EFFICIENT AND SHARP OFF-POLICY LEARNING UNDER UNOBSERVED CONFOUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a novel method for personalized off-policy learning in scenarios with unobserved confounding. Thereby, we address a key limitation of standard policy learning: standard policy learning assumes unconfoundedness, meaning that no unobserved factors influence both treatment assignment and outcomes. However, this assumption is often violated, because of which standard policy learning produces biased estimates and thus leads to policies that can be harmful. To address this limitation, we employ causal sensitivity analysis and derive a *semi-parametrically efficient estimator for a sharp bound on the value function under unobserved confounding*. Our estimator has three advantages: (1) Unlike existing works, our estimator avoids unstable minimax optimization based on inverse propensity weighted outcomes. (2) Our estimator is semi-parametrically efficient. (3) We prove that our estimator leads to the optimal confounding-robust policy. Finally, we extend our theory to the related task of policy improvement under unobserved confounding, i.e., when a baseline policy such as the standard of care is available. We show in experiments with synthetic and real-world data that our method outperforms simple plug-in approaches and existing baselines. Our method is highly relevant for decision-making where unobserved confounding can be problematic, such as in healthcare and public policy.

1 INTRODUCTION

Policy learning is crucial in many areas such as healthcare (Feuerriegel et al., 2024), education (Chan, 2023), and public policy (Ladi & Tsarouhas, 2020). However, collecting data through randomized experiments is often either infeasible or unethical. Instead, methods are needed that use observational data to inform decision-making. Here, we focus on *off-policy learning* to optimize decision policies from observational data (Athey & Wager, 2021).

The reliability of standard off-policy learning is compromised when *unobserved confounding* is present (Kallus et al., 2019). Unobserved confounding arises when factors affect both treatment choices and outcomes but are not recorded (Pearl, 2009). For example, the race of a patient may – unfortunately – affect the access to treatments (Obermeyer et al., 2019), yet race is typically not recorded in patient records. Hence, standard off-policy learning that relies on the assumption of no unobserved confounding will lead to *biased* estimates and may even generate *harmful* policies.

As a remedy, *confounding-robust policy learning* aims to find the optimal policy under worst-case unobserved confounding. This is typically achieved using the marginal sensitivity model (MSM) (Tan, 2006), a framework from causal sensitivity analysis that bounds the effect of unobserved confounding. However, the existing method for confounding-robust policy learning under the MSM (see (Kallus & Zhou, 2018a) for the conference paper and (Kallus & Zhou, 2021) for the journal version) has notable shortcomings. First, it must numerically optimize the worst-case effect on the regret function due to unobserved confounding. Such minimax optimization is based on inverse propensity weighted outcomes and hence unstable. Second, this method is statistically suboptimal: it lacks the property of semi-parametric efficiency and thus suffers from suboptimal variance properties.

In this paper, we address the above shortcomings by developing a novel *semi-parametrically efficient and sharp estimator for personalized off-policy learning under unobserved confounding*. Here, semi-parametric efficiency means the unbiased estimator with the lowest possible variance. Our key novelties are the following: (i) We derive a *closed-form expression* for a *sharp bound on the*

value function of a candidate policy under unobserved confounding.[1] As a result, we can thereby directly minimize our closed-form bound and, unlike existing works, avoid an unstable minimax optimization based on inverse propensity weighted outcomes. (ii) We propose an estimator that is semi-parametrically efficient. Hence, our estimator is the first to achieve the *lowest variance* among all unbiased estimators for our task.

Methodologically, we proceed as follows. We first derive a *sharp bound* on the value function for scenarios with unobserved confounding and, hence, avoid the unstable minimax optimization as in other methods. We then propose a novel *one-step bias-corrected estimator* to achieve semi-parametric efficiency and thus guarantee that our estimator has the lowest variance among all unbiased estimators. For this, we derive the corresponding efficient influence function of the sharp bound on the value function. We finally provide theoretical guarantees that minimizing our estimated sharp bound on the value function ensures that our method yields the *optimal* confounding-robust policy. Such guarantees are particularly crucial in high-stakes applications such as medicine or public policy, where unreliable policies can lead to harmful consequences.

Our work makes the following **contributions**[2]: (i) We propose a novel *efficient estimator for our sharp bound on the value function*. (ii) We derive an estimator for our bounds that is *semi-parametrically efficient*. (iii) We generalize our theoretical findings to the related task of *confounding-robust policy improvement*. (iv) Through extensive experiments using synthetic and real-world datasets, we show that our method consistently *outperforms* simple plug-in estimators and existing baselines.

2 Related work

We provide an overview of three literature streams particularly relevant to our work, namely, standard off-policy learning (i) with and (ii) without unobserved confounding as well as (iii) causal sensitivity analysis. We provide an extended related work in Appendix A (where we distinguish our work from other streams such as, e.g., unobserved confounding in reinforcement learning).

	Robust under unobserved conf.	Sharp bounds	Discrete treatments	Efficient for policy learning
Oprescu et al. (2023)	✓	✓	✗	✗
Swaminathan & Joachims (2015)	✗	✗	✓	✗
Athey & Wager (2021); Dudik et al. (2011)	✗	✗	✓	✓
Kallus & Zhou (2018a; 2021)	✓	✗	✓	✗
Efficient & sharp (ours)	✓	✓	✓	✓

Table 1: Overview of related methods. Oprescu et al. (2023) is designed for a different task, and standard policy learners ignore the issue of unobserved confounding. Only Kallus & Zhou (2018a; 2021) deals with our setting, but provides neither sharp bounds nor an efficient estimator. Only our work can deal with unobserved confounding, discrete treatments, provides sharp bounds, and an efficient estimator.

Off-policy learning under unconfoundedness: Off-policy learning aims to optimize the policy value, which needs to be estimated from data. For this, there are three major approaches: (i) the direct method (DM) (Qian & Murphy, 2011) leverages estimates of the response functions; (ii) inverse propensity weighting (IPW) (Swaminathan & Joachims, 2015) re-weights the data such that in order to resemble samples under the evaluation policy; and (iii) the doubly robust method (DR) (Athey & Wager, 2021; Dudik et al., 2011). The latter is based on the efficient influence function of the policy value (Robins et al., 1994) and is asymptotically efficient (Chernozhukov et al., 2018; van der Vaart, 1998).

Several works aim at improving the finite sample performance of these methods, for instance, via re-weighting (Kallus, 2018; 2021) or targeted maximum likelihood estimation (Bibaut et al., 2019). Further, several methods have been proposed for off-policy learning in specific settings involving, for example, distributional robustness (Kallus et al., 2022), fairness (Frauen et al., 2024b), interpretability (Tschernutter et al., 2022), and continuous treatments (Kallus & Zhou, 2018b; Schweisthal et al., 2023). However, all of the works assume unconfoundedness and, therefore, do **not** account for unobserved confounding.

Off-policy learning under unobserved confounding: In scenarios with unobserved confounding, standard approaches for off-policy learning are biased (Kallus & Zhou, 2018a; 2021), which can lead

[1] We use the term "sharp" as in earlier work from causal sensitivity analysis (Frauen et al., 2023): a valid upper (lower) bound of a causal quantity is *sharp* if there does not exist another valid bound that is strictly smaller (larger).

[2] Code is available at `https://anonymous.4open.science/r/CBCE` and will be released upon acceptance of the paper.

to harmful decisions. The reason is that, under unobserved confounding, the policy value *cannot* be identified from observational data. As a remedy, previous works leverage causal sensitivity analysis or related methods to obtain bounds on the unidentified policy value (Bellot & Chiappa, 2024; Guerdan et al., 2024; Huang & Wu, 2024; Joshi et al., 2024; Namkoong et al., 2020; Zhang & Bareinboim, 2024), which can then be used to learn an optimal worst-case policy. Optimizing such bounds is often termed "confounding-robust policy learning". However, these works do *not* consider sharp bounds under unobserved confounding and do *not* provide semi-parametrically efficient estimators.

Closest to our work is (Kallus & Zhou, 2018a) with an extended version published in (Kallus & Zhou, 2021). Therein, the authors propose a method for confounding-robust policy improvement, yet with two notable shortcomings: (i) it is *not* based on closed-form solutions for the bounds, and (ii) it is *not* based on a semi-parametrically efficient estimator of these bounds. Therefore, Kallus & Zhou (2018a; 2021) require solving a minimax optimization problem that is relies in inverse propensity weighted outcomes, which is *unstable*. Further, their estimator is suboptimal because it fails to achieve semi-parametric efficiency, meaning it does *not* achieve the lowest-possible variance among all unbiased estimators.

Causal sensitivity analysis: Causal sensitivity analysis (Cornfield et al., 1959) allows practitioners to account for unobserved confounding by using so-called sensitivity models (Jin et al., 2022; Rosenbaum, 1987), which incorporate domain knowledge on the strength of unobserved confounding. As a result, the sensitivity model allows to obtain bounds on a causal quantity of interest, which, if the sensitivity model is correctly specified, can then be used for consequential decision-making (Jesson et al., 2021).

A prominent sensitivity model is the MSM (Tan, 2006). The MSM gained popularity in recent years and, for instance, was used to obtain bounds on the conditional average treatment effect (CATE) through machine learning (Jesson et al., 2021; Kallus et al., 2019; Yin et al., 2022). Only recently, sharp bounds on the CATE have been derived (Bonvini et al., 2022; Dorn & Guo, 2022; Frauen et al., 2024a; 2023; Jin et al., 2023). Other works have considered the estimation of such bounds (Dorn et al., 2024; Oprescu et al., 2023). However, these works only consider causal sensitivity analysis for CATE but **not** policy learning. Further, their works are limited to **binary** treatments. In contrast, our method can handle *discrete* treatments, and, therefore, requires entirely different influence functions.

Research gap: To the best of our knowledge, we are the first to derive a *semi-parametrically efficient estimator for a sharp bound on the value function using the MSM*. Thereby, we enable optimal confounding-robust off-policy learning.

3 Problem Setup

Data: Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be our outcome of interest, such as the health condition of a patient. We follow the convention that w.l.o.g. *lower* values correspond to *better* outcomes (Kallus & Zhou, 2018a). Further, let $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ denote covariates that contain additional information, such as age, gender, or disease-related information. Finally, let $A \in \mathcal{A} = \{0, 1, \dots, d_a - 1\}$ be the assigned treatment (or action). Note that we do not restrict our setting to binary treatments but allow for arbitrary, discrete treatments. For the product space, we use $\mathcal{D} = \mathcal{Y} \times \mathcal{X} \times \mathcal{A}$. In the following, we assume that we have access to an observational dataset $\mathcal{D}_n = \{(Y_i, X_i, A_i)\}_{i=1}^n$ that consists of n i.i.d. copies of $(Y, X, A) \in \mathcal{D}$.

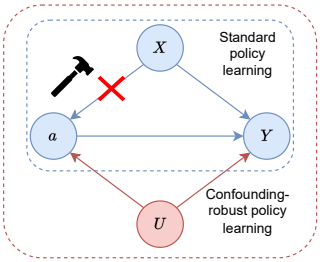

Figure 1: We can only block backdoor paths for observed confounders X. Hence, under unobserved confounding U, we cannot point-identify the potential outcome $Y[a]$ and related quantities such as the value function $V(\pi)$.

Policy value: Policy learning aims to find the best policy for assigning treatments, given covariates. Formally, a *policy* $\pi(a \mid x)$ is a conditional probability mass function $\pi : \mathcal{A} \times \mathcal{X} \to [0, 1]$ with $\sum_{a \in \mathcal{A}} \pi(a \mid x) = 1$, corresponding to the probability of receiving treatment $A = a$, given covariates $X = x$. The *value* $V(\pi)$ of a policy is defined as $V(\pi) = \mathbb{E}\Big[\sum_{a \in \mathcal{A}} \pi(a \mid X) Y[a]\Big]$, where $Y[a]$ denotes the potential outcome for Y when intervening on treatment $A = a$ (Neyman, 1923; Rubin, 1978). Hence, the policy value $V(\pi)$ is the expected average potential outcome when adhering to the policy π.

Standard off-policy learning: Off-policy learning aims to find a policy π that has the best policy value among all $\pi \in \Pi$ for some policy class Π. Of note, it is standard in the literature (Frauen et al., 2024b; Hatt et al., 2022; Kallus & Zhou, 2018a) to restrict the analysis to policy classes Π with finite complexity such as neural networks.

The value function is identifiable under the following three assumptions (Rubin, 1978): (i) *Consistency:* $Y[A] = Y$; (ii) *Positivity:* $0 < p(A = a \mid X = x) < 1 \; \forall \; a \in \mathcal{A}, x \in \mathcal{X}$; (iii) *Unconfoundedness:* $Y[a] \perp\!\!\!\perp A \mid X \; \forall \; a \in \mathcal{A}$. Then, the policy value is identified from the observational data via $V(\pi) = \mathbb{E}\Big[\sum_{a \in \mathcal{A}} \pi(a \mid X) Q(a, X)\Big]$, where $Q(a, x) = \mathbb{E}[Y \mid X = x, A = a]$ is the conditional average potential outcome function.

The optimal policy can then be learned via

$$\pi^*_{\text{standard}} = \arg\min_{\pi \in \Pi} \hat{V}(\pi), \tag{1}$$

where $\hat{V}(\pi)$ is an estimator of the identified policy value. Recall that we follow the convention in (Kallus & Zhou, 2018a; 2021) that lower Y are better, so we aim to *minimize* the value function.

Allowing for unobserved confounding: The assumption of (iii) *unconfoundedness* is problematic and often unrealistic (Hemkens et al., 2018): Unconfoundedness requires that the observed covariates X capture *all* factors that affect both treatment choice and outcome. In this work, we do **not** rely on the *unconfoundedness* assumption, which is restrictive and oftentimes unrealistic. Instead, we allow for *unobserved confounding*, which we denote by a random variable $U \in \mathcal{U} \subset \mathbb{R}$ (see Figure 2).

Importantly, under unobserved confounding, we cannot point-identify the value function $V(\pi)$. Instead, we aim to *partially* identify the value function $V(\pi)$ by leveraging causal sensitivity analysis. Specifically, we adopt the MSM (Tan, 2006) to bound the ratio between the *nominal propensity score*

$$e(a, x) = p(A = a \mid X = x), \tag{2}$$

which can be estimated from data $\mathcal{D}_n$ and the *true propensity score*

$$e(a, x, u) = p(A = a \mid X = x, U = u), \tag{3}$$

which is fundamentally unobserved. Formally, the MSM assumes

$$\Gamma^{-1} \leq \frac{e(a, x)}{1 - e(a, x)} \frac{1 - e(a, x, u)}{e(a, x, u)} \leq \Gamma \tag{4}$$

for some $\Gamma \geq 1$ that can be chosen by domain domain knowledge (Frauen et al., 2023; Kallus et al., 2019) or data-driven heuristics (Hatt et al., 2022) (see Supplement B).

Intuitively, Γ close to 1 implies that the impact of unobserved variables U on the treatment decision is small, whereas a large Γ means that observed variables X do not contain sufficient information to fully capture the treatment decision. In particular, $\Gamma = 1$ implies that the true propensity score coincides with the nominal propensity score. Hence, there is no unobserved confounding and our scenario simplifies to the naïve unconfoundedness setting. Conversely, if we let $\Gamma > 1$, the true and the nominal propensity scores differ, and, therefore, we account for additional unobserved confounding.

Formally, the marginal sensitivity model gives rise to a set of distributions $\mathcal{P}(\Gamma)$ over $\mathcal{D} \times \mathcal{U}$ that are compatible with the constraints in Equation 4. This set is defined as

$$\mathcal{P}(\Gamma) = \left\{ \tilde{p} \in \mathcal{P}(\mathcal{D} \times \mathcal{U}) : \int_{\mathcal{U}} \tilde{p}(d, u) \, \mathrm{d}u = p(d) \forall d \in \mathcal{D}, \; \Gamma^{-1} \leq \frac{\tilde{e}(A, X)}{1 - \tilde{e}(A, X)} \frac{1 - \tilde{e}(A, X, U)}{\tilde{e}(A, X, U)} \leq \Gamma \right\}, \tag{5}$$

where $\mathcal{P}(\mathcal{D} \times \mathcal{U})$ is the set of all possible joint distributions of the observables and the unobserved confounders, and where the nominal propensity score $\tilde{e}(A, X)$ and the true propensity score $\tilde{e}(A, X, U)$ result from $\tilde{p}$ as in Equation 2 and Equation 3, respectively.

Different from standard off-policy learning under the unconfoundedness assumption, we can *not* point-identify the value function, and, hence, optimizing the objective in Equation 1 is *biased*. Instead, we need to account for the worst-case scenario that can occur under unobserved confounding. That is, we are interested in: *Which policy yields the optimal value under the worst-case unobserved confounding, given our sensitivity constraints?*

Objective: Formally, the optimal confounding-robust policy π^* is the solution to the minimax problem

$$\pi^* = \arg\min_{\pi \in \Pi} \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} V(\pi). \tag{6}$$

However, the existing method (Kallus & Zhou, 2018a; 2021) for our task has key limitations: (i) It requires directly solving the minimax optimization problem, which can be unstable due to inverse propensity weighting. Instead, we later derive a *closed-form expression for the inner supremum* (i.e., an upper bound), which reduces Equation 6 to a simple minimization task. (ii) This method is *not* semi-parametrically efficient, thus leading to suboptimal finite-sample performance. As a remedy, we later derive an estimator that is *semi-parametrically efficient*.

4 Sharp bounds and efficient estimation

In this section, we introduce our estimator for sharp bounds of the value function under unobserved confounding. For this, we first derive a *closed-form solution* for the sharp bounds of the value function (Section 4.1), which directly solves the inner maximization in Equation 6. Then, we present our estimator for these bounds (Section 4.2), which is based on non-trivial derivations of the efficient influence function to offer *semi-parametric efficiency*. Further, we provide *learning guarantees* when optimizing the bounds of the value function (Section 4.3). Finally, we propose an *extension of our method* for scenarios where the aim is to optimize the relative improvement of a policy over a given baseline policy such as the standard of care in medicine (Section 4.4).

4.1 Sharp bounds for the value function

We now derive our sharp bound for the value function under unobserved confounding, given our sensitivity constraints $\mathcal{P}(\Gamma)$ in Equation 5. Recall that our aim is to *minimize* the value function $V(\pi)$, and, hence, we are interested in an *upper bound* for $V(\pi)$. That is, we seek to find the value function in the worst-case confounding scenario under the MSM, which is given by $V^{+,*}(\pi) = \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} V(\pi)$.

By definition, a closed-form solution to this maximization problem ensures that (i) the bound is *valid*, i.e., $V^{+,*}(\pi) \geq V(\pi)$ for all $\tilde{p} \in \mathcal{P}(\Gamma)$, and that (ii) the bound is *sharp*, i.e., there does **not** exist a valid upper bound $V^{+,\dagger}(\pi)$ such that $V^{+,\dagger}(\pi) < V^{+,*}(\pi)$.

In order to derive $V^{+,*}(\pi)$, we first introduce the *conditional average potential outcome* function

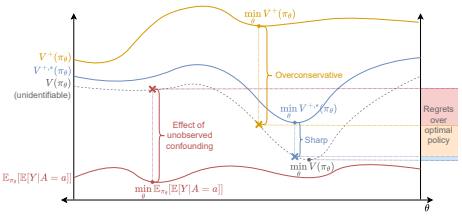

Figure 2: Under unobserved confounding, the value function is unidentifiable, and the ground-truth optimal policy is unknown. Ignoring unobserved confounding can lead to a policy with *large regret*, and may introduce harm. Further, optimizing w.r.t. a suboptimal bound can lead to an *overconservative* policy. Instead, we seek to find the optimal confounding-robust policy by minimizing a *sharp bound* on the worst-case effect of unobserved confounding.

$$Q(a, x) = \mathbb{E}[Y[a] \mid X = x], \tag{7}$$

which is the expected potential outcome for treatment $A = a$, given covariate information $X = x$. Importantly, because we do **not** make Assumption (iii) of *unconfoundedness*, the quantity $Q(a, x)$ is not point-identified.

We now state our first theorem, which provides a sharp upper bound of the value function under our sensitivity constraints $\mathcal{P}(\Gamma)$. Further, we also provide the sharp lower bound $V^{-,*} = \inf_{\tilde{p} \in \mathcal{P}(\Gamma)} V(\pi)$, which we later need for our extensions in Section 4.4.

Proposition 4.1. *Let $Q^{+,*}(a, x) = \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} Q(a, x)$ and $Q^{-,*}(a, x) = \inf_{\tilde{p} \in \mathcal{P}(\Gamma)} Q(a, x)$ be the sharp upper and lower bound for the conditional average potential outcome, respectively, given our sensitivity constraints $\mathcal{P}(\Gamma)$. Then, the sharp upper bound $\sup_{\tilde{p} \in \mathcal{P}(\Gamma)} V(\pi) = V^{+,*}(\pi)$ and the sharp lower bound $\inf_{\tilde{p} \in \mathcal{P}(\Gamma)} V(\pi) = V^{-,*}(\pi)$ for the value function $V(\pi)$ are given by*

$$V^{\pm,*}(\pi) = \int_{\mathcal{X}} \sum_{a} Q^{\pm,*}(a, x) \pi(a \mid x) \, \mathrm{d}p(x). \tag{8}$$

Proof. See Supplement D.1. □

Our above derivation of the closed-form solution has a crucial advantage over existing works (Kallus & Zhou, 2018a; 2021): we avoid an unstable minimax optimization that is based on inverse propensity weighted outcomes, and, instead, we can directly work with $V^{+,*}(\pi)$, which simplifies Equation 6 to $\pi^* = \arg\min_{\pi\in\Pi} V^{+,*}(\pi)$. As a result, we have reduced the original minimax problem to a much simpler *minimization task*.

4.2 Semi-parametrically efficient estimator for the sharp upper bound

In this section, we derive a semi-parametrically efficient estimator for our sharp upper bound $V^{+,*}(\pi)$ of the value function $V(\pi)$. semi-parametrically efficient estimators are desirable because they achieve the *lowest possible variance among all unbiased estimators* (Hines et al., 2022; Kennedy, 2022).

In order to derive such an estimator of $V^{+,*}(\pi)$, we first need to decompose the estimand $Q^{\pm,*}(a,x)$ in Proposition 4.1.

Definition 4.2 ((Dorn & Guo, 2022; Frauen et al., 2023))**.** Sharp bounds $Q^{\pm,*}(a,x)$ of the conditional average potential outcome $Q(a,x)$ function are given by

$$Q^{\pm,*}(a,x) = c^{\mp}(a,x)\underline{\mu}^{\pm}(a,x) + c^{\pm}(a,x)\bar{\mu}^{\pm}(a,x), \tag{9}$$

where we let

$$c^{\pm}(a,x) = b^{\pm}e(a,x) + \Gamma^{\pm 1},\ b^{\pm} = (1-\Gamma^{\pm 1}) \tag{10}$$

and

$$\underline{\mu}^{\pm}(a,x) = \mathbb{E}[Y\underline{\Delta}^{\pm}(Y,A,X) \mid X=x, A=a], \quad \bar{\mu}^{\pm}(a,x) = \mathbb{E}[Y\bar{\Delta}^{\pm}(Y,A,X) \mid X=x, A=a] \tag{11}$$

with

$$\underline{\Delta}^{\pm}(y,a,x) = \mathbb{1}_{\{y\le F^{-1}_{x,a}(\alpha^{\pm})\}}, \qquad \bar{\Delta}^{\pm}(y,a,x) = \mathbb{1}_{\{y\ge F^{-1}_{x,a}(\alpha^{\pm})\}}, \tag{12}$$

where $\alpha^+ = \Gamma/(1+\Gamma)$ and $\alpha^- = 1/(1+\Gamma)$, and where $F^{-1}_{x,a}(q)$ is the conditional quantile function

$$F^{-1}_{x,a}(q) = \inf\{y\in\mathcal{Y}:\ p(Y\le y \mid X=x, A=a)\ge q\}. \tag{13}$$

In order to achieve semi-parametric efficiency for the sharp upper bound $V^{+,*}(\pi)$, we need to carefully take into account the nuisance functions in Equation 8 and Equation 9, respectively. That is, the key difficulty lies in that $V^{+,*}(\pi)$ depends on several nuisance functions

$$\eta = \{e(a,x), F^{-1}_{a,x}(\alpha^{\pm}), \bar{\mu}^{\pm}(a,x), \underline{\mu}^{\pm}(a,x)\}. \tag{14}$$

If we followed a naïve plug-in approach (i.e., if we estimated $\hat{\eta}$ from data $\mathcal{D}_n$ and plugged them into Equation 8 and thus Equation 9), our final estimator $\hat{V}^{+,*}(\pi)$ would suffer from **first-order bias** due to estimation errors in the nuisance functions.

As a remedy, we present a *one-step bias-corrected* estimator. That is, we estimate the first-order bias and subtract it from our plug-in estimate (Kennedy, 2022; van der Vaart, 1998). To obtain a one-step bias-corrected estimator for our task, we need to make non-trivial derivations of the efficient influence function of $V^{+,*}(\pi)$ below. In the following, we let $\mathbb{P}_n\{\cdot\}$ denote the *sample average* for a dataset $\mathcal{D}_n$. Further, we use the short notation $f_x = f(x)$ for any function $f(\cdot)$ to improve readability.

Theorem 4.3. *An estimator for the sharp upper bound of the value function is given by*

$$\begin{aligned}\hat{V}^{+,*}(\pi) = \mathbb{P}_n\Big\{ &\sum_a \pi_{a,X}\Big[\hat{Q}^{+,*}_{a,X} - \hat{e}_{a,X}\Big(b^-\hat{\underline{\mu}}^+_{a,X} + b^+\hat{\bar{\mu}}^+_{a,X}\Big)\Big] + \pi_{A,X}\Big(b^-\hat{\underline{\mu}}^+_{A,X} + b^+\hat{\bar{\mu}}^+_{A,X}\Big) \\ &+ \frac{\pi_{A,X}}{\hat{e}_{A,X}}\Big[\Big(\hat{c}^-_{A,X} - \hat{c}^+_{A,X}\Big)\Big(\hat{F}^{-1}_{X,A}(\alpha^+)(\hat{\underline{\Delta}}^+_{Y,A,X} - \alpha^+)\Big) \\ &+ \hat{c}^-_{A,X}\Big(Y\hat{\underline{\Delta}}^+_{Y,A,X} - \hat{\underline{\mu}}^+_{A,X}\Big) + \hat{c}^+_{A,X}\Big(Y\hat{\bar{\Delta}}^+_{Y,A,X} - \hat{\bar{\mu}}^+_{A,X}\Big)\Big]\Big\}.\end{aligned}$$

*Further, the above estimator is **semi-parametrically efficient**.*

Proof. See Supplement D.2. □

We now have a *semi-parametrically efficient estimator* for the sharp upper bound of the value function under unobserved confounding. Algorithm 1 presents a flexible procedure to learn confounding-robust policies for parametric policy classes Π_θ (e.g., neural networks).

4.3 Learning guarantees

In this section, we provide asymptotic learning guarantees in the form of generalization bounds when learning the confounding-robust policy π via our Algorithm 1. Of note, it is *not* obvious that minimizing the *estimated* sharp upper bound $\hat{V}^{+,*}(\pi)$ provides a meaningful, confounding-robust policy π^*. Hence, we provide learning guarantees where we show that, with high probability, *minimizing our estimated sharp upper bound yields the optimal policy*.

For this, we show that minimizing the *estimated* sharp upper bound $\hat{V}^{+,*}(\pi)$ with respect to π indeed minimizes the true, unknown value function $V(\pi)$ on population level. Fortunately, our method only requires one additional assumption, namely, boundedness of the outcome $|Y| \leq C_y$. This is a very mild restriction and reasonable in practice.

Algorithm 1 Confounding-robust policy learning

Input: Data $\mathcal{D}_n = \{(Y_i, A_i, X_i)\}_{i=1}^n$, sensitivity parameter $\Gamma \geq 1$, sample split $\rho \in (0, 1)$, learning rate λ, parametric policy class Π_θ, training iterations K

Output: $\hat{V}^{+,*}(\pi)$

```
Perform sample split D^η_{⌈ρn⌉}, D^{V+,*}_{⌊(1−ρ)n⌋}
Estimate nuisance functions η̂ on D^η_{⌈ρn⌉}
Evaluate η̂ on D^{V+,*}_{⌊(1−ρ)n⌋}
Initialize policy π_θ^(0) ∈ Π_θ
for k = 0 to K − 1 do
    Estimate V^{+,*}(π_θ^(k)) as in (2) (using evaluated η̂)
    Update policy parameters:
        θ^(k+1) ← θ^(k) − λ∇_θ V^{+,*}(π_θ^(k))
end for
Return: Robust policy π* ← π_θ^(K)
```

We express the flexibility of our policy class Π in terms of the Rademacher complexity $\mathcal{R}_n(\pi)$, which is a common choice in the literature (Athey & Wager, 2021; Frauen et al., 2024b; Hatt et al., 2022; Kallus & Zhou, 2018a). Importantly, parametric policy classes $\Pi = \Pi_\theta$ such as neural networks have vanishing Rademacher complexity $\mathcal{R}_n(\Pi) \in \mathcal{O}(n^{-1/2})$.

Theorem 4.4. *Assume Y is bounded by a constant C_y, i.e. $|Y| \leq C_y$. Then, for any policy $\pi \in \Pi$, it holds that*

$$V(\pi) \leq \hat{V}^{+,*}(\pi) + 2C_v\Big(\mathcal{R}_n(\Pi) + \frac{5}{2}\sqrt{\frac{1}{2n}\log\Big(\frac{2}{\delta}\Big)}\Big) \tag{15}$$

with probability $1 - \delta$, where $C_v = 2C_y(1 + \Gamma^{-1} + \Gamma)$ and $\mathcal{R}_n(\pi)$ is the empirical Rademacher complexity of policy class Π.

Proof. See Supplement D.3. □

The above Theorem 4.4 has the following implication: given our sensitivity constraints $\mathcal{P}(\Gamma)$, our estimated sharp upper bound $\hat{V}^{+,*}(\pi)$ correctly bounds the true, unknown value function $V(\pi)$ on population level with high probability. Therefore, given sufficient data $\mathcal{D}_n$, minimizing $\hat{V}^{+,*}(\pi)$ with respect to π also minimizes $V(\pi)$ and, hence, yields the optimal π^*.

In sum, we have derived (i) a novel *sharp upper bound of the value function*, which circumvents unstable minimax optimization based on inverse propensity weighted outcomes. Further, we have proposed (ii) an estimator for this bound that is *semi-parametrically efficient*, i.e., an unbiased estimator with the lowest possible variance. Finally, we have derived (iii) *learning guarantees*, which show that minimizing our estimated bound via Algorithm 1 indeed optimizes the true, unknown population value.

4.4 Extension to policy improvement

Our main results from above focus on optimizing the value function $V(\pi)$, which is common in practice (e.g., Dudik et al., 2011; Hatt et al., 2022). However, in some scenarios, an established baseline policy π_0 may be available; then, one may aim to make a small relative improvement yet with certain guarantees. This setting is commonly termed as *policy improvement* (Kallus & Zhou, 2018a; 2021; Laroche et al., 2019; Thomas et al., 2015). We extend our theory to policy improvement under unobserved confounding in **Supplement C**.

5 EXPERIMENTS

	$\Gamma^* = 2$	$\Gamma^* = 4$	$\Gamma^* = 6$	$\Gamma^* = 8$	$\Gamma^* = 10$	$\Gamma^* = 12$	$\Gamma^* = 14$	$\Gamma^* = 16$
Standard IPW estimator	$\mathbf{-1.31 \pm 0.02}$	-0.60 ± 0.15	-0.09 ± 0.01	-0.07 ± 0.01	-0.06 ± 0.01	-0.06 ± 0.01	-0.05 ± 0.01	-0.03 ± 0.01
Standard DR estimator	-1.30 ± 0.04	-0.71 ± 0.02	-0.18 ± 0.13	-0.07 ± 0.01	-0.07 ± 0.01	-0.06 ± 0.01	-0.05 ± 0.01	-0.04 ± 0.01
Kallus & Zhou (2018a; 2021)	-1.21 ± 0.10	-0.70 ± 0.06	-0.40 ± 0.06	-0.22 ± 0.04	-0.16 ± 0.02	-0.14 ± 0.02	-0.10 ± 0.01	-0.08 ± 0.01
Efficient + sharp estimator (ours)	-1.12 ± 0.08	$\mathbf{-1.00 \pm 0.08}$	$\mathbf{-0.89 \pm 0.13}$	$\mathbf{-0.66 \pm 0.14}$	$\mathbf{-0.64 \pm 0.14}$	$\mathbf{-0.58 \pm 0.17}$	$\mathbf{-0.50 \pm 0.20}$	$\mathbf{-0.30 \pm 0.22}$
Absolute improvement	+0.19	−0.29	−0.49	−0.44	−0.48	−0.44	−0.40	−0.22

Table 2: **Varying confounding strength.** We vary the confounding parameter Γ^* in the DGP along with the sensitivity parameter Γ in both our efficient estimator and the baseline (Kallus & Zhou, 2018a; 2021). Then, we report the regret over a randomized policy (*lower values are better*). As confounding increases, our estimator is the only method that is robust and thus performs best.

In the following, we evaluate the performance of our method against: (1) the minimax optimization approach by Kallus & Zhou (2018a; 2021) and standard methods for policy learning, namely, (2) the IPW estimator (Swaminathan & Joachims, 2015) and (3) the DR estimator (Athey & Wager, 2021; Dudik et al., 2011). Importantly, the approach by Kallus & Zhou (2018a; 2021) is the **only** baseline that can deal with confounding-robust policy learning with the MSM and thus the **only** baseline for our task. To ensure a *fair comparison*, we use the same neural instantiations for all models in terms of (i) the nuisance functions $\hat{\eta}$ and (ii) the policy π_θ (see Supp. E). All results are averaged over 10 seeds.

• **Synthetic Data:** As is standard in causal inference literature (Hess & Feuerriegel, 2025; Hess et al., 2024; Kallus et al., 2019), we evaluate our method on synthetic data in order to have access to ground-truth counterfactuals. Here, we use an established data-generating process from the literature (Kallus et al., 2019): First, we simulate observed confounders $X \sim \text{Unif}[-2, 2]$ and unobserved confounders $U \sim \text{Ber}(1/2)$. The potential outcomes $Y[a]$ are then given by $Y[a] = (2a-1)X + (2a-1) - 2\sin(2(2a-1)X) - 2(2U-1)(1+0.5X) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,1)$ is random noise. Further, we assume a binary treatment, i.e., $d_a = 2$. For this, we first fix a ground-truth Γ^*. Then, we let the *true propensity score* be given by $e(1, x, u) = \frac{u}{\rho(x;1/\Gamma^*)} + \frac{1-u}{\rho(x;\Gamma^*)}$, where $\rho(x;\gamma) = 1 + (1/e(1,x) - 1)\gamma$, and $e(1,x) = \sigma(0.75x + 0.5)$ is the nominal propensity score.

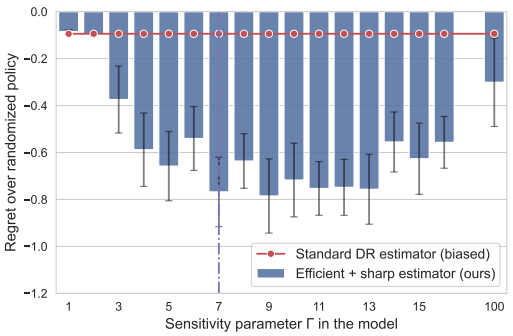

Figure 3: **Robustness analysis.** We set $\Gamma^* = 7$ in the data-generating process but use mis-specified sensitivity parameters Γ in our estimator (i.e., $\Gamma = 7$ is correctly specified, while $\Gamma \neq 7$ is mis-specified). We report the regret over a randomized policy *(lower values are better)*. Our estimator significantly improves upon the standard DR estimator, even for a completely mis-specified Γ.

Varying confounding strength: First, we demonstrate the performance of our method for increasing levels of unobserved confounding. For this, we increase the confounding parameter Γ^* in the data-generating process. We compare the regret of each method over a randomized baseline policy. In our method and (Kallus & Zhou, 2018a; 2021), we set the sensitivity parameter Γ equal to Γ^*.

Our results are shown in **Table** 2: (i) As expected, the standard methods for off-policy learning (i.e., a standard IPW estimator and a standard DR estimator) perform well for zero to very low levels of confounding. However, the standard methods are *biased* and thus become ineffective for $\Gamma^* > 1$. (ii) The method by Kallus & Zhou (2018a; 2021) performs well under low levels of confounding. Yet, the performance quickly deteriorates. (iii) Our proposed method performs clearly best for increasing Γ^*. Here, our method achieves a *relative performance gain by up to a factor of* 4.

Robustness analysis: Next, we show that our method is robust to mis-specification of the sensitivity parameter Γ. We thus fix the confounding strength to $\Gamma^* = 7$ in the DGP. We increase Γ from 1 (which corresponds to unconfoundedness) up to 100 (which mirrors almost assumption-free bounds). We again compute the regret of our learned policy over a randomized baseline policy to showcase the improvement. The method by Kallus & Zhou (2018a; 2021) has only a regret of -0.27 ± 0.06, even for correctly specified $\Gamma = 7$, and is hence not competitive; we thus removed it from the plot for better visualization. **Figure** 3 shows that our approach yields *robust results even for mis-specified*

Γ (i.e., when $\Gamma \neq 7$). Further, even under the (almost) no assumptions constraint of $\Gamma = 100$, our method provides significant improvements over the biased DR estimator.

Semi-parametrically efficient estimation: Finally, we show the benefits of our efficient estimation strategy over simple plug-in estimators. A semi-parametrically efficient estimator is the unbiased estimator with the lowest variance (Kennedy, 2022; van der Vaart, 1998). Hence, policies based on efficiently estimated bounds are learned better in low sample settings than those based on plug-in approaches. Therefore, we report the performance when we vary the number of training samples $\mathcal{D}_n$. Here, we compare our method against a naïve plug-in estimator of our sharp bounds based on Proposition 4.1. We again report the regret over a randomized baseline policy. **Figure** 4 shows that our method performs better in low sample settings, and achieves larger performance gains when increasing the sample size.

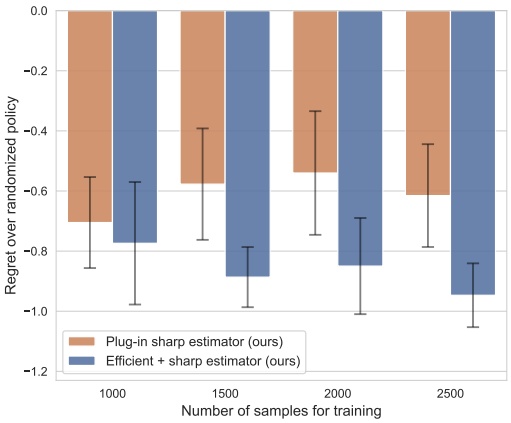

Figure 4: **Property of semi-parametrically efficient estimation.** We compare our efficient estimator with a simple plug-in estimator of our sharp upper bound from Proposition 4.1. We report the regret over a randomized policy *(lower values are better)*. Our efficient estimator leads to a lower regret and benefits from increasing sample size due to its optimal estimation properties.

• Real-world medical data: We evaluate our method on a real-world medical case study. For this, we use data from the International Stroke Trial (Sandercock et al., 2011), which is a randomized control trial (RCT) that examines the outcomes for early administration of aspirin, heparin, a combination of both, or none on acute ischaemic stroke. Importantly, this means that there are *four different* treatments available. The advantage of an RCT over observational data is that we can estimate the ground truth value function without bias, as the true propensity score is known. Our aim is to find the optimal treatment strategy based on patient covariates in order to prolong the *time-to-death* (TD) outcome variable (in days).

For this, we artificially introduce unobserved confounding as follows: In the training dataset, we randomly drop 60% of the untreated patients whose diastolic blood pressure is larger than the average, as well as 60% of the patients who received aspirin and whose blood pressure is lower than average. Then, we remove the diastolic blood pressure variable. Thereby, we introduce *unobserved* confounding in the training dataset.

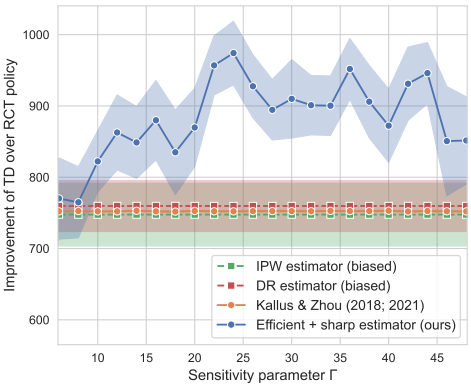

Figure 5: **Real-world medical data.** We compare our efficient estimator against the previous baselines based on data from the International Stroke Trial. Our method yields the best treatment policy and is robust over different Γ.

Results: We report the estimated improvement of the TD outcome of all methods over the randomized policy in Figure 5. Here, we vary the sensitivity parameter for both the method by Kallus & Zhou (2018a; 2021) and our method. Our method performs best at $\Gamma = 24$. Further, our method has the overall best treatment strategy, whereas all baselines fail to improve upon the randomized policy. Importantly, our method is stable for different parameterizations of Γ, which confirms the effectiveness of our method, and shows its applicability to medical scenarios.

Conclusion: We develop a novel semi-parametrically efficient estimator for sharp bounds on the value function under unobserved confounding. Further, our approach yields provably optimal, confounding-robust policies and avoids the instability of existing minimax-based methods. Our results provide a principled way for reliable decision-making from observational data, and show that robust policy learning can improve decision-making in sensitive applications such as healthcare.

REFERENCES

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

Alexis Bellot and Silvia Chiappa. Towards estimating bounds on the effect of policies under unobserved confounding. In *NeurIPS*, 2024.

Andrew Bennett, Nathan Kallus, Lihong Li, and Ali Mousavi. Off-policy evaluation in infinite horzon reinforcement learning with latent confounders. In *AISTATS*, 2021.

Andrew Bennett, Nathan Kallus, Miruna Oprescu, Wen Sun, and Kaiwen Wang. Efficient and sharp off-policy evaluation in robust Markov decision processes. In *NeurIPS*, 2024.

Aurelien Bibaut, Ivana Malenica, Nikos Vlassis, and Mark van der Laan. More efficient off-policy evaluation through regularized targeted learning. In *ICML*, 2019.

Matteo Bonvini, Edward Kennedy, Valerie Ventura, and Larry Wasserman. Sensitivity analysis for marginal structural models. *arXiv preprint*, arXiv:2210.04681, 2022.

Cecilia Ka Yuk Chan. A comprehensive AI policy framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, 20(38), 2023.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society*, 82(1):39–67, 2020.

James Cornfield, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernst L. Wynder. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22(1):173–203, 1959.

Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544):2645–2657, 2022.

Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/ doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *Journal of the American Statistical Association*, 2024.

Miroslav Dudik, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *ICML*, 2011.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 2024.

Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Sharp bounds for generalized causal sensitivity analysis. In *NeurIPS*, 2023.

Dennis Frauen, Fergus Imrie, Alicia Curth, Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. A neural framework for generalized causal sensitivity analysis. In *ICLR*, 2024a.

Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Fair off-policy learning from observational data. In *ICML*, 2024b.

Luke Guerdan, Amanda Coston, Kenneth Holstein, and Zhiwei Steven Wu. Predictive performance comparison of decision policies under confounding. In *ICML*, 2024.

Tobias Hatt, Daniel Tschernutter, and Stefan Feuerriegel. Generalizing off-policy learning under sample selection bias. In *UAI*, 2022.

Lars G. Hemkens, Hannah Ewald, Florian Naudet, Aviv Ladanie, Jonathan G. Shaw, Gautam Sajeev, and John P. A. Ioannidis. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. *Journal of Clinical Epodemiology*, 93, 2018.

Konstantin Hess and Stefan Feuerriegel. Stabilized neural prediction of potential outcomes in continuous time. In *ICLR*, 2025.

Konstantin Hess, Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bayesian neural controlled differential equations for treatment effect estimation. In *ICLR*, 2024.

Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3):292–304, 2022.

Wen Huang and Xintau Wu. Robustly improving bandit algorithms with confounded and selection biased offline data: A causal approach. In *AAAI*, 2024.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *ICML*, 2021.

Ying Jin, Zhimei Ren, and Zhengyuan Zhou. Sensitivity analysis under the f-sensitivity models: A distributional robustness perspective. *arXiv preprint*, arXiv:2203.04373, 2022.

Ying Jin, Zhimei Ren, and Emmanuel J. Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences (PNAS)*, 120(6):e2214889120, 2023.

Shalmali Joshi, Junzhe Zhang, and Elias Bareinboim. Towards safe policy learning under partial identifiability: A causal approach. In *AAAI*, 2024.

Nathan Kallus. Balanced policy evaluation and learning. In *NeurIPS*, 2018.

Nathan Kallus. More efficient policy learning via optimal retargeting. *Journal of the American Statistical Association*, 116(534):646–658, 2021.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *NeurIPS*, 2018a.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *AISTATS*, 2018b.

Nathan Kallus and Angela Zhou. Confounding robust policy evaluation in infinite-horizon reinforcement learning. In *NeurIPS*, 2020.

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

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *AISTATS*, 2019.

Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. In *ICML*, 2022.

Chinmaya Kausik, Yangyi Lu, Kevin Tan, Maggie Maker, Yixin Wang, and Ambuj Tewari. Offline policy evaluation and optimization under confounding. In *AISTATS*, 2024.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint*, 2022.

Stella Ladi and Dimitris Tsarouhas. EU economic governance and Covid-19: policy learning and windows of opportunity. *Journal of European Integration*, 42(8):1041–1056, 2020.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *ICML*, 2019.

Michel Ledoux and Michel Talagrand. Comparison theorems, random geometry and some limit theorems for empirical processes. *Annals of Probability*, 17(2):596–631, 1989.

Alex Luedtke. Simplifying debiased inference via automatic differentiation and probabilistic programming. *arXiv preprint*, 2405.08675, 2024.

Colin McDiarmid. *On the method of bounded differences*, pp. 148–188. Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference. Cambridge University Press, 1989.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In *NeurIPS*, 2020.

Jerzy Neyman. On the application of probability theory to agricultural experiments. *Annals of Agricultural Sciences*, 10:1–51, 1923.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Miruna Oprescu, Jacob Dorn, Marah Ghoummaid, Andrew Jesson, Nathan Kallus, and Uri Shalit. B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In *ICML*, 2023.

Alizée Pace, Hugo Yèche, Bernhard Schölkopf, Gunnar Rätsch, and Guy Tennenholtz. Delphic offline reinforcement learning under nonidentifiable hidden confounding. In *ICLR*, 2024.

Judea Pearl. *Causality*. Cambridge University Press, New York City, 2009. ISBN 9780521895606.

Min Qian and Susan A. Murphy. Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2):1180–1210, 2011.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of reression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–688, 1994.

Paul R. Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

Donald B. Rubin. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6(1):34–58, 1978.

Peter AG Sandercock, Maciej Niewada, and Anna Członkowska. The international stroke trial database. *Trials*, 12(101), 2011.

Jonas Schweisthal, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Reliable off-policy learning for dosage combinations. In *NeurIPS*, 2023.

Chengchun Shi, Masatoshi Uehara, Jiawei Huang, and Nan Jiang. A minimax learning approach to off-policy evaluation in confounded partially observable Markov decision processese. In *ICML*, 2022.

Chengchun Shi, Jin Zhu, Ye Shen, Shikai Luo, Hongtu Zhu, and Rui Song. Off-policy confidence interval estimation with confounded Markov decision process. *Journal of the American Statistical Association*, 119(545):273–284, 2024.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, 2015.

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

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *ICML*, 2015.

Daniel Tschernutter, Tobias Hatt, and Stefan Feuerriegel. Interpretable off-policy learning via hyperbox search. In *ICML*, 2022.

Aart van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, 1998. ISBN 0521496039.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Provably efficient causal reinforcement learning with confounded observational data. In *NeurIPS*, 2021.

Mingzhang Yin, Claudia Shi, Yixin Wang, and David M. Blei. Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, pp. 1–14, 2022.

Junzhe Zhang and Elias Bareinboim. Eligibility traces for confounding robust off-policy evaluation. *OpenReview preprint*, 2024.

A Extended related work

Offline reinforcement learning under unobserved confounding: Offline reinforcement learning deals with the problem of learning the optimal policy when the reward (value) function is defined over an infinite time horizon. Therefore, these works rely upon techniques that are different from ours.

Some works focus on off-policy evaluation under unobserved confounding (Bennett et al., 2021; Kallus & Zhou, 2020) and even propose computationally efficient algorithms for this task (Kausik et al., 2024). However, these methods primarily focus on the identification of policy value bounds without semi-parametrically efficient estimation procedures. In contrast, Bennett et al. (2024) propose an efficient estimator for offline reinforcement learning. Different from our work, however, they require estimation of density ratios in order to evaluate the policy value. Further, Pace et al. (2024) propose a heuristic approach that learns representations of the unobserved confounders but does not provide theoretical guarantees for efficiency or unbiasedness. Shi et al. (2022) propose an approach that involves the approximation of bridge functions in partially observable Markov decision processes (POMDPs). Additionally, Shi et al. (2024) use mediators as auxiliary variables to construct confidence intervals for policy evaluation under unobserved confounding. Finally, Wang et al. (2021) improve sample efficiency in offline reinforcement learning under both observed and unobserved confounding.

B CHOOSING THE SENSITIVITY PARAMETER IN THE MSM

In this work, we adopt the MSM (Tan, 2006) in order to bound the ratio between the *nominal propensity score*

$$e(a, x) = p(A = a \mid X = x), \tag{16}$$

and the *true propensity score*

$$e(a, x, u) = p(A = a \mid X = x, U = u). \tag{17}$$

Here, the nominal propensity score can be estimated from data, whereas the true propensity score is fundamentally unobservable. In particular, the MSM is given by

$$\Gamma^{-1} \leq \frac{e(a, x)}{1 - e(a, x)} \frac{1 - e(a, x, u)}{e(a, x, u)} \leq \Gamma \tag{18}$$

for some sensitivity parameter $\Gamma \geq 1$.

Typically, the sensitivity constraints Γ are chosen via domain knowledge (Frauen et al., 2023; Kallus et al., 2019) or data-driven heuristics (Hatt et al., 2022). For example, in practical applications, one typically has a benchmark variable (e.g., hours with sunlight) that is a known cause of the outcome (e.g., vitamin D deficiency), and one then wants to study how strong a confounder (e.g., other ecological activities such as nutrition) must be to explain away the effect of the benchmark variables. (Cinelli & Hazlett, 2020) term this the robustness value, which quantifies the strength of unobserved confounding needed to change conclusions.

Hence, to achieve this, a commonly used strategy for selecting Γ is the following: We can search for the smallest Γ such that the partially identified interval for the causal quantity of interest includes 0. Then, we can interpret Γ as a measure of the minimal deviation from unconfoundedness required to invalidate the effect of an intervention (Jesson et al., 2021; Jin et al., 2023).

C Extension to policy improvement

Our main results from above focus on optimizing the value function $V(\pi)$, which is common in practice (e.g., Dudik et al., 2011; Hatt et al., 2022). However, in some scenarios, an established baseline policy π_0 may be available; then, one may aim to make a small relative improvement yet with certain guarantees. This setting is commonly termed as *policy improvement* (Kallus & Zhou, 2018a; 2021; Laroche et al., 2019; Thomas et al., 2015).

Hence, we no longer aim to minimize bounds on the value function $V(\pi)$ but, instead, bounds on the regret of a candidate policy against a baseline policy (Kallus & Zhou, 2018a). Specifically, we can define the *regret* of policy π over baseline π_0 as $R_{\pi_0}(\pi) = V(\pi) - V(\pi_0)$. Hence, a negative regret implies that policy π improves upon π_0. Importantly, the optimal confounding-robust policy π^* in Equation 6 can also be defined as the policy π that achieves the best relative improvement over baseline π_0 in the worst-case scenario, that is,

$$\pi^* = \arg\min_{\pi \in \Pi} \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} R_{\pi_0}(\pi). \tag{19}$$

This definition is *equivalent* to Equation 6. Nevertheless, the above objective may be preferred in practice when aiming at policy improvement.

We now show in the following three corollaries that our results directly generalize to policy improvement. First, we provide a closed-form solution for an upper bound of the regret function $R_{\pi_0}(\pi)$, given our sensitivity constraints $\mathcal{P}(\Gamma)$.

Corollary C.1. *An upper bound for the regret function $R_{\pi_0}(\pi)$ is given by $R^+_\pi(\pi) = \int_{\mathcal{X}} \sum_a \Big(Q^{+,*}(a,x)\pi(a \mid x) - Q^{-,*}(a,x)\pi_0(a \mid x) \Big) \, \mathrm{d}p(x)$.*

Proof. See Supplement D.4. □

Next, we derive a semi-parametrically efficient, one-step bias-corrected estimator, which is based on the efficient influence function.

Corollary C.2. *A semi-parametrically efficient estimator for the upper regret bound is given by*

$$\begin{aligned}\hat{R}^+_{\pi_0}(\pi) = \sum_{\pm \in \{-,+\}} \pm \mathbb{P}_n \Big\{ \sum_a \pi^{\pm}_{a,X} \Big[\hat{Q}^{\pm,*}_{a,X} - \hat{e}_{a,X} \Big(b^{\mp} \hat{\underline{\mu}}^{\pm}_{a,X} + b^{\pm} \hat{\bar{\mu}}^{\pm}_{a,X} \Big) \Big] \\ + \pi^{\pm}_{A,X} \Big(b^{\mp} \hat{\underline{\mu}}^{\pm}_{A,X} + b^{\pm} \hat{\bar{\mu}}^{\pm}_{A,X} \Big) + \frac{\pi^{\pm}_{A,X}}{\hat{e}_{A,X}} \Big[\Big(\hat{c}^{\mp}_{A,X} - \hat{c}^{\pm}_{A,X} \Big) \Big(\hat{F}^{-1}_{X,A}(\alpha^{\pm})(\hat{\underline{\Delta}}^{\pm}_{Y,A,X} - \alpha^{\pm}) \Big) \\ + \hat{c}^{\mp}_{A,X} \Big(Y \hat{\underline{\Delta}}^{\pm}_{Y,A,X} - \hat{\underline{\mu}}^{\pm}_{A,X} \Big) + \hat{c}^{\pm}_{A,X} \Big(Y \hat{\bar{\Delta}}^{\pm}_{Y,A,X} - \hat{\bar{\mu}}^{\pm}_{A,X} \Big) \Big] \Big\},\end{aligned}$$

where we let $\pi^+ = \pi$ and $\pi^- = \pi_0$ for readability.

Proof. See Supplement D.5. □

Finally, we provide improvement guarantees: given a baseline policy π_0 (e.g., the standard of care), if the empirical estimator $\hat{R}_{\pi_0}(\pi)^+$ is *negative*, which we can check by evaluating it, we are *guaranteed* that π improves upon π_0 and introduces *no harm*.

Corollary C.3. *Under the same assumption as in Theorem 4.4, for any policy $\pi \in \Pi$ and baseline policy $\pi_0 \in \Pi$, it holds, with probability $1-\delta$, that $R_{\pi_0}(\pi) \leq \hat{R}^+_{\pi_0}(\pi) + 4C_v \Big(\mathcal{R}_n(\Pi) + \frac{5}{2} \sqrt{\frac{1}{2n} \log\Big(\frac{2}{\delta}\Big)} \Big)$.*

Proof. See Supplement D.6. □

D PROOFS

D.1 SHARP BOUND OF THE VALUE FUNCTION

Proposition D.1. *Let $Q^{\pm,*}(a, x)$ be the sharp upper/lower bound for the conditional average potential outcome, given our sensitivity constraints $\mathcal{P}(\Gamma)$. Then, sharp bounds for the value function $V(\pi)$ are given by*

$$V^{\pm,*}(\pi) = \int_{\mathcal{X}} \sum_{a} Q^{\pm,*}(a, x)\, \pi(a \mid x)\, \mathrm{d}p(x). \tag{20}$$

Proof. We provide the proof for the sharp upper bound $V^{+,*}(\pi)$. The lower bound follows completely analogously by swapping the signs and replacing the supremum with an infimum.

We start by noting that the upper bound on the value function depends on the set of admissible distributions $\mathcal{P}(\Gamma)$ induced by the sensitivity model, that is,

$$V^{+,*}(\pi) = V^{+,*}(\pi, \mathcal{P}(\Gamma)). \tag{21}$$

Hence, we can write

$$\begin{aligned}
& V^{+,*}(\pi) && (22)\\
=& V^{+,*}(\pi, \mathcal{P}(\Gamma)) && (23)\\
=& \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} V(\pi, \tilde{p}) && (24)\\
=& \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} \int_{\mathcal{X}} \sum_{a} Q(a, x, \tilde{p})\, \pi(a \mid x)\, \mathrm{d}\tilde{p}(x) && (25)\\
=& \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} \int_{\mathcal{X}} \sum_{a} Q(a, x, \tilde{p})\, \pi(a \mid x)\, \mathrm{d}p(x), && (26)
\end{aligned}$$

where Equation 26 follows from the equality $p(\mathcal{D}) = \tilde{p}(\mathcal{D})$ for all $\tilde{p} \in \mathcal{P}(\Gamma)$.

Clearly, by definition of the optimal bounds $Q^{+,*}(a, x)$, we have that

$$Q(a, x, \tilde{p}) \leq \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} Q(a, x, \tilde{p}) = Q^{+,*}(a, x, \mathcal{P}(\Gamma)) \tag{27}$$

for all $\tilde{p} \in \mathcal{P}(\Gamma)$, and since $Q^{+,*}(a, x) \in L^1(\pi, p)$, we know by dominated convergence (Frauen et al., 2023) that

$$\sup_{\tilde{p} \in \mathcal{P}(\Gamma)} \int_{\mathcal{X}} \sum_{a} Q(a, x, \tilde{p})\, \pi(a \mid x)\, \mathrm{d}p(x) = \int_{\mathcal{X}} \sum_{a} Q^{+,*}(a, x)\, \pi(a \mid x)\, \mathrm{d}p(x). \tag{28}$$

□

D.2 Efficient estimator of the sharp bound of the value function

Theorem D.2. *The efficient estimator for the sharp upper bound of the value function is given by*

$$\hat{V}^{+,*}(\pi) \tag{29}$$

$$=\mathbb{P}_n\Big\{\sum_a \pi_{a,X}\Big[\hat{Q}^{+,*}_{a,X} - \hat{e}_{a,X}\Big(b^-\hat{\underline{\mu}}^+_{a,X} + b^+\hat{\bar{\mu}}^+_{a,X}\Big)\Big] + \pi_{A,X}\Big(b^-\hat{\underline{\mu}}^+_{A,X} + b^+\hat{\bar{\mu}}^+_{A,X}\Big) \tag{30}$$

$$+\frac{\pi_{A,X}}{\hat{e}_{A,X}}\Big[\Big(\hat{c}^-_{A,X} - \hat{c}^+_{A,X}\Big)\Big(\hat{F}^{-1}_{X,A}(\alpha^+)(\hat{\underline{\Delta}}^+_{Y,A,X} - \alpha^+)\Big) + \hat{c}^-_{A,X}\Big(Y\hat{\underline{\Delta}}^+_{Y,A,X} - \hat{\underline{\mu}}^+_{A,X}\Big) + \hat{c}^+_{A,X}\Big(Y\hat{\bar{\Delta}}^+_{Y,A,X} - \hat{\bar{\mu}}^+_{A,X}\Big)\Big]\Big\}. \tag{31}$$

Proof. The sharp upper bound of the value function is given by

$$V^{\pm,*}(\pi) = \int_{\mathcal{X}} \sum_a Q^{\pm,*}(a,x)\pi(a \mid x)\,\mathrm{d}p(x). \tag{32}$$

In the following, we derive the efficient estimator for this quantity. Therein, we make use of the chain rule for deriving efficient influence function (Kennedy, 2022). *A proof of the validity of the chain rule for deriving efficient influence functions is provided by Luedtke (2024)* **(Lemma S3)**.

In order to avoid notational overload and for the sake of clarity, we do not use additional variables such as $b^\pm$, $c^\pm$, $\underline{\Delta}^\pm$, $\bar{\mu}^\pm$, etc. until the final steps, such that the derivation becomes easier to follow. Moreover, we make the dependency on nuisance functions $\eta \subseteq \{e(a,x), F^{-1}_{a,x}(\alpha^\pm), \bar{\mu}^\pm(a,x), \underline{\mu}^\pm(a,x)\}$ explicit by writing, for example, $V^{+,*}(\pi;\eta)$ for $V^{+,*}(\pi)$.

The influence function of $V^{+,*}(\pi;\eta)$ is given by

$$\mathbb{IF}\Big(V^{+,*}(\pi;\eta)\Big) \tag{33}$$

$$=\mathbb{IF}\Big(\int_{\mathcal{X}}\sum_a Q^{+,*}(a,x;\eta)\pi(a \mid x)\,\mathrm{d}p(x)\Big) \tag{34}$$

$$=\sum_a \int_{\mathcal{X}} \pi(a \mid x)\,\mathbb{IF}\Big(p(x)Q^{+,*}(a,x;\eta)\Big)\,\mathrm{d}x \tag{35}$$

$$=\sum_a \int_{\mathcal{X}} \pi(a \mid x)\,\mathbb{IF}\Big(p(x)\Big)Q^{+,*}(a,x;\eta) + \pi(a \mid x)\,p(x)\,\mathbb{IF}\Big(Q^{+,*}(a,x;\eta)\Big)\,\mathrm{d}x \tag{36}$$

$$=\sum_a \int_{\mathcal{X}} \pi(a \mid x)\Big(\mathbb{1}_{\{X=x\}} - p(x)\Big)Q^{+,*}(a,x;\eta)\,\mathrm{d}x + \sum_a \int_{\mathcal{X}} \pi(a \mid x)\,p(x)\,\mathbb{IF}\Big(Q^{+,*}(a,x;\eta)\Big)\,\mathrm{d}x \tag{37}$$

$$=\sum_a \pi(a \mid X)\,Q^{+,*}(a,X;\eta) - V^{+,*}(\pi;\eta) + \sum_a \int_{\mathcal{X}} \pi(a \mid x)\,p(x)\,\mathbb{IF}\Big(Q^{+,*}(a,x;\eta)\Big)\,\mathrm{d}x \tag{38}$$

Hence, in Equation 38, we are left to compute the efficient influence function of $Q^{+,*}(a,x)$, that is, the sharp upper bound of the CAPO. With $\alpha^+ = \Gamma/(1+\Gamma)$, the sharp upper bound $Q^{+,*}(a,x)$ is given by

$$Q^{+,*}(a,x;\eta) \tag{39}$$

$$=\Big((1-\Gamma^{-1})e(a,x) + \Gamma^{-1}\Big)\mathbb{E}\Big[Y\mathbb{1}_{\{Y\le F^{-1}_{X,A}(\alpha^+)\}} \mid X=x, A=a\Big] \tag{40}$$

$$+\Big((1-\Gamma)e(a,x) + \Gamma\Big)\mathbb{E}\Big[Y\mathbb{1}_{\{Y\ge F^{-1}_{X,A}(\alpha^+)\}} \mid X=x, A=a\Big]. \tag{41}$$

Hence, the influence function is given by

$$\mathbb{IF}\Big(Q^{+,*}(a,x;\eta)\Big) \tag{42}$$

$$= \underbrace{\mathbb{IF}\Big((1-\Gamma^{-1})e(a,x)+\Gamma^{-1}\Big)}_{(a)} \mathbb{E}\Big[Y\mathbb{1}_{\{Y\leq F^{-1}_{X,A}(\alpha^+)\}} \mid X=x, A=a\Big] \tag{43}$$

$$+ \Big((1-\Gamma^{-1})e(a,x)+\Gamma^{-1}\Big) \underbrace{\mathbb{IF}\Big(\mathbb{E}\Big[Y\mathbb{1}_{\{Y\leq F^{-1}_{X,A}(\alpha^+)\}} \mid X=x, A=a\Big]\Big)}_{(c)} \tag{44}$$

$$+ \underbrace{\mathbb{IF}\Big((1-\Gamma)e(a,x)+\Gamma\Big)}_{(b)} \mathbb{E}\Big[Y\mathbb{1}_{\{Y\geq F^{-1}_{X,A}(\alpha^+)\}} \mid X=x, A=a\Big] \tag{45}$$

$$+ \Big((1-\Gamma)e(a,x)+\Gamma\Big) \underbrace{\mathbb{IF}\Big(\mathbb{E}\Big[Y\mathbb{1}_{\{Y\geq F^{-1}_{X,A}(\alpha^+)\}} \mid X=x, A=a\Big]\Big)}_{(d)}. \tag{46}$$

We start with (a) and (b). For (a), we obtain

$$\mathbb{IF}\Big((1-\Gamma^{-1})e(a,x)+\Gamma^{-1}\Big) \tag{47}$$

$$=(1-\Gamma^{-1})\mathbb{IF}\Big(e(a,x)\Big) \tag{48}$$

$$=(1-\Gamma^{-1})\frac{\mathbb{1}_{\{X=x\}}}{p(x)}\Big(\mathbb{1}_{\{A=a\}}-e(a,x)\Big), \tag{49}$$

and, similarly for (b), we yield

$$\mathbb{IF}\Big((1-\Gamma)e(a,x)+\Gamma\Big) \tag{50}$$

$$=(1-\Gamma)\frac{\mathbb{1}_{\{X=x\}}}{p(x)}\Big(\mathbb{1}_{\{A=a\}}-e(a,x)\Big). \tag{51}$$

Next, we compute the influence function of (c) via

$$\mathbb{IF}\Big(\mathbb{E}\Big[Y\mathbb{1}_{\{Y\leq F^{-1}_{X,A}(\alpha^+)\}} \mid X=x, A=a\Big]\Big) \tag{52}$$

$$=\mathbb{IF}\Big(\int_{\mathcal{Y}} \mathbb{1}_{\{y\leq F^{-1}_{x,a}(\alpha^+)\}}\, y\, p(y\mid x,a)\,\mathrm{d}y\Big) \tag{53}$$

$$= \int_{\mathcal{Y}} \mathbb{IF}\Big(\mathbb{1}_{\{y\leq F^{-1}_{x,a}(\alpha^+)\}}\Big)\, y\, p(y\mid x,a) + \mathbb{1}_{\{y\leq F^{-1}_{x,a}(\alpha^+)\}} y \mathbb{IF}\Big(p(y\mid x,a)\Big)\,\mathrm{d}y \tag{54}$$

$$= \underbrace{\int_{\mathcal{Y}} \mathbb{IF}\Big(\mathbb{1}_{\{y\leq F^{-1}_{x,a}(\alpha^+)\}}\Big)\, y\, p(y\mid x,a)\,\mathrm{d}y}_{(c_1)} + \underbrace{\int_{\mathcal{Y}} \mathbb{1}_{\{y\leq F^{-1}_{x,a}(\alpha^+)\}}\, y\, \mathbb{IF}\Big(p(y\mid x,a)\Big)\,\mathrm{d}y}_{(c_2)}. \tag{55}$$

For (c_1), we first note that

$$\mathbb{IF}\Big(F_{x,a}(\alpha^+)\Big) \tag{56}$$

$$=\mathbb{IF}\Big(\mathbb{P}(Y\leq y \mid X=x, A=a)\Big) \tag{57}$$

$$=\mathbb{IF}\Big(\mathbb{E}[\mathbb{1}_{\{Y\leq y\}} \mid X=x, A=a]\Big) \tag{58}$$

$$=\frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big(\mathbb{1}_{\{Y\leq y\}}-\mathbb{E}[\mathbb{1}_{\{Y\leq y\}} \mid X=x, A=a]\Big) \tag{59}$$

$$= \underbrace{\frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big(\mathbb{1}_{\{Y\leq y\}}-F_{x,a}(y)\Big)}_{(*)}. \tag{60}$$

Then, we can simplify (c_1) via

$$\int_{\mathcal{Y}} \mathbb{IF}\Big(\mathbb{1}_{\{y\le F^{-1}_{x,a}(\alpha^+)\}}\Big)\, y\, p(y \mid x,a)\,\mathrm{d}y \tag{61}$$

$$=\int_{\mathcal{Y}} \delta\Big(y - F^{-1}_{x,a}(\alpha^+)\Big)\mathbb{IF}\Big(F^{-1}_{x,a}(\alpha^+)\Big)\, y\, p(y \mid x,a)\,\mathrm{d}y \tag{62}$$

$$=\mathbb{IF}\Big(F^{-1}_{x,a}(\alpha^+)\Big)\int_{\mathcal{Y}} \delta\Big(y - F^{-1}_{x,a}(\alpha^+)\Big)\, y\, p(y \mid x,a)\,\mathrm{d}y \tag{63}$$

$$=\mathbb{IF}\Big(F^{-1}_{x,a}(F_{x,a}(y^*))\Big)\int_{\mathcal{Y}} \delta\Big(y - F^{-1}_{x,a}(\alpha^+)\Big)\, y\, p(y \mid x,a)\,\mathrm{d}y \tag{64}$$

$$=\frac{\mathrm{d}}{\mathrm{d}q}F^{-1}_{x,a}(q)\Big|_{q=F_{x,a}(y^*)}\mathbb{IF}\Big(F_{x,a}(y^*)\Big)\int_{\mathcal{Y}} \delta\Big(y - F^{-1}_{x,a}(\alpha^+)\Big)\, y\, p(y \mid x,a)\,\mathrm{d}y \tag{65}$$

$$=\frac{1}{F'_{x,a}(F^{-1}_{x,a}(F_{x,a}(y^*)))}\mathbb{IF}\Big(F_{x,a}(y^*)\Big)\int_{\mathcal{Y}} \delta\Big(y - F^{-1}_{x,a}(\alpha^+)\Big)\, y\, p(y \mid x,a)\,\mathrm{d}y \tag{66}$$

$$=\frac{1}{p(y^* \mid x,a)}\mathbb{IF}\Big(F_{x,a}(y^*)\Big)\int_{\mathcal{Y}} \delta\Big(y - F^{-1}_{x,a}(\alpha^+)\Big)\, y\, p(y \mid x,a)\,\mathrm{d}y \tag{67}$$

$$\overset{(*)}{=}\frac{1}{p(y^* \mid x,a)}\frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big(\mathbb{1}_{\{Y\le y^*\}} - F_{x,a}(y^*)\Big)\int_{\mathcal{Y}} \delta(y-y^*)\, y\, p(y \mid x,a)\,\mathrm{d}y \tag{68}$$

$$=\frac{1}{p(y^* \mid x,a)}\frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big(\mathbb{1}_{\{Y\le y^*\}} - \alpha^+\Big)\int_{\mathcal{Y}} \delta(y-y^*)\, y\, p(y \mid x,a)\,\mathrm{d}y \tag{69}$$

$$=\frac{1}{p(y^* \mid x,a)}\frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big(\mathbb{1}_{\{Y\le y^*\}} - \alpha^+\Big)\, y^*\, p(y^* \mid x,a) \tag{70}$$

$$=\frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}F^{-1}_{x,a}(\alpha^+)(\mathbb{1}_{\{Y\le F^{-1}_{x,a}(\alpha^+)\}} - \alpha^+), \tag{71}$$

for some $y^* \in \mathcal{Y}$ such that $F_{x,a}(y^*) = \alpha^+$, where $\delta(\cdot)$ is the Dirac-delta function.

Next, we simplify (c_2) via

$$\int_{\mathcal{Y}} \mathbb{1}_{\{y\le F^{-1}_{x,a}(\alpha^+)\}} y\mathbb{IF}\Big(p(y \mid x,a)\Big)\,\mathrm{d}y \tag{72}$$

$$=\int_{\mathcal{Y}} \mathbb{1}_{\{y\le F^{-1}_{x,a}(\alpha^+)\}}\, y\,\mathbb{IF}\Big(\mathbb{E}[\mathbb{1}_{\{Y=y\}} \mid X=x, A=a]\Big)\,\mathrm{d}y \tag{73}$$

$$=\int_{\mathcal{Y}} \mathbb{1}_{\{y\le F^{-1}_{x,a}(\alpha^+)\}}\, y\,\frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big(\mathbb{1}_{\{Y=y\}} - p(y \mid x,a\Big)\,\mathrm{d}y \tag{74}$$

$$=\frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big(Y\,\mathbb{1}_{\{Y\le F^{-1}_{x,a}(\alpha^+)\}} - \mathbb{E}[Y\,\mathbb{1}_{\{Y\le F^{-1}_{X,A}(\alpha^+)\}} \mid X=x, A=a]\Big). \tag{75}$$

Then, combining (c_1) and (c_2), we get

$$\mathbb{IF}\Big(\mathbb{E}\Big[Y\,\mathbb{1}_{\{Y\le F^{-1}_{X,A}(\alpha^+)\}} \mid X=x, A=a\Big]\Big) \tag{76}$$

$$=\frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big(Y\,\mathbb{1}_{\{Y\le F^{-1}_{x,a}(\alpha^+)\}} - \mathbb{E}[Y\,\mathbb{1}_{\{Y\le F^{-1}_{x,a}(\alpha^+)\}} \mid X=x, A=a] + F^{-1}_{x,a}(\alpha^+)(\mathbb{1}_{\{Y\le F^{-1}_{x,a}(\alpha^+)\}} - \alpha^+)\Big). \tag{77}$$

Finally, we compute the influence function of (d) analogously to (c) via

$$\mathbb{IF}\Big(\mathbb{E}\Big[Y\,\mathbb{1}_{\{Y\ge F^{-1}_{X,A}(\alpha^+)\}} \mid X=x, A=a\Big]\Big) \tag{78}$$

$$=\underbrace{\int_{\mathcal{Y}} \mathbb{IF}\Big(\mathbb{1}_{\{y\ge F^{-1}_{x,a}(\alpha^+)\}}\Big)\, y\, p(y \mid x,a)\,\mathrm{d}y}_{(d_1)} + \underbrace{\int_{\mathcal{Y}} \mathbb{1}_{\{y\ge F^{-1}_{x,a}(\alpha^+)\}}\, y\,\mathbb{IF}\Big(p(y \mid x,a)\Big)\,\mathrm{d}y}_{(d_2)}. \tag{79}$$

Again, we start with (d_1) via

$$\int_{\mathcal{Y}} \mathbb{IF}\Big(\mathbb{1}_{\{y\geq F^{-1}_{x,a}(\alpha^+)\}}\Big)\, y\, p(y \mid x, a)\, \mathrm{d}y \tag{80}$$

$$= \int_{\mathcal{Y}} \mathbb{IF}\Big((1 - \mathbb{1}_{\{y\leq F^{-1}_{x,a}(\alpha^+)\}})\Big)\, y\, p(y \mid x, a)\, \mathrm{d}y \tag{81}$$

$$= -\int_{\mathcal{Y}} \mathbb{IF}\Big(\mathbb{1}_{\{y\leq F^{-1}_{x,a}(\alpha^+)\}}\Big)\, y\, p(y \mid x, a)\, \mathrm{d}y \tag{82}$$

$$= \frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big(-F^{-1}_{x,a}(\alpha^+)\Big)(\mathbb{1}_{\{Y\leq F^{-1}_{x,a}(\alpha^+)\}} - \alpha^+), \tag{83}$$

using the result for (c_1) in Equation 71. Further, for (d_2), we get that

$$\int_{\mathcal{Y}} \mathbb{1}_{\{y\geq F^{-1}_{x,a}(\alpha^+)\}}\, y\, \mathbb{IF}\Big(p(y \mid x, a)\Big)\, \mathrm{d}y \tag{84}$$

$$= \frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big(Y\mathbb{1}_{\{Y\geq F^{-1}_{x,a}(\alpha^+)\}} - \mathbb{E}[Y\mathbb{1}_{\{Y\geq F^{-1}_{X,A}(\alpha^+)\}} \mid X = x, A = a]\Big). \tag{85}$$

Combining (d_1) and (d_2), we obtain that

$$\mathbb{IF}\Big(\mathbb{E}\Big[Y\mathbb{1}_{\{Y\geq F^{-1}_{x,a}(\alpha^+)\}} \mid X = x, A = a\Big]\Big) \tag{86}$$

$$= \frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big(Y\mathbb{1}_{\{Y\geq F^{-1}_{x,a}(\alpha^+)\}} - \mathbb{E}[Y\mathbb{1}_{\{Y\geq F^{-1}_{X,A}(\alpha^+)\}} \mid X = x, A = a] - F^{-1}_{x,a}(\alpha^+)(\mathbb{1}_{\{Y\leq F^{-1}_{x,a}(\alpha^+)\}} - \alpha^+)\Big). \tag{87}$$

Finally, we can state the influence function of $Q^{+,*}(a,x)$ by combining (a)–(d) in Equation 49, Equation 51, Equation 77, and Equation 87. We get

$$\mathbb{IF}\Big(Q^{+,*}(a,x;\eta)\Big) \tag{88}$$

$$= \frac{\mathbb{1}_{\{X=x\}}}{p(x)}(1-\Gamma^{-1})\Big(\mathbb{1}_{\{A=a\}} - e(a,x)\Big)\mathbb{E}\Big[Y\mathbb{1}_{\{Y\leq F^{-1}_{X,A}(\alpha^+)\}} \mid X = x, A = a\Big] \tag{89}$$

$$+ \frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big((1-\Gamma^{-1})e(a,x) + \Gamma^{-1}\Big) \tag{90}$$

$$\times \Big(Y\mathbb{1}_{\{Y\leq F^{-1}_{x,a}(\alpha^+)\}} - \mathbb{E}[Y\mathbb{1}_{\{Y\leq F^{-1}_{X,A}(\alpha^+)\}} \mid X = x, A = a] + F^{-1}_{x,a}(\alpha^+)(\mathbb{1}_{\{Y\leq F^{-1}_{x,a}(\alpha^+)\}} - \alpha^+)\Big) \tag{91}$$

$$+ \frac{\mathbb{1}_{\{X=x\}}}{p(x)}(1-\Gamma)\Big(\mathbb{1}_{\{A=a\}} - e(a,x)\Big)\mathbb{E}\Big[Y\mathbb{1}_{\{Y\geq F^{-1}_{X,A}(\alpha^+)\}} \mid X = x, A = a\Big] \tag{92}$$

$$+ \frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big((1-\Gamma)e(a,x) + \Gamma\Big) \tag{93}$$

$$\times \Big(Y\mathbb{1}_{\{Y\geq F^{-1}_{x,a}(\alpha^+)\}} - \mathbb{E}[Y\mathbb{1}_{\{Y\geq F^{-1}_{X,A}(\alpha^+)\}} \mid X = x, A = a] - F^{-1}_{x,a}(\alpha^+)(\mathbb{1}_{\{Y\leq F^{-1}_{x,a}(\alpha^+)\}} - \alpha^+)\Big). \tag{94}$$

In order to simplify the above lengthy equation as in our main paper, we introduce the following variables:

- $b^{\pm} = (1 - \Gamma^{\pm 1})$
- $c^{\pm}(a,x;\eta) = (b^{\pm}e(a,x) + \Gamma^{\pm 1})$
- $\underline{\Delta}^{\pm}(y,a,x;\eta) = \mathbb{1}_{\{y\leq F^{-1}_{x,a}(\alpha^{\pm})\}}$
- $\bar{\Delta}^{\pm}(y,a,x;\eta) = \mathbb{1}_{\{y\geq F^{-1}_{x,a}(\alpha^{\pm})\}}$
- $\underline{\mu}^{\pm}(a,x;\eta) = \mathbb{E}[Y\underline{\Delta}^{\pm}(Y,A,X;\eta) \mid X = x, A = a]$

- $\bar{\mu}^{\pm}(a, x; \eta) = \mathbb{E}[Y\bar{\Delta}^{\pm}(Y, A, X; \eta) \mid X = x, A = a]$

Then, Equation 94 simplifies to

$$\mathbb{IF}\Big(Q^{+,*}(a, x; \eta)\Big) \tag{95}$$

$$= \frac{\mathbb{1}_{\{X=x\}}}{p(x)} b^{-}\Big(\mathbb{1}_{\{A=a\}} - e(a, x)\Big)\underline{\mu}^{+}(a, x; \eta) \tag{96}$$

$$+ \frac{\mathbb{1}_{\{X=x, A=a\}}}{p(x)e(a, x)} c^{-}(a, x; \eta)\Big(Y\underline{\Delta}^{+}(Y, a, x; \eta) - \underline{\mu}^{+}(a, x; \eta) + F^{-1}_{x,a}(\alpha^{+})(\underline{\Delta}^{+}(Y, a, x; \eta) - \alpha^{+})\Big) \tag{97}$$

$$+ \frac{\mathbb{1}_{\{X=x\}}}{p(x)} b^{+}\Big(\mathbb{1}_{\{A=a\}} - e(a, x)\Big)\bar{\mu}^{+}(a, x; \eta) \tag{98}$$

$$+ \frac{\mathbb{1}_{\{X=x, A=a\}}}{p(x)e(a, x)} c^{+}(a, x; \eta)\Big(Y\bar{\Delta}^{+}(Y, a, x; \eta) - \bar{\mu}^{+}(a, x; \eta) - F^{-1}_{x,a}(\alpha^{+})(\underline{\Delta}^{+}(Y, a, x; \eta) - \alpha^{+})\Big) \tag{99}$$

$$= \frac{\mathbb{1}_{\{X=x\}}}{p(x)}\Big\{\Big(\mathbb{1}_{\{A=a\}} - e(a, x)\Big)\Big(b^{-}\underline{\mu}^{+}(a, x; \eta) + b^{+}\bar{\mu}^{+}(a, x; \eta)\Big) \tag{100}$$

$$+ \frac{\mathbb{1}_{\{A=a\}}}{e(a, x)}\Big[\Big(c^{-}(a, x; \eta) - c^{+}(a, x; \eta)\Big)\Big(F^{-1}_{x,a}(\alpha^{+})(\underline{\Delta}^{+}(Y, a, x; \eta) - \alpha^{+})\Big) \tag{101}$$

$$+ c^{-}(a, x; \eta)\Big(Y\underline{\Delta}^{+}(Y, a, x; \eta) - \underline{\mu}^{+}(a, x; \eta)\Big) + c^{+}(a, x; \eta)\Big(Y\bar{\Delta}^{+}(Y, a, x; \eta) - \bar{\mu}^{+}(a, x; \eta)\Big)\Big]\Big\} \tag{102}$$

Finally, we can combine Equation 38 and Equation 102. That is, the efficient influence function of $V^{+,*}(\pi)$ is given by

$$\mathbb{IF}\Big(V^{+,*}(\pi; \eta)\Big) \tag{103}$$

$$= \sum_{a} \pi(a \mid X) Q^{+,*}(a, X) - V^{+,*}(\pi) \tag{104}$$

$$+ \sum_{a} \pi(a \mid X)\Big(\mathbb{1}_{\{A=a\}} - e(a, X)\Big)\Big(b^{-}\underline{\mu}^{+}(a, X; \eta) + b^{+}\bar{\mu}^{+}(a, X; \eta)\Big) \tag{105}$$

$$+ \sum_{a} \pi(a \mid X)\frac{\mathbb{1}_{\{A=a\}}}{e(a, X)}\Big[\Big(c^{-}(a, X; \eta) - c^{+}(a, X; \eta)\Big)\Big(F^{-1}_{x,a}(\alpha^{+})(\underline{\Delta}^{+}(Y, a, X; \eta) - \alpha^{+})\Big) \tag{106}$$

$$+ c^{-}(a, X; \eta)\Big(Y\underline{\Delta}^{+}(Y, a, X; \eta) - \underline{\mu}^{+}(a, X; \eta)\Big) + c^{+}(a, X; \eta)\Big(Y\bar{\Delta}^{+}(Y, a, X; \eta) - \bar{\mu}^{+}(a, X; \eta)\Big)\Big] \tag{107}$$

$$= - V^{+,*}(\pi) + \sum_{a} \pi(a \mid X)\Big[Q^{+,*}(a, X) - e(a, X)\Big(b^{-}\underline{\mu}^{+}(a, X; \eta) + b^{+}\bar{\mu}^{+}(a, X; \eta)\Big)\Big] \tag{108}$$

$$+ \pi(A \mid X)\Big(b^{-}\underline{\mu}^{+}(A, X; \eta) + b^{+}\bar{\mu}^{+}(A, X; \eta)\Big) \tag{109}$$

$$+ \frac{\pi(A \mid X)}{e(A, X)}\Big[\Big(c^{-}(A, X; \eta) - c^{+}(A, X; \eta)\Big)\Big(F^{-1}_{X,A}(\alpha^{+})(\underline{\Delta}_{\alpha^{+}}(Y, A, X; \eta) - \alpha^{+})\Big) \tag{110}$$

$$+ c^{-}(A, X; \eta)\Big(Y\underline{\Delta}^{+}(Y, A, X; \eta) - \underline{\mu}^{+}(A, X; \eta)\Big) + c^{+}(A, X; \eta)\Big(Y\bar{\Delta}^{+}(Y, A, X; \eta) - \bar{\mu}^{+}(A, X; \eta)\Big)\Big] \tag{111}$$

We can derive the efficient estimator for the bounds of the value function through one-step bias correction via

$$V^{+,*}(\pi;\hat{\eta}) + \mathbb{P}_n\Big\{V^{+,*}(\pi;\hat{\eta})\Big\} \tag{112}$$

$$=\mathbb{P}_n\Big\{\sum_a \pi(a \mid X)\Big[Q^{+,*}(a,X;\hat{\eta}) - \hat{e}(a,X)\Big(b^-\hat{\underline{\mu}}^+(a,X;\hat{\eta}) + b^+\hat{\bar{\mu}}^+(a,X;\hat{\eta})\Big)\Big] \tag{113}$$

$$+ \pi(A \mid X)\Big(b^-\hat{\underline{\mu}}^+(A,X;\hat{\eta}) + b^+\hat{\bar{\mu}}^+(A,X;\hat{\eta})\Big) \tag{114}$$

$$+ \frac{\pi(A \mid X)}{\hat{e}(A,X)}\Big[\Big(c^-(A,X;\hat{\eta}) - c^+(A,X;\hat{\eta})\Big)\Big(\hat{F}^{-1}_{X,A}(\alpha^+)(\underline{\Delta}^+(Y,A,X;\hat{\eta}) - \alpha^+)\Big) \tag{115}$$

$$+ c^-(A,X;\hat{\eta})\Big(Y\underline{\Delta}^+(Y,A,X;\hat{\eta}) - \hat{\underline{\mu}}^+(A,X;\hat{\eta})\Big) + c^+(A,X;\hat{\eta})\Big(Y\bar{\Delta}^+(Y,A,X;\hat{\eta}) - \hat{\bar{\mu}}^+(A,X;\hat{\eta})\Big)\Big]\Big\} \tag{116}$$

$$=\mathbb{P}_n\Big\{\sum_a \pi_{a,X}\Big[\hat{Q}^{+,*}_{a,X} - \hat{e}_{a,X}\Big(b^-\hat{\underline{\mu}}^+_{a,X} + b^+\hat{\bar{\mu}}^+_{a,X}\Big)\Big] + \pi_{A,X}\Big(b^-\hat{\underline{\mu}}^+_{A,X} + b^+\hat{\bar{\mu}}^+_{A,X}\Big) \tag{117}$$

$$+ \frac{\pi_{A,X}}{\hat{e}_{A,X}}\Big[\Big(\hat{c}^-_{A,X} - \hat{c}^+_{A,X}\Big)\Big(\hat{F}^{-1}_{X,A}(\alpha^+)(\hat{\underline{\Delta}}^+_{Y,A,X} - \alpha^+)\Big) + \hat{c}^-_{A,X}\Big(Y\hat{\underline{\Delta}}^+_{Y,A,X} - \hat{\underline{\mu}}^+_{A,X}\Big) + \hat{c}^+_{A,X}\Big(Y\hat{\bar{\Delta}}^+_{Y,A,X} - \hat{\bar{\mu}}^+_{A,X}\Big)\Big]\Big\}$$

using our short-hand notation from the main paper.

□

D.3 Learning guarantee: Value function

Theorem D.3. *Assume Y is bounded by a constant C_y, i.e. $|Y| \leq C_y$. Then, for any policy $\pi \in \Pi$, it holds, with probability $1-\delta$, that*

$$V(\pi) \leq \hat{V}^{+,*}(\pi) + 2C_v\Big(\mathcal{R}_n(\Pi) + \frac{5}{2}\sqrt{\frac{1}{2n}\log\Big(\frac{2}{\delta}\Big)}\Big), \tag{118}$$

where $C_v = 2C_y(1+\Gamma^{-1}+\Gamma)$ and $\mathcal{R}_n(\pi)$ is the empirical Rademacher complexity of policy class Π.

Proof. We start by bounding the sharp upper bound $V^{+,*}(\pi)$ of the value function. By assumption, we have that $|Y| \leq C_y$. Hence, for any $\pi \in \Pi$, we can bound $V^{+,*}(\pi)$ via

$$\begin{aligned}
&|V^{+,*}(\pi)| && (119)\\
=&\Big|\int_{\mathcal{X}}\sum_a Q^{+,*}(a,x)\pi(a \mid x)\,\mathrm{d}p(x)\Big| && (120)\\
\leq&\Big|\sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} Q^{+,*}(a,x)\Big| && (121)\\
=&\sup_{(x,a)\in\mathcal{X}\times\mathcal{A}}\Big|\Big((1-\Gamma^{-1})e(a,x)+\Gamma^{-1}\Big)\mathbb{E}\Big[Y\mathbb{1}_{\{Y\leq F^{-1}_{x,a}(\alpha^+)\}} \mid X=x, A=a\Big] && (122)\\
&+\Big((1-\Gamma)e(a,x)+\Gamma\Big)\mathbb{E}\Big[Y\mathbb{1}_{\{Y\geq F^{-1}_{x,a}(\alpha^+)\}} \mid X=x, A=a\Big]\Big| && (123)\\
\leq&\sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} C_y\Big(\Big((1-\Gamma^{-1})e(a,x)+\Gamma^{-1}\Big)+\Big((1-\Gamma)e(a,x)+\Gamma\Big)\Big) && (124)\\
\leq&C_y\Big(2+2\Gamma^{-1}+2\Gamma\Big) && (125)\\
=&C_v. && (126)
\end{aligned}$$

Now, for the main result, we seek to find an upper bound for

$$\sup_{\pi\in\Pi}\hat{V}^{+,*}(\pi) - V(\pi). \tag{127}$$

By adding a zero and sublinearity of the supremum operator, we have that

$$\sup_{\pi\in\Pi}\hat{V}^{+,*}(\pi) - V(\pi) \leq \sup_{\pi\in\Pi}\Big(\hat{V}^{+,*}(\pi) - V^{+,*}(\pi)\Big) + \sup_{\pi\in\Pi}\Big(V^{+,*}(\pi) - V(\pi)\Big). \tag{128}$$

Further, by validity of our bounds, we know that

$$\sup_{\pi\in\Pi}V^{+,*}(\pi) - V(\pi) \geq 0, \tag{129}$$

such that we only need to focus on

$$D = \sup_{\pi\in\Pi}\hat{V}^{+,*}(\pi) - V^{+,*}(\pi). \tag{130}$$

Since

$$\hat{V}^{+,*}(\pi) = \frac{1}{n}\sum_{i=1}^n V_i^{+,*}(\pi) = \frac{1}{n}\sum_{i=1}^n\sum_a Q^{+,*}(X_i,a)\pi(a \mid X_i) \tag{131}$$

is a sample average with $|V_i^{+,*}(\pi)| \leq C_v$, we know that D satisfies the bounded difference with C_v/n. Hence, we can apply McDiarmid's inequality (McDiarmid, 1989), which yields

$$\mathbb{P}\Big(D - \mathbb{E}[D] \geq \epsilon\Big) \leq \underbrace{\exp\Big(-\frac{2\epsilon^2 n}{C_v^2}\Big)}_{=p_1}. \tag{132}$$

Then, solving for ϵ gives us

$$\epsilon = \sqrt{\frac{C_v^2}{2n} \log\Big(\frac{1}{p_1}\Big)}. \tag{133}$$

Hence, we know, with probability at least $1 - p_1$, that

$$D \leq \mathbb{E}[D] + \sqrt{\frac{C_v^2}{2n} \log\Big(\frac{1}{p_1}\Big)}. \tag{134}$$

As in (Athey & Wager, 2021; Frauen et al., 2024b; Hatt et al., 2022; Kallus & Zhou, 2018a), we express the flexibility of our policy class Π in terms of the Rademacher complexity. For this, we first note that a standard symmetrization argument yields

$$\mathbb{E}[D] \leq \mathbb{E}\left[\frac{1}{2^n} \sum_{\sigma \in \{-1,+1\}} \sup_{\pi \in \Pi} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i V_i^{+,*}(\pi) \right| \right], \tag{135}$$

where $\sigma_i \sim_{\text{iid}} \text{Unif}\{-1, +1\}$. Then, using the Rademacher-comparison lemma (Ledoux & Talagrand, 1989), we have that

$$\mathbb{E}[D] \leq 2C_v \mathbb{E}\Big[\mathcal{R}_n(\Pi)\Big], \tag{136}$$

where

$$\mathcal{R}_n(\Pi) = \mathbb{E}_\sigma \left[\sup_{\pi \in \Pi} \frac{1}{n} \sum_{i=1}^{n} \sigma_i V_i^{+,*} \right] \tag{137}$$

is the empirical Rademacher complexity of policy class Π. Again, $\mathcal{R}_n(\Pi)$ satisfies bounded difference with $2/n$, such that we can apply McDiarmid's inequality (McDiarmid, 1989). This gives

$$\mathbb{P}\Big(\mathcal{R}_n(\Pi) - \mathbb{E}[\mathcal{R}_n(\Pi)] \geq \epsilon\Big) \leq \underbrace{\exp\Big(-\frac{\epsilon^2 n}{2}\Big)}_{=p_2}. \tag{138}$$

Solving for ϵ yields

$$\epsilon = \sqrt{\frac{2}{n} \log\Big(\frac{1}{p_2}\Big)}, \tag{139}$$

such that, with probability at least $1 - p_2$, we have that

$$\mathbb{E}\Big[\mathcal{R}_n(\Pi)\Big] \leq \mathcal{R}_n(\Pi) + \sqrt{\frac{2}{n} \log\Big(\frac{1}{p_2}\Big)}. \tag{140}$$

Combining Equation 140 with our previous result in Equation 134, we have, with probability of at least $1 - p_1 - p_2$, that

$$\sup_{\pi \in \Pi} \Big(\hat{V}^{+,*}(\pi) - V^{+,*}(\pi)\Big) \tag{141}$$

$$\leq 2C_v \mathcal{R}_n(\Pi) + \sqrt{\frac{C_v^2}{2n} \log\Big(\frac{1}{p_1}\Big)} + 2C_v \sqrt{\frac{2}{n} \log\Big(\frac{1}{p_2}\Big)} \tag{142}$$

$$= 2C_v \Big(\mathcal{R}_n(\Pi) + \sqrt{\frac{1}{8n} \log\Big(\frac{1}{p_1}\Big)} + \sqrt{\frac{2}{n} \log\Big(\frac{1}{p_2}\Big)}\Big). \tag{143}$$

Now let $p_1 = p_2 = \delta/2$. Then, we know that with probability at least $1 - \delta$,

$$\sup_{\pi \in \Pi} \Big(\hat{V}^{+,*}(\pi) - V^{+,*}(\pi)\Big) \tag{144}$$

$$\leq 2C_v \Big(\mathcal{R}_n(\Pi) + \sqrt{\frac{1}{8n} \log\Big(\frac{1}{\delta}\Big)} + \sqrt{\frac{2}{n} \log\Big(\frac{1}{\delta}\Big)}\Big) \tag{145}$$

$$= 2C_v \Big(\mathcal{R}_n(\Pi) + \frac{5}{2} \sqrt{\frac{1}{2n} \log\Big(\frac{2}{\delta}\Big)}\Big), \tag{146}$$

or, equivalently,

$$V^{+,*}(\pi) \leq \hat{V}^{+,*}(\pi) + 2C_v\Big(\mathcal{R}_n(\Pi) + \frac{5}{2}\sqrt{\frac{1}{2n}\log\Big(\frac{2}{\delta}\Big)}\Big) \tag{147}$$

for all $\pi \in \Pi$, which concludes the proof. □

D.4 BOUND OF THE REGRET FUNCTION

Corollary D.4. *Given $Q^{\pm,*}(a,x)$ and our sensitivity constraints $\mathcal{P}(\Gamma)$ as in Proposition 4.1, an upper bound for the regret function $R_{\pi_0}(\pi)$ is given by*

$$R^+_{\pi}(\pi) = \int_{\mathcal{X}} \sum_a \Big(Q^{+,*}(a,x)\pi(a \mid x) - Q^{-,*}(a,x)\pi_0(a \mid x) \Big) \, \mathrm{d}p(x). \tag{148}$$

Proof. Our proof follows similar steps as in the proof of Proposition 4.1. For clarity, we repeat the same steps such that the proof is self-contained.

Again, we start by noting that the upper bound on the regret function depends on the choice of our sensitivity constraints and, hence, the set of distributions $\mathcal{P}(\Gamma)$. Therefore, we can write that

$$R^+_{\pi_0}(\pi) = R^+_{\pi_0}(\pi, \mathcal{P}(\Gamma)). \tag{149}$$

Following similar steps as in Proposition 4.1, we can write

$$\begin{aligned}
& R^+_{\pi_0}(\pi) && (150)\\
=& R^+_{\pi_0}(\pi, \mathcal{P}(\Gamma)) && (151)\\
=& \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} R_{\pi_0}(\pi, \tilde{p}) && (152)\\
=& \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} \Big(V(\pi, \tilde{p}) - V(\pi_0, \tilde{p}) \Big) && (153)\\
=& \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} \int_{\mathcal{X}} \sum_a \Big(Q(a,x,\tilde{p})\pi(a \mid x) - Q(a,x,\tilde{p})\pi_0(a \mid x) \Big) \, \mathrm{d}\tilde{p}(x) && (154)\\
=& \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} \int_{\mathcal{X}} \sum_a \Big(Q(a,x,\tilde{p})\pi(a \mid x) - Q(a,x,\tilde{p})\pi_0(a \mid x) \Big) \, \mathrm{d}p(x), && (155)
\end{aligned}$$

where Equation 155 again follows from $p(\mathcal{D}) = \tilde{p}(\mathcal{D})$ for all $\tilde{p} \in \mathcal{P}(\Gamma)$.

Since the optimal bounds $Q^{\pm,*}(a,x)$ are those $Q(a,x,\tilde{p})$, $\tilde{p} \in \mathcal{P}(\Gamma)$, for which the supremum/infimum are attained, we have that

$$Q(a,x,\tilde{p}) \leq \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} Q(a,x,\tilde{p}) = Q^{+,*}(a,x,\mathcal{P}(\Gamma)) \tag{156}$$

and

$$Q(a,x,\tilde{p}) \geq \inf_{p \in \mathcal{P}(\Gamma)} Q(a,x,\tilde{p}) = Q^{-,*}(a,x,\mathcal{P}(\Gamma)) \tag{157}$$

for all $\tilde{p} \in \mathcal{P}(\Gamma)$. Then, since $Q^{\pm,*}(a,x) \in L^1(\pi, p)$, it follows by dominated convergence that

$$\begin{aligned}
& \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} \int_{\mathcal{X}} \sum_a \Big(Q(a,x,\tilde{p})\pi(a \mid x) - Q(a,x,\tilde{p})\pi_0(a \mid x) \Big) \, \mathrm{d}p(x) && (158)\\
=& \int_{\mathcal{X}} \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} \sum_a \Big(Q(a,x,\tilde{p})\pi(a \mid x) - Q(a,x,\tilde{p})\pi_0(a \mid x) \Big) \, \mathrm{d}p(x) && (159)\\
\leq& \int_{\mathcal{X}} \sum_a \sup_{\tilde{p} \in \mathcal{P}(\Gamma)} Q(a,x,\tilde{p})\pi(a \mid x) \, \mathrm{d}p(x) - \int_{\mathcal{X}} \sum_a \inf_{\tilde{p} \in \mathcal{P}(\Gamma)} Q(a,x,\tilde{p})\pi_0(a \mid x) \, \mathrm{d}p(x) && (160)\\
=& \int_{\mathcal{X}} \sum_a \Big(Q^{+,*}(a,x)\pi(a \mid x) - Q^{-,*}(a,x)\pi_0(a \mid x) \Big) \, \mathrm{d}p(x), && (161)
\end{aligned}$$

where Equation 160 follows from the sublinearity of the supremum/infimum operator. □

D.5 Efficient estimator of the regret bound

Corollary D.5. *The efficient estimator for the upper bound of the regret function is given by*

$$\hat{R}^{+}_{\pi_0}(\pi) \tag{162}$$

$$= \sum_{\pm\in\{-,+\}} \pm\mathbb{P}_n\Big\{ \sum_a \pi^{\pm}_{a,X}\Big[\hat{Q}^{\pm,*}_{a,X} - \hat{e}_{a,X}\Big(b^{\mp}\hat{\underline{\mu}}^{\pm}_{a,X} + b^{\pm}\hat{\bar{\mu}}^{\pm}_{a,X}\Big)\Big] + \pi^{\pm}_{A,X}\Big(b^{\mp}\hat{\underline{\mu}}^{\pm}_{A,X} + b^{\pm}\hat{\bar{\mu}}^{\pm}_{A,X}\Big) \tag{163}$$

$$+ \frac{\pi^{\pm}_{A,X}}{\hat{e}_{A,X}}\Big[\Big(\hat{c}^{\mp}_{A,X} - \hat{c}^{\pm}_{A,X}\Big)\Big(\hat{F}^{-1}_{X,A}(\alpha^{\pm})(\hat{\underline{\Delta}}^{\pm}_{Y,A,X} - \alpha^{\pm})\Big) + \hat{c}^{\mp}_{A,X}\Big(Y\hat{\underline{\Delta}}^{\pm}_{Y,A,X} - \hat{\underline{\mu}}^{\pm}_{A,X}\Big) + \hat{c}^{\pm}_{A,X}\Big(Y\hat{\bar{\Delta}}^{\pm}_{Y,A,X} - \hat{\bar{\mu}}^{\pm}_{A,X}\Big)\Big]\Big\}, \tag{164}$$

where we use $\pi^{+} = \pi$ and $\pi^{-} = \pi_0$ for readability.

Proof. The upper bound of the regret function is given by

$$R^{+}_{\pi_0} = V^{+,*}(\pi) - V^{-,*}(\pi_0), \tag{165}$$

where

$$V^{\pm,*}(\pi) = \int_{\mathcal{X}} \sum_a Q^{\pm,*}(a,x)\pi(a \mid x)\,\mathrm{d}p(x). \tag{166}$$

By additivity of the efficient influence function, we know that

$$\mathbb{IF}\Big(R^{+,*}_{\pi_0}(\pi)\Big) = \mathbb{IF}\Big(V^{+,*}(\pi)\Big) - \mathbb{IF}\Big(V^{-,*}(\pi_0)\Big), \tag{167}$$

such that we can focus on both terms separately in Supplements D.5.1 and D.5.2 and then plug them together in Supplement D.5.3 in order to obtain our efficient estimator.

D.5.1 Efficient influence function of $V^{+,*}(\pi)$

We already have the efficient influence function of $V^{+,*}(\pi)$ from the proof of Theorem 4.3 in Supplement D.2. It is given by

$$\mathbb{IF}\Big(V^{+,*}(\pi;\eta)\Big) \tag{168}$$

$$= -V^{+,*}(\pi) + \sum_a \pi(a \mid X)\Big[Q^{+,*}(a,X) - e(a,X)\Big(b^{-}\underline{\mu}^{+}(a,X;\eta) + b^{+}\bar{\mu}^{+}(a,X;\eta)\Big)\Big] \tag{169}$$

$$+ \pi(A \mid X)\Big(b^{-}\underline{\mu}^{+}(A,X;\eta) + b^{+}\bar{\mu}^{+}(A,X;\eta)\Big) \tag{170}$$

$$+ \frac{\pi(A \mid X)}{e(A,X)}\Big[\Big(c^{-}(A,X;\eta) - c^{+}(A,X;\eta)\Big)\Big(F^{-1}_{X,A}(\alpha^{+})(\underline{\Delta}_{\alpha^{+}}(Y,A,X;\eta) - \alpha^{+})\Big) \tag{171}$$

$$+ c^{-}(A,X;\eta)\Big(Y\underline{\Delta}^{+}(Y,A,X;\eta) - \underline{\mu}^{+}(A,X;\eta)\Big) + c^{+}(A,X;\eta)\Big(Y\bar{\Delta}^{+}(Y,A,X;\eta) - \bar{\mu}^{+}(A,X;\eta)\Big)\Big]. \tag{172}$$

D.5.2 Efficient influence function of $V^{-,*}(\pi)$

We can derive the sharp lower bound for the value function analogously to $V^{+,*}$. Again, we make use of the chain rule for deriving efficient influence function (Kennedy, 2022), a proof of which can be found in (Luedtke, 2024) **(Lemma S3)**.

For this, let $\alpha^- = 1/(1+\Gamma)$. Then, the sharp lower bound of the CAPO is given by

$$
\begin{aligned}
& Q^{-,*}(a,x;\eta) && (173)\\
=& \Big((1-\Gamma)e(a,x)+\Gamma\Big)\mathbb{E}\Big[Y\mathbb{1}_{\{Y\leq F^{-1}_{X,A}(\alpha^-)\}} \mid X=x, A=a\Big] && (174)\\
& + \Big((1-\Gamma^{-1})e(a,x)+\Gamma^{-1}\Big)\mathbb{E}\Big[Y\mathbb{1}_{\{Y\geq F^{-1}_{X,A}(\alpha^-)\}} \mid X=x, A=a\Big]. && (175)
\end{aligned}
$$

Hence, the influence function of $Q^{-,*}(a,x)$ is given by

$$
\begin{aligned}
& \mathbb{IF}\Big(Q^{-,*}(a,x;\eta)\Big) && (176)\\
=& \frac{\mathbb{1}_{\{X=x\}}}{p(x)}(1-\Gamma)\Big(\mathbb{1}_{\{A=a\}}-e(a,x)\Big)\mathbb{E}\Big[Y\mathbb{1}_{\{Y\leq F^{-1}_{X,A}(\alpha^-)\}} \mid X=x, A=a\Big] && (177)\\
& + \frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big((1-\Gamma)e(a,x)+\Gamma\Big) && (178)\\
& \quad \times \Big(Y\mathbb{1}_{\{Y\leq F^{-1}_{x,a}(\alpha^-)\}} - \mathbb{E}[Y\mathbb{1}_{\{Y\leq F^{-1}_{X,A}(\alpha^-)\}} \mid X=x, A=a] + F^{-1}_{x,a}(\alpha^-)(\mathbb{1}_{\{Y\leq F^{-1}_{x,a}(\alpha^-)\}}-\alpha^-)\Big) && (179)\\
& + \frac{\mathbb{1}_{\{X=x\}}}{p(x)}(1-\Gamma^{-1})\Big(\mathbb{1}_{\{A=a\}}-e(a,x)\Big)\mathbb{E}\Big[Y\mathbb{1}_{\{Y\geq F^{-1}_{X,A}(\alpha^-)\}} \mid X=x, A=a\Big] && (180)\\
& + \frac{\mathbb{1}_{\{X=x,A=a\}}}{p(a,x)}\Big((1-\Gamma^{-1})e(a,x)+\Gamma^{-1}\Big) && (181)\\
& \quad \times \Big(Y\mathbb{1}_{\{Y\geq F^{-1}_{x,a}(\alpha^-)\}} - \mathbb{E}[Y\mathbb{1}_{\{Y\geq F^{-1}_{X,A}(\alpha^-)\}} \mid X=x, A=a] - F^{-1}_{x,a}(\alpha^-)(\mathbb{1}_{\{Y\leq F^{-1}_{x,a}(\alpha^-)\}}-\alpha^-)\Big) && (182)\\
=& \frac{\mathbb{1}_{\{X=x\}}}{p(x)}\Big\{\Big(\mathbb{1}_{\{A=a\}}-e(a,x)\Big)\Big(b^+\underline{\mu}^-(a,x;\eta)+b^-\bar{\mu}^-(a,x;\eta)\Big) && (183)\\
& + \frac{\mathbb{1}_{\{A=a\}}}{e(a,x)}\Big[\Big(c^+(a,x;\eta)-c^-(a,x;\eta)\Big)\Big(F^{-1}_{x,a}(\alpha^-)(\underline{\Delta}^-(Y,a,x;\eta)-\alpha^-)\Big) && (184)\\
& + c^+(a,x;\eta)\Big(Y\underline{\Delta}^-(Y,a,x;\eta)-\underline{\mu}^-(a,x;\eta)\Big)+c^-(a,x;\eta)\Big(Y\bar{\Delta}^-(Y,a,x;\eta)-\bar{\mu}^-(a,x;\eta)\Big)\Big]\Big\} && (185)
\end{aligned}
$$

Following the same steps as for Equation 38, the influence function of $V^{-,*}(\pi)$ is given by

$$
\begin{aligned}
& \mathbb{IF}\Big(V^{-,*}(\pi;\eta)\Big) && (186)\\
=& -V^{-,*}(\pi;\eta)+\sum_a \pi(a\mid X)\Big[Q^{-,*}(a,X;\eta)-e(a,X)\Big(b^+\underline{\mu}^-(a,X;\eta)+b^-\bar{\mu}^-(a,X;\eta)\Big)\Big] && (187)\\
& + \pi(A\mid X)\Big(b^+\underline{\mu}^-(A,X;\eta)+b^-\bar{\mu}^-(A,X;\eta)\Big) && (188)\\
& + \frac{\pi(A\mid X)}{e(A,X)}\Big[\Big(c^+(A,X;\eta)-c^-(A,X;\eta)\Big)\Big(F^{-1}_{X,A}(\alpha^-)(\underline{\Delta}^-(Y,A,X;\eta)-\alpha^-)\Big) && (189)\\
& + c^+(A,X;\eta)\Big(Y\underline{\Delta}^-(Y,A,X;\eta)-\underline{\mu}^-(A,X;\eta)\Big)+c^-(A,X;\eta)\Big(Y\bar{\Delta}^-(Y,A,X;\eta)-\bar{\mu}^-(A,X;\eta)\Big)\Big]. && (190)
\end{aligned}
$$

D.5.3 EFFICIENT ESTIMATOR OF $R^+_{\pi_0}(\pi)$

We can derive the efficient estimator for the bounds of the value function through one-step bias correction using our results form Supplements D.5.1 and D.5.2 via

$$V^{\pm,*}(\pi;\hat{\eta}) + \mathbb{P}_n\Big\{V^{\pm,*}(\pi;\hat{\eta})\Big\} \tag{191}$$

$$=\mathbb{P}_n\Big\{\sum_a \pi(a \mid X)\Big[Q^{\pm,*}(a,X;\hat{\eta}) - \hat{e}(a,X)\Big(b^{\mp}\hat{\underline{\mu}}^{\pm}(a,X;\hat{\eta}) + b^{\pm}\hat{\bar{\mu}}^{\pm}(a,X;\hat{\eta})\Big)\Big] \tag{192}$$

$$+ \pi(A \mid X)\Big(b^{\mp}\hat{\underline{\mu}}^{\pm}(A,X;\hat{\eta}) + b^{\pm}\hat{\bar{\mu}}^{\pm}(A,X;\hat{\eta})\Big) \tag{193}$$

$$+ \frac{\pi(A \mid X)}{\hat{e}(A,X)}\Big[\Big(c^{\mp}(A,X;\hat{\eta}) - c^{\pm}(A,X;\hat{\eta})\Big)\Big(\hat{F}^{-1}_{X,A}(\alpha^{\pm})(\underline{\Delta}^{\pm}(Y,A,X;\hat{\eta}) - \alpha^{\pm})\Big) \tag{194}$$

$$+ c^{\mp}(A,X;\hat{\eta})\Big(Y\underline{\Delta}^{\pm}(Y,A,X;\hat{\eta}) - \hat{\underline{\mu}}^{\pm}(A,X;\hat{\eta})\Big) + c^{\pm}(A,X;\hat{\eta})\Big(Y\bar{\Delta}^{\pm}(Y,A,X;\hat{\eta}) - \hat{\bar{\mu}}^{\pm}(A,X;\hat{\eta})\Big)\Big]\Big\} \tag{195}$$

$$=\mathbb{P}_n\Big\{\sum_a \pi_{a,X}\Big[\hat{Q}^{\pm,*}_{a,X} - \hat{e}_{a,X}\Big(b^{\mp}\hat{\underline{\mu}}^{\pm}_{a,X} + b^{\pm}\hat{\bar{\mu}}^{\pm}_{a,X}\Big)\Big] + \pi_{A,X}\Big(b^{\mp}\hat{\underline{\mu}}^{\pm}_{A,X} + b^{\pm}\hat{\bar{\mu}}^{\pm}_{A,X}\Big) \tag{196}$$

$$+ \frac{\pi_{A,X}}{\hat{e}_{A,X}}\Big[\Big(\hat{c}^{\mp}_{A,X} - \hat{c}^{\pm}_{A,X}\Big)\Big(\hat{F}^{-1}_{X,A}(\alpha^{\pm})(\hat{\underline{\Delta}}^{\pm}_{Y,A,X} - \alpha^{\pm})\Big) + \hat{c}^{\mp}_{A,X}\Big(Y\hat{\underline{\Delta}}^{\pm}_{Y,A,X} - \hat{\underline{\mu}}^{\pm}_{A,X}\Big) + \hat{c}^{\pm}_{A,X}\Big(Y\hat{\bar{\Delta}}^{\pm}_{Y,A,X} - \hat{\bar{\mu}}^{\pm}_{A,X}\Big)\Big]\Big\} \tag{197}$$

using our short-hand notation from the main paper. Hence, the efficient estimator of the upper bound of the regret function is given by

$$\hat{R}^+_{\pi_0}(\pi) \tag{198}$$

$$=R^+_{\pi_0}(\pi;\hat{\eta}) + \mathbb{P}_n\Big\{\mathbb{IF}\Big(R^{+,*}_{\pi_0}(\pi;\hat{\eta})\Big)\Big\} \tag{199}$$

$$=\Big(V^{+,*}(\pi;\hat{\eta}) - V^{-,*}(\pi_0;\hat{\eta})\Big) + \mathbb{P}_n\Big\{\mathbb{IF}\Big(V^{+,*}(\pi;\hat{\eta})\Big) - \mathbb{IF}\Big(V^{-,*}(\pi_0;\hat{\eta})\Big)\Big\} \tag{200}$$

$$= \sum_{\pm\in\{+,-\}} \pm\mathbb{P}_n\Big\{\sum_a \pi^{\pm}_{a,X}\Big[\hat{Q}^{\pm,*}_{a,X} - \hat{e}_{a,X}\Big(b^{\mp}\hat{\underline{\mu}}^{\pm}_{a,X} + b^{\pm}\hat{\bar{\mu}}^{\pm}_{a,X}\Big)\Big] + \pi^{\pm}_{A,X}\Big(b^{\mp}\hat{\underline{\mu}}^{\pm}_{A,X} + b^{\pm}\hat{\bar{\mu}}^{\pm}_{A,X}\Big) \tag{201}$$

$$+ \frac{\pi^{\pm}_{A,X}}{\hat{e}_{A,X}}\Big[\Big(\hat{c}^{\mp}_{A,X} - \hat{c}^{\pm}_{A,X}\Big)\Big(\hat{F}^{-1}_{X,A}(\alpha^{\pm})(\hat{\underline{\Delta}}^{\pm}_{Y,A,X} - \alpha^{\pm})\Big) + \hat{c}^{\mp}_{A,X}\Big(Y\hat{\underline{\Delta}}^{\pm}_{Y,A,X} - \hat{\underline{\mu}}^{\pm}_{A,X}\Big) + \hat{c}^{\pm}_{A,X}\Big(Y\hat{\bar{\Delta}}^{\pm}_{Y,A,X} - \hat{\bar{\mu}}^{\pm}_{A,X}\Big)\Big]\Big\}, \tag{202}$$

where we let $\pi^+ = \pi$ and $\pi^- = \pi_0$ for readability. □

D.6 IMPROVEMENT GUARANTEE: REGRET FUNCTION

Corollary D.6. *Under the same assumption as in Theorem 4.4, for any policy $\pi \in \Pi$ and baseline policy $\pi_0 \in \Pi$, it holds, with probability $1-\delta$, that*

$$R_{\pi_0}(\pi) \leq \hat{R}^{+}_{\pi_0}(\pi) + 4C_v\Big(\mathcal{R}_n(\Pi) + \frac{5}{2}\sqrt{\frac{1}{2n}\log\Big(\frac{2}{\delta}\Big)}\Big), \tag{203}$$

where $C_v = 2C_y(1+\Gamma^{-1}+\Gamma)$ and $\mathcal{R}_n(\pi)$ is the empirical Rademacher complexity of policy class Π.

Proof. In order to show the main result, we note that

$$R^{+}_{\pi_0}(\pi) = V^{+,*}(\pi) - V^{-,*}(\pi_0) \tag{204}$$

for arbitrary $\pi, \pi_0 \in \Pi$.

From Theorem 4.4, we know that

$$V^{+,*}(\pi) \leq \hat{V}^{+,*}(\pi) + 2C_v\Big(\mathcal{R}_n(\Pi) + \frac{5}{2}\sqrt{\frac{1}{2n}\log\Big(\frac{2}{\delta}\Big)}\Big). \tag{205}$$

Since $\pi, \pi_0 \in \Pi$ are arbitrary, we can repeat the same arguments for $V^{-,*}(\pi_0)$ and obtain

$$V^{-,*}(\pi_0) \geq \hat{V}^{-,*}(\pi_0) - 2C_v\Big(\mathcal{R}_n(\Pi) + \frac{5}{2}\sqrt{\frac{1}{2n}\log\Big(\frac{2}{\delta}\Big)}\Big). \tag{206}$$

Then, we conclude the proof by

$$\begin{aligned}
& R_{\pi_0}(\pi) && (207)\\
=& V(\pi) - V(\pi_0) && (208)\\
\leq& V^{+,*}(\pi) - V^{-,*}(\pi_0) && (209)\\
=& \Big[\hat{V}^{+,*}(\pi) + 2C_v\Big(\mathcal{R}_n(\Pi) + \frac{5}{2}\sqrt{\frac{1}{2n}\log\Big(\frac{2}{\delta}\Big)}\Big)\Big] - \Big[\hat{V}^{-,*}(\pi_0) - 2C_v\Big(\mathcal{R}_n(\Pi) + \frac{5}{2}\sqrt{\frac{1}{2n}\log\Big(\frac{2}{\delta}\Big)}\Big)\Big] && (210)\\
=& \hat{R}^{+}_{\pi_0}(\pi) + 4C_v\Big(\mathcal{R}_n(\Pi) + \frac{5}{2}\sqrt{\frac{1}{2n}\log\Big(\frac{2}{\delta}\Big)}\Big). && (211)
\end{aligned}$$

□

E IMPLEMENTATION DETAILS

We summarize the neural instantiations of all estimators in Section 5.

Runtime: Training the first and second stage models for our method took in total approximately 20 seconds using $n = 1000$ synthetic data samples and a standard computer with AMD Ryzen 7 Pro CPU and 32GB of RAM.634. All baselines have a comparable runtime.

Nuisance function	**Hyperparameter**	Configuration	Standard methods		Kallus & Zhou (2018a; 2021)	Plug-in sharp (ours)	Efficient sharp (ours)
			IPW estimator	DR estimator			
Propensity score	Hidden layers	3	✓	✓	✓	✗	✓
	Layer size	$\{64, 64, 32\}$					
	Hidden activation	ReLU					
	Learning rate	0.001					
	Number of epochs	300					
	Early stopping patience	10					
	Batch size	64					
Conditional quantile function	Hidden layers	3	✗	✗	✗	✗	✓
	Layer size	$\{64, 64, 32\}$					
	Hidden activation	ReLU					
	Learning rate	0.001					
	Number of epochs	300					
	Early stopping patience	10					
	Batch size	64					
(Masked) CAPO model	Hidden layers	3	✗	✓	✗	✓	✓
	Layer size	$\{64, 64, 32\}$					
	Hidden activation	ReLU					
	Learning rate	0.001					
	Number of epochs	300					
	Early stopping patience	10					
	Batch size	64					
Parametric policy	Hidden layers	3	✓	✓	✓	✓	✓
	Layer size	$\{64, 64, 32\}$					
	Hidden activation	ReLU					
	Learning rate	0.001					
	Number of epochs	300					
	Early stopping patience	10					
	Batch size	64					

Table 3: Neural instantiations of estimated nuisance functions $\hat{\eta}$ and parametric policy π_θ. To ensure a fair comparison, all methods share the same nuisance function where applicable. For all models, we set the split parameter for training the nuisance and the policy model to $\rho = 0.5$.