phi,n} = \inf\left\{c > 0 : \mathbb{E}\left[\phi\left(|\mathbb{G}_n^o f|/c\right)|\{Z_i\}_{i=1}^n\right] \leq 1\right\}.$$

For more details of Orlicz norm, see in Chapter 2.2 in Vaart and Wellner (1996).

Conditionally on $\{Z_i\}_{i=1}^n$, the process $\mathbb{G}_n^o$ is sub-Gaussian for the $L^2(\mathbb{P}_n)$-seminorm $\|\cdot\|_n$ by Hoeffding's inequality. Formally, for any $f, g \in \mathcal{F}_n$,

$$\|\mathbb{G}_n^o f - \mathbb{G}_n^o g\|_{\phi,n} \leq \|f-g\|_n \equiv \sqrt{\frac{1}{n}\sum_{i=1}^n |f(Z_i)-g(Z_i)|^2}.$$

The value $\eta_n = \sup_{f\in\mathcal{F}_n}\|f\|_n$ is an upper bound for the radius of $\mathcal{F}_n \cup \{0\}$ with respect to this norm. The maximal inequality Theorem 2.2.4 in Vaart and Wellner (1996) gives

$$\begin{aligned}\left\|\sup_{f\in\mathcal{F}_n}|\mathbb{G}_n^o f|\right\|_{\phi,n} &\leq K_\phi \int_0^{\eta_n}\sqrt{1+\log N\left(\epsilon, \mathcal{F}_n, L^2(\mathbb{P}_n)\right)}\mathrm{d}\epsilon \\ &\leq K_\phi\|F\|_n\int_0^1\sqrt{1+\log N\left(\epsilon\|F\|_n, \mathcal{F}_n, L^2(\mathbb{P}_n)\right)}\mathrm{d}\epsilon.\end{aligned}$$

where $K_\phi$ is a universal constant only depending on the function $\phi$; see Theorem 2.2.4 in Vaart and Wellner (1996) for more details. By Problem 2.2.5 in Vaart and Wellner (1996), we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_n}|\mathbb{G}_n^o f|^2\,\middle|\,\{Z_i\}_{i=1}^n\right] \leq 4\log 2\left\|\sup_{f\in\mathcal{F}_n}|\mathbb{G}_n^o f|\right\|_{\phi,n}^2.$$

Consequently, applying Jensen's inequality gives that there is a constant $K > 0$ not depending on $n$ such that for all $n \in \mathbb{N}^+$,

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}_n}|\mathbb{G}_n^o f|\,\middle|\,\{Z_i\}_{i=1}^n\right] &\leq K\|F\|_n\int_0^1\sqrt{1+\log N\left(\epsilon\|F\|_n, \mathcal{F}_n, L^2(\mathbb{P}_n)\right)}\mathrm{d}\epsilon \\ &\leq K\|F\|_n\sup_Q\int_0^1\sqrt{1+\log N\left(\epsilon\|F\|_{Q,2}, \mathcal{F}_n, L^2(Q)\right)}\mathrm{d}\epsilon \\ &\leq K\|F\|_n\sqrt{\mathrm{VC}(\Pi_n)},\end{aligned}$$

where $K > 0$ is a constant independent of $n$. Taking expectation on both hand sides and applying Lemma 2.3.1 in Vaart and Wellner (1996) gives

$$\begin{aligned}\mathbb{E}\left[\sup_{\psi\in\mathcal{F}_n}\left|\left(\mathbb{E}_n-\mathbb{E}\right)\psi\right|\right] &\leq 2n^{-1/2}\mathbb{E}\left[\sup_{f\in\mathcal{F}_n}\left|\mathbb{G}_n^o f\right|\right]\\ &\lesssim \sqrt{\mathrm{VC}(\Pi_n)/n}\mathbb{E}\sqrt{\frac{1}{n}\sum_{i=1}^n |F(Z_i)|^2}\\ &\leq \|F\|_{\mathbb{P},2}\sqrt{\mathrm{VC}(\Pi_n)/n},\end{aligned}$$

where the last step follows from Jensen's inequality. The desired result follows. □

# D Proofs for the Identification under Multi-Valued Treatment

Before presenting the formal proofs of Proposition B.1 and Theorem B.1, we first establish a more general result that extends Proposition 1B of Dorn and Guo (2023) to the multi-valued treatment setting. As shown below, Proposition B.1 follows as a direct corollary of Lemma D.1.

**Lemma D.1.** For any $a \in \mathcal{A}$ and any random variable $E \in (0,1)$ satisfying

$$\mathbb{E}\left[\mathbb{1}\{A=a\}/E|X=x\right]=1,$$

and

$$\frac{\left(e_a(X)+\left[1-e_a(X)\right]/\Lambda\right)\mathbb{1}\{A=a\}}{e_a(X)} \leq \frac{\mathbb{1}\{A=a\}}{E} \leq \frac{\left(e_a(X)+\left[1-e_a(X)\right]\Lambda\right)\mathbb{1}\{A=a\}}{e_a(X)}, \tag{D.1}$$

we can construct random variables $(X, Y(a_1), Y(a_2), \ldots, Y(a_d), A, U)$ defined on the same probability space as $(X, Y, A, E)$ and an associated putative propensity score $e_a(x,u) = \mathbb{P}[A=a|X=x, U=u]$ such that

(1) $Y = \sum_{j=1}^d \mathbb{1}\{A=a_j\}Y(a_j)$.

(2) $(Y(a_1), Y(a_2), \ldots, Y(a_d)) \perp\!\!\!\perp A \mid (X,U)$ and

$$e_a(X)/\left(e_a(X)+\left[1-e_a(X)\right]\Lambda\right) \leq e_a(X,U) \leq e_a(X)/\left(e_a(X)+\left[1-e_a(X)\right]/\Lambda\right).$$

(3) $\mathbb{1}\{A=a\}/e_a(X,U) = \mathbb{1}\{A=a\}/E$.

*Proof of Lemma D.1.* Define the following conditional distribution functions:

$$F(y|x,a)=\mathbb{P}(Y\leq y|X=x,A=a),$$

$$G(y|x,a,e)=\mathbb{P}(Y\leq y|X=x,A=a,E=e),$$

$$H(e|x,a)=\mathbb{P}(E\leq e|X=x,A=a),$$

$$K_a(u|x)=\int_{-\infty}^{u}\frac{e_a(x)}{1-e_a(x)}\frac{1-e}{e}\mathrm{d}H(e|x,a).$$

Given $\mathbb{E}\left[\mathbb{1}\{A=a\}/E \mid X=x\right]=1$, $K_a(u|x)$ is a valid CDF for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$. For simplicity, we provide the construction for $a=a_1$; identical arguments apply to other treatments. Let $V, V_{a_1}, V_{a_2}, \ldots, V_{a_d}$ be i.i.d. Uniform$[0,1]$ random variables, independent of $(X,Y,A,E)$. We

construct $Y(a_1), Y(a_2), \ldots, Y(a_d)$ and $U$ as:

$$Y(a_1) = \mathbb{1}\{A = a_1\}Y + \mathbb{1}\{A \neq a_1\}G^{-1}(V_{a_1}|X, a_1, U),$$

$$Y(a_j) = \mathbb{1}\{A \neq a_j\}F^{-1}(V_{a_j}|X, a_j) + \mathbb{1}\{A = a_j\}Y \quad \text{for} \quad j \in \{2, 3, \ldots, d\},$$

$$U = \mathbb{1}\{A = a_1\}E + \mathbb{1}\{A \neq a_1\}K_{a_1}^{-1}(V|X).$$

We verify that the constructed random variables satisfy the three required properties. Property (1) follows directly from the construction. We now proceed to examine the remaining properties. To verify property (2), we consider the following derivation. Conditional on $X, U, A = a_1$, we have

$$\begin{aligned}
&\mathbb{P}(Y(a_1) \leq y_1, Y(a_2) \leq y_2, \ldots, Y(a_d) \leq y_d|X, U, A = a_1) \\
=&\mathbb{P}\left(Y \leq y_1, F^{-1}(V_{a_2}|X, a_2) \leq y_2, \ldots, F^{-1}(V_{a_d}|X, a_d) \leq y_d|X, U, A = a_1\right) \\
=&\mathbb{P}\left(Y \leq y_1|X, U, A = a_1\right)\prod_{j=2}^{d}\mathbb{P}\left(F^{-1}(V_{a_j}|X, a_j) \leq y_j|X, U, A = a_1\right) \\
=&G(y_1|X, a_1, U)\prod_{j=2}^{d}F(y_j|X, a_j).
\end{aligned}$$

Conditional on $X, U, A = a_j$ for $j \neq 1$, we have

$$\begin{aligned}
&\mathbb{P}(Y(a_1) \leq y_1, Y(a_2) \leq y_2, \ldots, Y(a_d) \leq y_d|X, U, A = a_j) \\
=&\mathbb{P}\left(G^{-1}(V_{a_1}|X, a_1, U) \leq y_1, F^{-1}(V_{a_2}|X, a_2) \leq y_2, \ldots, Y \leq y_j, \ldots, F^{-1}(V_{a_d}|X, a_d) \leq y_d|X, U, A = a_j\right) \\
=&\mathbb{P}\left(G^{-1}(V_{a_1}|X, a_1, U) \leq y_1|X, U, A = a_j\right)\prod_{\ell=2, \ell\neq j}^{d}\mathbb{P}\left(F^{-1}(V_{a_\ell}|X, a_\ell) \leq y_\ell|X, U, A = a_\ell\right) \\
&\times \mathbb{P}\left(Y \leq y_j|X, U, A = a_j\right) \\
=&G(y_1|X, a_1, U)\prod_{\ell=2, \ell\neq j}^{d}F(y_\ell|X, a_\ell) \times \mathbb{P}\left(Y \leq y_j|X, K^{-1}(V|X), A = a_j\right) \\
=&G(y_1|X, a_1, U)\prod_{j=2}^{d}F(y_j|X, a_j),
\end{aligned}$$

Combining two expression above gives $(Y(a_1), Y(a_2), \ldots, Y(a_d)) \perp\!\!\!\perp A \mid (X, U)$. By Bayes' rule,

$$\begin{aligned}
e_{a_1}(x, u) =& e_{a_1}(x)\frac{\mathrm{d}\mathbb{P}(u|X = x, A = a_1)}{\mathrm{d}\mathbb{P}(u|X = x)} \\
=& e_{a_1}(x)\frac{\mathrm{d}\mathbb{P}(u|X = x, A = a_1)/\mathrm{d}H(u|x, a_1)}{\mathrm{d}\mathbb{P}(u|X = x)/\mathrm{d}H(u|x, a_1)} \\
=& \frac{e_{a_1}(x)}{e_{a_1}(x) + \frac{e_{a_1}(x)}{1-e_{a_1}(x)}\frac{1-u}{u}\left(1 - e_{a_1}(x)\right)} \\
=& u.
\end{aligned}$$

Since the support of $K_{a_1}(\cdot|x)$ is contained in that of $H(\cdot|x, a_1)$, Eq. (D.1) implies

$$e_{a_1}(X)/\left(e_{a_1}(X) + [1 - e_{a_1}(X)]\,\Lambda\right) \leq U \leq e_{a_1}(X)/\left(e_{a_1}(X) + [1 - e_{a_1}(X)]/\Lambda\right) \quad \text{a.s.}$$

Finally, property (3) is also easy to verify. The event $\mathbb{1}\{A = a_1\}$ implies $U = E$, so

$$\mathbb{1}\{A = a_1\}/e_{a_1}(X, U) = \mathbb{1}\{A = a_1\}/U = \mathbb{1}\{A = a_1\}/E.$$

□

*Proof of Proposition B.1.* The proof follows directly from that of Lemma D.1, using a similar argument as in Proposition A.1, and is therefore omitted for brevity. □

We now turn to proof Theorem B.1.

*Proof of Theorem B.1.* According to Proposition B.1, for any $Q \in \mathcal{P}_{\mathrm{M}}$ and $(x, a) \in \mathcal{X} \times \mathcal{A}$, we have

$$\mu_Q(x, a) \geq \inf_{Q \in \mathcal{P}_{\mathrm{M}}} \mu_Q(x, a) = \mu^-(x, a).$$

It follows that

$$\mathbb{E}\left[\sum_{a \in \mathcal{A}} \mu_Q(X, a)\pi_a(X)\right] \geq \mathbb{E}\left[\sum_{a \in \mathcal{A}} \mu^-(X, a)\pi_a(X)\right]$$

Since $Q \in \mathcal{P}_{\mathrm{M}}$ is arbitrary, the inequality remains valid after taking the infimum over $Q \in \mathcal{P}_{\mathrm{M}}$ on both sides. Therefore, we conclude that

$$\begin{aligned} W(\pi) &= \inf_{Q \in \mathcal{P}_{\mathrm{M}}} \mathbb{E}_Q\left[\sum_{a \in \mathcal{A}} Y(a)\pi_a(X)\right] = \inf_{Q \in \mathcal{P}_{\mathrm{M}}} \mathbb{E}\left[\sum_{a \in \mathcal{A}} \mu_Q(X, a)\pi_a(X)\right] \\ &\geq \mathbb{E}\left[\sum_{a \in \mathcal{A}} \mu^-(X, a)\pi_a(X)\right]. \end{aligned}$$

□