Positivity-free Policy Learning with Observational Data

Pan Zhao
PreMeDICaL, Inria,
Université de Montpellier,
France

Antoine Chambaz
Université Paris Cité,
CNRS, MAP5,
F-75006 Paris, France

Julie Josse
PreMeDICaL, Inria,
Université de Montpellier,
France

Shu Yang
Department of Statistics,
NC State University,
United States

Abstract

Policy learning utilizing observational data is pivotal across various domains, with the objective of learning the optimal treatment assignment policy while adhering to specific constraints such as fairness, budget, and simplicity. This study introduces a novel positivity-free (stochastic) policy learning framework designed to address the challenges posed by the impracticality of the positivity assumption in real-world scenarios. This framework leverages incremental propensity score policies to adjust propensity score values instead of assigning fixed values to treatments. We characterize these incremental propensity score policies and establish identification conditions, employing semiparametric efficiency theory to propose efficient estimators capable of achieving rapid convergence rates, even when integrated with advanced machine learning algorithms. This paper provides a thorough exploration of the theoretical guarantees associated with policy learning and validates the proposed framework's finite-sample performance through comprehensive numerical experiments, ensuring the identification of causal effects from observational data is both robust and reliable.

1 INTRODUCTION

Over the past decade, methodologies for learning treatment assignment policies have seen substantial advancements in fields like biostatistics (Luedtke and van der Laan, 2016b; Tsiatis et al., 2019), computer science (Uehara et al., 2022; Yu et al., 2022), and econometrics (Athey and Wager, 2021; Jia et al., 2023). The core objective of data-driven policy learning is to learn optimal policies that map individual characteristics to treatment assignments to optimize some utility or outcome functions. This is crucial for deriving robust and trustworthy policies in high-stakes decision-making settings, requiring adherence to standard causal assumptions: consistency, unconfoundedness, and positivity (van der Laan et al., 2011; Imbens and Rubin, 2015).

Various statistical and machine-learning methods have been developed to address policy learning tasks. Popular approaches include model-based methods such as Q-learning and A-learning (Murphy, 2003; Shi et al., 2018), and direct model-free policy search methods such as decision trees and outcome weighted learning (Zhang et al., 2012; Cui et al., 2017), among others (Bibaut et al., 2021; Zhou et al., 2023; Zhao and Cui, 2023). Another prevailing line of work concerns heterogeneous treatment effects estimation (Wager and Athey, 2018; Künzel et al., 2019; Nie and Wager, 2021; Kallus and Oprescu, 2023), where the sign of the conditional average treatment effects equivalently determines the optimal policy.

However, most methods depend heavily on the three standard causal assumptions to identify causal effects and optimal policies. Recent progress has been made to relax the consistency and unconfoundedness assumptions (Cortez et al., 2022; Kallus and Zhou, 2018), but advancements addressing the violation of the positivity assumption are scarce. For average treatment effects estimation, Yang and Ding (2018) and Branson et al. (2023) provide estimation and asymptotic inference results for propensity score trimming with binary and continuous treatments, Zhang et al. (2023) consider a missing-at-random mechanism without a positivity condition for generalizable and double robust inference for average treatment effects under selection bias with decaying overlap, Liu et al.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

(2023) propose the overlap weighted average treatment effect on the treated under lack of positivity, and Visconti and Zubizarreta (2018) use cardinality matching to handle limited overlap in observational studies. For policy evaluation, Khan et al. (2023) provide partial identification results for off-policy evaluation under non-parametric Lipschitz smoothness assumptions on the conditional mean function, and thus avoid assuming either overlap or a well-specified model. In the machine learning literature, Lawrence et al. (2017) consider counterfactual learning from deterministic bandit logs under lack of sufficient exploration. Gui and Veitch (2023) use supervised representation learning to estimate causal effects for text data with apparent overlap violation. Jin et al. (2022) use pessimism and generalized empirical Bernstein's inequality to study offline policy learning without assuming any uniform overlap condition. To our knowledge, our work is the first to consider learning treatment assignment policies while avoiding the positivity assumption.

This study introduces a novel positivity-free policy learning framework focusing on dynamic and stochastic policies, which are practical. We propose *incremental propensity score policies* that shift propensity scores by an individualized parameter, requiring only the consistency and unconfoundedness causal assumptions. Our approach enhances the concept of incremental intervention effects, as proposed by Kennedy (2019), adapting it to individual treatment policy contexts.

We also use semiparametric theory to characterize the efficient influence function (Bickel et al., 1993; van der Laan and Robins, 2003), which serves as the foundation to construct estimators with favorable properties, such as double/multiple robustness and asymptotically negligible second-order bias (also called Neyman orthogonality in double machine learning (Chernozhukov et al., 2018) or orthogonal statistical learning (Foster and Syrgkanis, 2023)). Thus, our proposed estimators can attain fast parametric $\sqrt{n}$ convergence rates, even when nuisance parameters are estimated at slower rates such as $n^{1/4}$ via flexible machine learning algorithms.

Based on the above efficient off-policy evaluation results, we propose approaches to learning the optimal policy by maximizing the value function, possibly under application-specific constraints. Several examples are provided in Section 4, including fairness and resource limit. While it remains an open problem to provide finite sample or asymptotic regret bounds as Athey and Wager (2021) for stochastic policy learning with constraints, which is out of the scope of this article, we establish asymptotic guarantees for our proposed policy learning methods under alternative (stronger) conditions.

2 STATISTICAL FRAMEWORK

We first introduce the notations and setup. Let X denote the p-dimensional vector of covariates that belongs to a covariate space $\mathcal{X} \subset \mathbb{R}^p$, $A \in \mathcal{A} = \{0, 1\}$ denote the binary treatment, $Y \in \mathbb{R}$ denote the outcome of interest. Without loss of generality, we assume throughout that larger values of Y are more desirable. Our observed data structure is $O = (X, A, Y)$. Suppose that our collected random sample $(O_1, \ldots, O_n)$ of size n are independent and identically distributed (i.i.d.) observations of $O \sim P$, where P denote the true distribution of the observed data.

Now, we are in the position to introduce different types of policies or interventions commonly used in the literature: (i) under *static* policies, the same treatments would be applied indiscriminately, while *dynamic* policies depend on individual characteristics; (ii) *deterministic* policies recommend one specific treatment and *stochastic* policies output probabilities of prescribing each treatment level. This article focuses on dynamic and stochastic policies, which are more practical in various settings and have received substantial recent interest. Typical examples include point exposures (Dudík et al., 2014), longitudinal studies (Tian, 2008; Murphy et al., 2001; van der Laan and Petersen, 2007), natural stochastic policies in reinforcement learning (Kallus and Uehara, 2020), and particularly interventions that depend on the observational treatment process (Muñoz and van Der Laan, 2012; Haneuse and Rotnitzky, 2013; Young et al., 2014); but none of the existing intervention effects both avoids positivity conditions entirely and is completely nonparametric.

We use the potential outcomes framework (Neyman, 1923; Rubin, 1974) to define causal effects. Let $Y(a)$ denote the potential outcome had the treatment a been assigned. A policy $d : \mathcal{X} \to \{0, 1\}$ is deterministic if it maps individual characteristics x to a treatment assignment 0 or 1, and the output of a stochastic policy $d : \mathcal{X} \to [0, 1]$ is the probability of assigning treatment 1. Let $\mathcal{D}$ denote a pre-specified class of policies of interest, where each policy $d \in \mathcal{D}$ induces the value function defined by

$$V(d) = E[Y(d)] = E[Y(1)d(X) + Y(0)(1 - d(X))],$$

where $Y(d)$ is the potential outcome under the policy d. In Remark 1, we briefly review standard (deterministic) policy learning methods. In our framework, we focus on dynamic and stochastic policies. Our goal is to directly search for the optimal policy d^* that maximizes the value function $V(d)$, possibly under application-specific constraints $c(d) \leq 0$. See

Section 4 for detailed examples.

2.1 Causal Assumptions

We make the following identification assumptions.

Assumption 1 (Consistency). $Y = Y(A)$.

Assumption 2 (Unconfoundedness). $A \perp Y(a) \mid X$ *for* $a = 0, 1$.

Assumption 1 is also known as the stable unit treatment value assumption, which says there should be no multiple versions of the treatment and no interference between units. Assumption 2 states that there are no unmeasured confounders so that treatment assignment is as good as random conditional on the covariates X. In this article, we entirely avoid the positivity assumption which requires that each unit has a positive probability of receiving both treatment levels, i.e., $c < Pr(A = 1 \mid X) < 1 - c$ for some constant $c > 0$.

Remark 1. Standard policy learning methods need all of Assumptions 1, 2 and the positivity assumption to identify the value function of deterministic policies $d : \mathcal{X} \to \mathcal{A}$ by the outcome regression (OR), inverse probability weighting (IPW) and augmented IPW (AIPW) formulas:

$$
\begin{aligned}
V_{\mathrm{OR}}(d) &= E[E[Y \mid X, A = d(X)]], \\
V_{\mathrm{IPW}}(d) &= E\left[\frac{I\{A = d(X)\}Y}{Pr(A = d(X) \mid X)}\right], \\
V_{\mathrm{AIPW}}(d) &= E\left[E[Y \mid X, A = d(X)] + \right. \\
&\qquad \left. \frac{I\{A = d(X)\}(Y - E[Y \mid X, A = d(X)])}{Pr(A = d(X) \mid X)}\right],
\end{aligned}
$$

thus the optimal policies are given by $d^*_{\mathrm{OR}} = \arg\max_{d \in \mathcal{D}} V_{\mathrm{OR}}(d)$, $d^*_{\mathrm{IPW}} = \arg\max_{d \in \mathcal{D}} V_{\mathrm{IPW}}(d)$, and $d^*_{\mathrm{AIPW}} = \arg\max_{d \in \mathcal{D}} V_{\mathrm{AIPW}}(d)$, possibly under application-specific constraints. When the positivity is violated, it is error-prone to rely on the outcome regression model's extrapolation, and the IPW and AIPW estimators would fail due to division by zero.

2.2 Incremental Propensity Score Policies

Kennedy (2019) propose a new class of stochastic dynamic intervention, called incremental propensity score interventions, and show that these interventions are nonparametrically identified without requiring any positivity restrictions on the propensity scores. Specifically, their proposed intervention replaces the observational propensity score $\pi(x) = Pr(A = 1 \mid X = x)$ with a shifted version based on multiplying the odds of receiving treatment, $\delta\pi(x)/\{\delta\pi(x) + 1 - \pi(x)\}$, where the increment parameter $\delta \in (0, \infty)$ is user-specified and dictates the extent to which the propensity scores fluctuate from their actual observational values. Some motivation and examples, efficiency theory, and estimators for mean outcomes under these interventions are studied in detail by Kennedy (2019).

We propose a positivity-free (stochastic) policy learning framework based on the incremental propensity score interventions. Specifically, we consider the stochastic policy $d : \mathcal{X} \to [0, 1]$ that assigns treatment 1 with probability

$$
d(x) = \frac{\delta(x)\pi(x)}{\delta(x)\pi(x) + 1 - \pi(x)}, \tag{1}
$$

where $\delta(x)$ enables individualized treatment assignment. We note that the choice of $d(x)$ in (1) is motivated by its interpretability and positivity-free. In particular, whenever $0 < \pi(x) < 1$, $\delta(x) = [d(x)/\{1 - d(x)\}]/[\pi(x)/\{1 - \pi(x)\}]$ is simply an odds ratio, indicating how the policy changes the odds of receiving treatment. When positivity is violated, we have that $d(x) = 0$ if $\pi(x) = 0$, and $d(x) = 1$ if $\pi(x) = 1$.

3 IDENTIFICATION AND EFFICIENCY THEORY

3.1 Identification

We first give formal identification results for the value function of incremental propensity score policies, which require no conditions on the propensity scores.

Proposition 1 (Identification formulas). *Under Assumptions 1 and 2, the value function $V(d)$ can be nonparametrically identified by the outcome regression with incremental propensity score (OR-IPS) formula:*

$$
V_{\mathrm{OR-IPS}}(d) = E\left[\frac{\delta(X)\pi(X)\mu_1(X) + \{1 - \pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)}\right], \tag{2}
$$

where $\mu_a(X) = E[Y \mid X, A = a], a = 0, 1$ are the outcome regression functions or the inverse probability weighting of incremental propensity score (IPW-IPS) formula:

$$
V_{\mathrm{IPW-IPS}}(d) = E\left[\frac{Y\{\delta(X)A + 1 - A\}}{\delta(X)\pi(X) + 1 - \pi(X)}\right]. \tag{3}
$$

In the potential outcomes framework, under non-binary policy d, the law of $Y(d)$ is given by: sample X; compute $d(X)$ and sample A from the Bernoulli distribution with success probability $d(X)$; set $Y(d) = Y(A)$. The value of d is then $E[Y(d)] = E[Y(1)d(X) + Y(0)(1 - d(X)]$. In Equation (3), $Y = Y(A)$. Proposition 1 shows that the value function can be identified by (i) a weighted average of the outcome regression functions μ_0, μ_1, where the weight on μ_1 is given by

the incremental propensity score $d(x)$ and the weight on μ_0 is $1-d(x)$; (ii) inverse probability weighting where each treated is weighted by the (inverse of the) propensity score plus some fractional contribution of its complement, i.e., $\pi(x)+(1-\pi(x))/\delta(x)$, and untreated units are weighted by this same amount, except the entire weight is further down-weighted by a factor of $\delta(x)$.

3.2 Efficient Off-policy Evaluation

Despite that simple plug-in OR-IPS and IPW-IPS estimators can be easily constructed from (2) and (3), these estimators will only be $\sqrt{n}$-consistent when the outcome regression or propensity score models are correctly specified. This is usually unrealistic in practice. We use semiparametric efficiency theory to study the following statistical functional of P from a nonparametric statistical model $\mathcal{M}$:

$$\Psi(P) = E_P\left[\frac{\delta(X)\pi(X)\mu_1(X)+\{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X)+1-\pi(X)}\right],$$

and propose efficient estimators based on the efficient influence function.

Proposition 2 (Semiparametric Efficiency)**.** *The efficient influence function of* $\Psi(P)$ *is*

$$\begin{aligned}\phi(P)(O) = &\frac{A\delta(X)\{Y-\mu_1(X)\}+(1-A)\{Y-\mu_0(X)\}}{\delta(X)\pi(X)+1-\pi(X)}\\ &+\frac{\delta(X)\pi(X)\mu_1(X)+\{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X)+1-\pi(X)}\\ &+\frac{\delta(X)\tau(X)\{A-\pi(X)\}}{\{\delta(X)\pi(X)+1-\pi(X)\}^2}-\Psi(P),\end{aligned} \tag{4}$$

where $\tau(x)=\mu_1(x)-\mu_0(x)$.

By Proposition 2, the one-step bias-corrected estimator is given by

$$\hat{\Psi}_{\mathrm{OS}} = \Psi(\hat{P}) + P_n\phi(\hat{P})(O) = \frac{1}{n}\sum_{i=1}^{n}\xi(\hat{P})(O_i), \tag{5}$$

where we estimate P by $\hat{P}$, and let P_n denote the empirical distribution, and $\xi(P)(O)=\phi(P)(O)+\Psi(P)$ is the uncentered efficient influence function. This estimator can converge at fast parametric $\sqrt{n}$ rates and attain the efficiency bound, even when the propensity score $\pi(x)$ and outcome regression functions μ_0,μ_1 are modeled flexibly and estimated at rates slower than $\sqrt{n}$, as long as these nuisance functions are estimated consistently at rates faster than $n^{1/4}$. This allows much more flexible nonparametric methods and modern machine learning algorithms to be employed.

However, characterizing asymptotic properties of the estimator (5) requires some empirical process conditions that restrict the flexibility and complexity of the nuisance estimators; otherwise, we will have overfitting bias and intractable asymptotic behaviors. See the asymptotic analysis in Section 5 and proofs thereof. To accommodate the wide use of modern machine learning algorithms that usually fail to satisfy the required empirical process conditions, we apply the cross-fitting procedure to obtain asymptotically normal and efficient estimators (Zheng and van der Laan, 2010; Chernozhukov et al., 2018). Suppose we randomly split the data into K folds. Then the cross-fitting estimator is

$$\hat{\Psi}_{\mathrm{CF}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\Psi}_k = \frac{1}{K}\sum_{k=1}^{K}P_{n,k}\xi(P_{n,-k})(O), \tag{6}$$

where $P_{n,k}$ and $P_{n,-k}$ denote the empirical measures on data from the k-fold and excluding the k-fold, respectively. That is, for $k=1,\ldots,K$, nuisance estimators are constructed excluding the k-fold, and the value function $\hat{\Psi}_k$ is evaluated on the k-th fold; finally, the cross-fitting estimator is the average of the K value estimators from K folds.

4 FROM EFFICIENT POLICY EVALUATION TO LEARNING

In this section, we first present our proposed methods for policy learning.

As discussed in Section 2, given a pre-specified policy class $\mathcal{D}$ (e.g., linear decision rules), we propose estimating the optimal treatment assignment rule $\hat{d}$ that solves (i) $\hat{d} = \arg\max_{d\in\mathcal{D}}\hat{V}(d)$, where $\hat{V}(d)$ is a value function estimator by OR-IPS (2), IPW-IPS (3), one-step (5) or cross-fitting (6); or (ii) $\hat{d} = \arg\max_{d\in\mathcal{D}}\hat{V}(d)$ subject to $\hat{c}(d)\le c$, when an application-specific constraint $c(d)\le c$ is imposed, and $\hat{c}(d)$ is a constraint estimator which usually needs to be studied on a case-by-case basis.

We first review important examples of policy learning that fit into our framework.

Vanilla direct policy search. The first example is what most existing work on policy learning has focused on, primarily for deterministic policies with a binary treatment. When the policy class is unrestricted, the optimal treatment assignment rule depends on the sign of the conditional average treatment effect for each individual unit, which cannot be extended to stochastic policies. Our proposed optimal incremental propensity score policies maximize the value function.

Fair policy learning. In many decision-making scenarios, such as hiring, recommendation systems,

and criminal justice, concerns have been raised regarding the fairness of decisions from the learning process (Chzhen et al., 2020). Let $S \in \mathcal{S}$ denote the sensitive attribute. For randomized predictions $f : \mathcal{X} \times \mathcal{S} \to \Delta(\mathcal{A})$, popular fairness criteria include (i) demographic parity (DP) (Calders et al., 2009): $E[f(X,S) \mid S = s] = E[f(X,S) \mid S = s'], \forall s, s' \in \mathcal{S}$, which says that $f(X,S)$ is independent from S, or (ii) equal opportunity (EO) (Hardt et al., 2016): $E[f(X,S) \mid S = s, A = a] = E[f(X,S) \mid S = s', A = a], \forall s, s' \in \mathcal{S}, a \in \mathcal{A}$, which requires equal true positive and true negative rates. Following the same spirit, we consider fair policy learning tasks as the constrained optimization problem:

$$\max_{d \in \mathcal{D}} V(d), \text{ subject to } m(d) \leq b,$$

where $m(d)$ is either the DP or EO metrics, which can be estimated by

$$\hat{m}_{\mathrm{DP}}(d) = \left(\sum_{s \in \mathcal{S}} \left(\frac{\sum_{i=1}^n d(X_i) I\{S_i = s\}}{\sum_{i=1}^n I\{S_i = s\}} - \frac{\sum_{i=1}^n d(X_i)}{n} \right)^2 \right)^{1/2},$$

or

$$\hat{m}_{\mathrm{EO}}(d) = \left(\sum_{s \in \mathcal{S}} \left(\frac{\sum_{i=1}^n d(X_i) I\{S_i = s, A_i = 1\}}{\sum_{i=1}^n I\{S_i = s, A_i = 1\}} - \frac{\sum_{i=1}^n d(X_i) I\{A_i = 1\}}{\sum_{i=1}^n I\{A_i = 1\}} \right)^2 \right)^{1/2},$$

and b is a pre-specified tuning parameter.

Resource-limited policy learning. In many real-world applications, the proportion of individuals who can receive the treatment is a priori limited due to a budget or a capacity constraint. So we consider the resource-limited policy learning tasks as the constrained optimization problem:

$$\max_{d \in \mathcal{D}} V(d), \text{ subject to } E[d] \leq b,$$

where b is the pre-specified budget or capacity.

Protect the vulnerable. Since the optimal policy is typically defined as the maximizer of the expected potential outcome over the entire population, such a policy may be suboptimal or even detrimental to certain disadvantaged subgroups. Fang et al. (2022) propose the fairness-oriented optimal policy learning framework:

$$\max_{d \in \mathcal{D}} V(d), \text{ subject to } Q_\tau(Y(d)) \geq b,$$

where $Q_\tau(Y(d)) = \inf\{t : F_{Y(d)}(t) \geq \tau\}$ is the τ-th quantile of $Y(d)$, $F_{Y(d)}$ denotes the cumulative distribution function of $Y(d)$, and b is a pre-specified protection threshold. Note that the quantile function can be estimated by $\hat{Q}_\tau(Y(d)) = \arg\min_q n^{-1} \sum_{i=1}^n c_i(d) \rho_\tau(Y_i - q)$, where $\rho_\tau(u) = u(\tau - I\{u < 0\})$ is the quantile loss function, and $c_i(d) = A_i d(X_i) + (1 - A_i)(1 - d(X_i))$.

Other examples in the literature include the counterfactual no-harm criterion by the principal stratification method (Li et al., 2023), (weakly) NP-hard knapsack problem (Luedtke and van der Laan, 2016a), and instrumental variable methods (Qiu et al., 2021).

5 ASYMPTOTIC ANALYSIS OF POLICY EVALUATION AND LEARNING

In this section, we first characterize the asymptotic distributions of our proposed one-step estimator (5) and the cross-fitted estimator (6) for off-policy evaluation.

Theorem 1. *Assume the following conditions hold:* (i) $\|\hat{\pi}(x) - \pi(x)\|_{L_2} \times \|\hat{\mu}_a - \mu_a\|_{L_2} = o_p(n^{-1/2})$ *for* $a = 0, 1$*;* (ii) $\phi(P)$ *belongs to a Donsker class;* (iii) $|Y|$ *and* $|\delta(X)|$ *are bounded in probability. For the one-step estimator, we have that* $\sqrt{n}(\hat{\Psi}_{\mathrm{OS}} - \Psi(P)) \to \mathcal{N}(0, E[\phi^2])$.

Theorem 2. *Assume the following conditions hold:* (i) $\|\hat{\pi}(x) - \pi(x)\|_{L_2} \times \|\hat{\mu}_a - \mu_a\|_{L_2} = o_p(n^{-1/2})$ *for* $a = 0, 1$*;* (ii) $|Y|$ *and* $|\delta(X)|$ *are bounded in probability. For the cross-fitting estimator, we have that* $\sqrt{n}(\hat{\Psi}_{\mathrm{CF}} - \Psi(P)) \to \mathcal{N}(0, E[\phi^2])$.

Condition (i) of Theorems 1 and 2 is commonly assumed such that the second-order remainder term is $o_p(1)$ (Kennedy, 2022). Condition (ii) of Theorems 1 ensures the centered empirical process term is $o_p(1)$. Condition (iii) of Theorems 1 and condition (ii) of Theorems 2 are mild regularity conditions. The asymptotic variance of the one-step estimator can be consistently estimated by $\frac{1}{n} \sum_{i=1}^n \hat{\phi}^2(\hat{P})(O_i)$, and the asymptotic variance of the cross-fitting estimator can be consistently estimated by $\frac{1}{K} \sum_{k=1}^K P_{n,k} \phi^2(\hat{P}_{-k})(O)$.

Next, we prove asymptotic guarantees for the following generic off-policy learning problem:

$$\max_{d \in \mathcal{D}} \hat{V}(d), \text{ subject to } \hat{c}(d) \leq c,$$

where $\hat{V}(d)$ is a value estimator of our proposed incremental propensity score policies, $\hat{c}(d)$ is an estimate of the constraint, and c is a pre-specified criterion.

Consider a parametric policy class $\mathcal{D}(H)$ indexed by $\eta \in H$, where H is a compact set. Let η^* denote the true Euclidean parameter indexing the optimal policy. To simplify the notation, for $d(x;\eta) \in \mathcal{D}(H)$, we define $V(\eta) = V(d(x;\eta))$ and $c(\eta) = c(d(x;\eta))$.

Theorem 3. *Assume the following conditions hold:* (i) $d(x;\eta)$ *is a continuously differentiable and convex function with respect to* η*;* (ii) $\hat{V}(\eta)$ *and* $\hat{c}(\eta)$ *converge to* $V(\eta)$ *and* $c(\eta)$ *at rates* $\sqrt{n}$*. We have that* (i) $V(\hat{\eta}) - V(\eta^*) = O_p(n^{-1/2})$*;* (ii) $\hat{V}(\hat{\eta}) - V(\eta^*) = O_p(n^{-1/2})$.

Theorem 4. *Assume the following conditions hold:* (i) $\mathcal{D}$ *is a Glivenko–Cantelli class;* (ii) $\hat{\pi}(x)$ *and* $\hat{\mu}_a(x)$ *are uniformly consistent estimators of* $\pi(x)$ *and* $\mu_a(x)$ *for* $a = 0, 1$*;* (iii) $\forall d \in \mathcal{D}$*,* $m \in (0, 1)$*, it follows that* $md \in \mathcal{D}$*. We have that* (i) $V(\hat{d}) - V(d) = o_p(1)$*;* (ii) $\hat{V}(\hat{d}) - V(d) = o_p(1)$.

Theorem 3 and 4 follow a well-known canvas in the stochastic off-policy learning literature (Shapiro, 1991; Li et al., 2023). Theorem 3 (i) establishes that the regret of the learned policy attains the convergence rate of $n^{-1/2}$, and (ii) shows that $\hat{V}(\hat{\eta})$ is a $\sqrt{n}$-consistent estimator of the optimal value function for parametric and convex policy classes under mild assumptions. Theorem 4 (i) establishes that the regret of the learned policy vanishes, and (ii) shows $\hat{V}(\hat{\eta})$ is still a consistent estimator for GC classes.

6 EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the performance of our proposed positivity-free policy learning methods by comparison with standard policy learning methods. Replication code is available at GitHub.

6.1 Simulation

We consider the fair policy learning task under the demographic parity constraint and simulate

$$\begin{aligned} &S \sim \text{Bernoulli}(0.5),\ (X_1, X_2, X_3) \sim \text{Uniform}(0, 1), \\ &A \sim \text{Bernoulli}(\text{expit}(-1 - X_1 + 1.5X_2 - 0.25X_3 - 3.1S)), \\ &Y(0) \sim \mathcal{N}\{20(1 + X_1 - X_2 + X_3^2 + \exp(X_2)), 20^2\}, \\ &Y(1) \sim \mathcal{N}\{20(1 + X_1 - X_2 + X_3^2 + \exp(X_2)) + \\ &\qquad 25(3 - 5X_1 + 2X_2 - 3X_3 + S), 20^2\}, \end{aligned}$$

where $\text{expit} : x \mapsto 1/(1 + \exp(-x))$. We let S denote the sensitive attribute and X_1, X_2, X_3 the common non-sensitive attributes. The treatment assignment mechanism yields variable propensity scores that can degrade the performance of weighting-based estimators in standard policy learning methods. Specifically, half propensity scores are smaller than 0.06. For standard methods, we consider the policy class of linear rules $\mathcal{D}_{\text{linear}} = \{d(s, x) = I\{(1, s, x_1, x_2, x_3)\beta > 0\} : \beta \in \mathbb{R}^5, \|\beta\|_2 = 1\}$. For the incremental propensity score policies, we consider the class $\mathcal{D}_{\text{IPS}} = \{d(s, x) = \delta(s, x; \beta)\pi(s, x)/\{\delta(s, x; \beta)\pi(s, x) + 1 - \pi(s, x)\} : \beta \in \mathbb{R}^5\}$, which is indexed by $\delta(s, x; \beta) = \exp\{(1, s, x_1, x_2, x_3)\beta\}$.

We estimate the outcome regression model $\mu(s, x)$ and the propensity score $\pi(s, x)$ using the generalized random forests (Athey et al., 2019) implemented in the `R` package `grf`. The constrained optimization problems are solved by the derivative-free linear approximations algorithm (Powell, 1994), implemented in the `R` package `nloptr`. The sample size is $n = 1000$, and the demographic parity threshold is $\tau = 0.01$. Note that the policy learning approach where the optimal policy is directly determined by the conditional average treatment effects cannot fit into our stochastic policy learning framework since it fails to satisfy the constraint such as fairness.

We compare the true values of the estimated optimal policies using test data with sample size $N = 10^5$. On the test data where the counterfactual outcomes are known, we use the same derivative-free linear approximations algorithm and solve the constrained optimization problem to approximate the true optimal value. Simulation results of 100 Monte Carlo repetition are reported in Figure 1a. When some estimated propensity scores are exactly 0, the IPW and AIPW estimators would fail, and `NA` is returned. The OR estimator can wrongly extrapolate. Three standard methods IPW, OR, and AIPW have the worst performance. The IPW-IPS estimator also has large variability, which is similarly reported in Kennedy (2019). The OR-IPS and efficient one-step estimators achieve the best performance with the highest value.

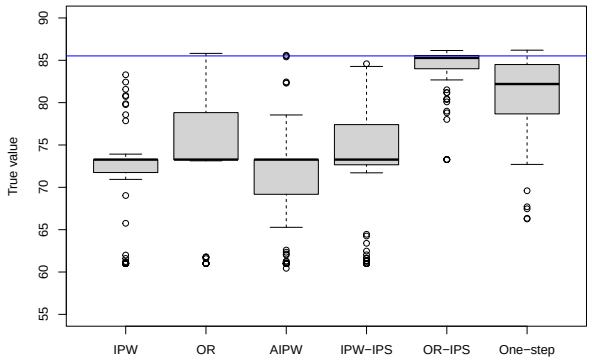
We also observe the presence of lower outliers for all the methods when subjected to positivity violation. This indicates a degree of instability in the derivative-free linear approximations algorithm employed in our stochastic policy learning tasks. Addressing this issue, we acknowledge the need for future research to develop more robust optimization algorithms tailored specifically to our proposed methods.

Additional simulation results are given in Section G of the Supplementary Material. Specifically, we illustrate that our proposed policy learning methods have comparable performance when there is no positivity violation, and also illustrate the better performance of our proposed methods when using parametric models.

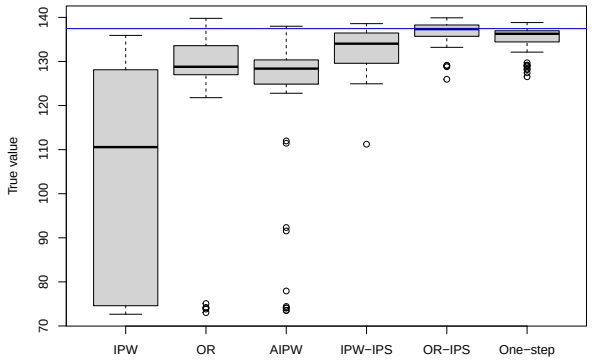
6.2 Data Application

We illustrate our proposed methods using semi-synthetic data from the `Fairlearn` open source project (Weerts et al., 2023). Additional information on our data analysis is provided in Section H of the Supplementary Material.

The Diabetes dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks (Strack et al., 2014), and contains hospital records of patients diagnosed with diabetes who underwent laboratory tests and medications and stayed up to 14 days. Our application aims to learn the optimal policy for prescribing diabetic medication



(a) Simulations.



(b) Diabetes data application.

Figure 1: Performance of optimal policies under three standard methods (IPW, OR, AIPW) and our proposed three methods (IPW-IPS, OR-IPS, One-step). The blue line is the (approximate) true optimal value.

by maximizing the expected outcome under the demographic parity constraint. The sensitive attribute is race, and we assess the positivity violation from the observation that many of the estimated propensity scores are very close to 0.

We include 7 baseline covariates: `race`, `gender`, `age`, `time_in_hospital` (number of days between admission and discharge), `num_lab_procedures` (number of lab tests performed during the encounter), `num_medications` (number of distinct generic names administered during the encounter) and `number_diagnoses` (number of diagnoses). Under positivity violation, we are unable to identify the value function, e.g. relying on the outcome regression's extrapolation to learn the counterfactual outcomes on test data. Thus the potential outcomes are simulated as follows: $Y(0) \sim \mathcal{N}\{20(1 + \texttt{gender} - \texttt{age} + \texttt{time\_in\_hospital} + \texttt{num\_lab\_procedures} + \texttt{num\_medications} + \texttt{num\_medications}^2 + \exp(\texttt{number\_diagnoses})), 20^2\}$, and $Y(1) \sim \mathcal{N}\{20(1 + \texttt{gender} - \texttt{age} + \texttt{time\_in\_hospital} + \texttt{num\_lab\_procedures} + \texttt{num\_medications} + \texttt{num\_medications}^2 + \exp(\texttt{number\_diagnoses})) + 25(3 - 5\texttt{age} + 2\texttt{time\_in\_hospital} - 3\texttt{num\_medications} + \texttt{race}), 20^2\}$. The estimation setup and policy classes are the same as previous simulations. The constrained optimization problems are solved by the derivative-free linear approximations algorithm, implemented in the `R` package `nloptr`. We run 50 repetitions; each time we randomly select 500 patients as training data to learn the optimal policy and 2000 patients as test data to evaluate the performance. Empirical results are reported in Figure 1b. When the positivity violation is severer, the IPW estimator has extremely large variability, and we also observe that our proposed methods perform consistently better than the standard methods.

7 DISCUSSION

This article proposes a general positivity-free stochastic policy learning framework using observational data, possibly subject to application-specific constraints. There are several interesting directions for future research. It is relevant to extend our methods to the more general case with multiple time points for treatment assignment, multiple treatment levels, or high-dimensional models (Wei et al., 2023; Sarvet et al., 2023), where positivity is even more likely to be violated. The incremental propensity score approach can also be extended to account for common issues such as covariate shift (Zhao et al., 2023; Lei et al., 2023), censoring and dropout (Cui et al., 2023), and truncation by death (Chu et al., 2023).

Acknowledgements

Pan Zhao and Julie Josse are supported in part by the French National Research Agency ANR-16-IDEX-0006. Shu Yang is partially supported by NSF SES 2242776, NIH 1R01AG066883 and 1R01ES031651.

References

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

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, 2019. doi: 10.1214/18-AOS1709. URL https://doi.org/10.1214/18-AOS1709.

Aurélien Bibaut, Nathan Kallus, Maria Di-

makopoulou, Antoine Chambaz, and Mark van Der Laan. Risk minimization from adaptively collected data: Guarantees for supervised and policy learning. *Advances in neural information processing systems*, 34:19261–19273, 2021.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

Zach Branson, Edward H Kennedy, Sivaraman Balakrishnan, and Larry Wasserman. Causal effect estimation after propensity score trimming with continuous treatments. *arXiv preprint arXiv:2309.00706*, 2023.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pages 13–18. IEEE, 2009.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221.

Jianing Chu, Shu Yang, and Wenbin Lu. Multiply robust off-policy evaluation and learning under truncation by death. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6195–6227. PMLR, 23–29 Jul 2023.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7321–7331. Curran Associates, Inc., 2020.

Mayleen Cortez, Matthew Eichhorn, and Christina Yu. Staggered rollout designs enable causal inference under interference without network knowledge. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 7437–7449. Curran Associates, Inc., 2022.

Yifan Cui, Ruoqing Zhu, and Michael Kosorok. Tree based weighted learning for estimating individualized treatment rules with censored data. *Electronic journal of statistics*, 11(2):3927, 2017.

Yifan Cui, Michael R Kosorok, Erik Sverdrup, Stefan Wager, and Ruoqing Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85 (2):179–211, 2023.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29:485–511, 2014.

Ethan X Fang, Zhaoran Wang, and Lan Wang. Fairness-oriented learning for optimal individualized treatment rules. *Journal of the American Statistical Association*, pages 1–14, 2022.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3): 879–908, 2023.

Lin Gui and Victor Veitch. Causal estimation for text data with (apparent) overlap violations. In *International Conference on Learning Representations*, 2023.

Sebastian Haneuse and Andrea Rotnitzky. Estimation of the effect of interventions that modify the received treatment. *Statistics in medicine*, 32(30):5260–5277, 2013.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

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

Zeyang Jia, Eli Ben-Michael, and Kosuke Imai. Bayesian safe policy learning with chance constrained optimization: Application to military security assessment during the vietnam war. *arXiv preprint arXiv:2307.08840*, 2023.

Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. Policy learning "without" overlap: Pessimism and generalized empirical bernstein's inequality. *arXiv preprint arXiv:2212.09900*, 2022.

Nathan Kallus and Miruna Oprescu. Robust and agnostic learning of conditional distributional treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 6037–6060. PMLR, 2023.

Nathan Kallus and Masatoshi Uehara. Efficient evaluation of natural stochastic policies in offline reinforcement learning. *arXiv preprint arXiv:2006.03886*, 2020.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. *Advances in neural information processing systems*, 31, 2018.

Edward H Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114 (526):645–656, 2019.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

Edward H. Kennedy, Sivaraman Balakrishnan, and Max G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics*, 48(4):2008 – 2030, 2020.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: partial identification through smoothness. *arXiv preprint arXiv:2305.11812*, 2023.

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

Carolin Lawrence, Artem Sokolov, and Stefan Riezler. Counterfactual learning from bandit feedback under deterministic logging: A case study in statistical machine translation. *arXiv preprint arXiv:1707.09118*, 2017.

Lihua Lei, Roshni Sahoo, and Stefan Wager. Policy learning under biased sample selection. *arXiv preprint arXiv:2304.11735*, 2023.

Haoxuan Li, Chunyuan Zheng, Yixiao Cao, Zhi Geng, Yue Liu, and Peng Wu. Trustworthy policy learning under the counterfactual no-harm criterion. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 20575–20598. PMLR, 23–29 Jul 2023.

Yi Liu, Huiyue Li, Yunji Zhou, and Roland Matsouaka. Average treatment effect on the treated, under lack of positivity. *arXiv preprint arXiv:2309.01334*, 2023.

Alexander R Luedtke and Mark J van der Laan. Optimal individualized treatments in resource-limited settings. *The international journal of biostatistics*, 12(1):283–303, 2016a.

Alexander R Luedtke and Mark J van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713, 2016b.

Iván Díaz Muñoz and Mark van Der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

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

Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96 (456):1410–1423, 2001.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51, 1923.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108 (2):299–319, 2021.

Michael JD Powell. *A direct search optimization method that models the objective and constraint functions by linear interpolation*. Springer, 1994.

Hongxiang Qiu, Marco Carone, Ekaterina Sadikova, Maria Petukhova, Ronald C Kessler, and Alex Luedtke. Optimal individualized decision rules using instrumental variable methods. *Journal of the American Statistical Association*, 116(533):174–191, 2021.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Aaron L. Sarvet, Kerollos N. Wanis, Jessica G. Young, Roberto Hernandez-Alejandro, and Mats J. Stensrud. Longitudinal Incremental Propensity Score Interventions for Limited Resource Settings. *Biometrics*, 79(4):3418–3430, 03 2023. ISSN 0006-341X. doi: 10.1111/biom.13859. URL https://doi.org/10.1111/biom.13859.

Jasjeet S Sekhon and Walter R Mebane. Genetic optimization using derivatives. *Political Analysis*, 7: 187–210, 1998.

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

Chengchun Shi, Ailin Fan, Rui Song, and Wenbin Lu. High-dimensional *A*-learning for optimal dynamic treatment regimes. *The Annals of Statistics*, 46(3): 925 – 957, 2018.

Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, John N Clore, et al. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

Jin Tian. Identifying dynamic sequential plans. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2008.

Anastasios A Tsiatis, Marie Davidian, Shannon T Holloway, and Eric B Laber. *Dynamic treatment regimes: Statistical methods for precision medicine*. CRC press, 2019.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

Mark J van der Laan and Maya L Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *The international journal of biostatistics*, 3(1), 2007.

Mark J van der Laan and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003.

Mark J van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

Giancarlo Visconti and José R Zubizarreta. Handling limited overlap in observational studies with cardinality matching. *Observational Studies*, 4(1):217–249, 2018.

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

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems. *Journal of Machine Learning Research*, 24 (257):1–8, 2023.

Waverly Wei, Yuqing Zhou, Zeyu Zheng, and Jingshen Wang. Inference on the best policies with many covariates. *Journal of Econometrics*, page 105460, 2023. ISSN 0304-4076.

Shu Yang and Peng Ding. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105 (2):487–493, 03 2018. URL https://doi.org/10.1093/biomet/asy008.

Jessica G Young, Miguel A Hernán, and James M Robins. Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiologic methods*, 3(1):1–19, 2014.

Christina Lee Yu, Edoardo M Airoldi, Christian Borgs, and Jennifer T Chayes. Estimating the total treatment effect in randomized experiments with unknown network structure. *Proceedings of the National Academy of Sciences*, 119(44):e2208975119, 2022.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

Yuqian Zhang, Abhishek Chakrabortty, and Jelena Bradic. Semi-supervised causal inference: Generalizable and double robust inference for average treatment effects under selection bias with decaying overlap. *arXiv preprint arXiv:2305.12789*, 2023.

Pan Zhao and Yifan Cui. A semiparametric instrumented difference-in-differences approach to policy learning. *arXiv preprint arXiv:2310.09545*, 2023.

Pan Zhao, Julie Josse, and Shu Yang. Efficient and robust transfer learning of optimal individualized treatment regimes with right-censored survival data. *arXiv preprint arXiv:2301.05491*, 2023.

Wenjing Zheng and Mark J. van der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. Working Paper Series Working Paper 273, U.C. Berkeley Division of Biostatistics, November 2010. URL https://biostats.bepress.com/ucbbiostat/paper273.

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

A Proof of Proposition 1

The proof of our identification results is straightforward, following similar arguments in Kennedy (2019). First, we prove the OR-IPS formula:

$$\begin{aligned}
V(d) &= E[Y(d)] \\
&= E[Y(1)d(X) + Y(0)(1-d(X))] \\
&= E[E[Y(1)d(X) + Y(0)(1-d(X)) \mid X]] \\
&= E[d(X)E[Y(1) \mid X] + (1-d(X))E[Y(0) \mid X]] \\
&= E[d(X)E[Y \mid X, A=1] + (1-d(X))E[Y \mid X, A=0]] \\
&= E\left[\frac{\delta(X)\pi(X)}{\delta(X)\pi(X)+1-\pi(X)}\mu_1(X) + \frac{1-\pi(X)}{\delta(X)\pi(X)+1-\pi(X)}\mu_0(X)\right] \\
&= E\left[\frac{\delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X)+1-\pi(X)}\right].
\end{aligned}$$

Next, we prove the IPW-IPS formula:

$$\begin{aligned}
&E\left[\frac{Y\{\delta(X)A+1-A\}}{\delta(X)\pi(X)+1-\pi(X)}\right] \\
&= E\left[\frac{YA\delta(X)}{\delta(X)\pi(X)+1-\pi(X)} + \frac{Y(1-A)}{\delta(X)\pi(X)+1-\pi(X)}\right] \\
&= E\left[\frac{Y(1)A\delta(X)}{\delta(X)\pi(X)+1-\pi(X)} + \frac{Y(0)(1-A)}{\delta(X)\pi(X)+1-\pi(X)}\right] \\
&= E\left[E\left[\frac{Y(1)A\delta(X)}{\delta(X)\pi(X)+1-\pi(X)} + \frac{Y(0)(1-A)}{\delta(X)\pi(X)+1-\pi(X)} \mid X\right]\right] \\
&= E\left[\frac{E[Y(1)A \mid X]\delta(X)}{\delta(X)\pi(X)+1-\pi(X)} + \frac{E[Y(0)(1-A) \mid X]}{\delta(X)\pi(X)+1-\pi(X)}\right] \\
&= E\left[Y(1)\frac{E[A \mid X]\delta(X)}{\delta(X)\pi(X)+1-\pi(X)} + Y(0)\frac{E[(1-A) \mid X]}{\delta(X)\pi(X)+1-\pi(X)}\right] \\
&= E[E[Y(1)d(X) + Y(0)(1-d(X)) \mid X]] \\
&= V(d).
\end{aligned}$$

B Proof of Proposition 2

We derive the efficient influence function for the following statistical functional:

$$\Psi(P) = E_P\left[\frac{\delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X)+1-\pi(X)}\right].$$

For a given distribution P in the nonparametric statistical model $\mathcal{M}$, we let p denote the density of P with respect to some dominating measure ν. For all bounded $h \in L_2(P)$, define the parametric submodel $p_\epsilon = (1+\epsilon h)p$, which is valid for small enough ϵ and has score h at $\epsilon = 0$. We would establish that $\Psi(P)$ is pathwise differentiable with respect to $\mathcal{M}$ at P with efficient influence function $\phi(P)$ if we have that for any $P \in \mathcal{M}$,

$$\frac{\partial}{\partial\epsilon}\Psi(P_\epsilon)\bigg|_{\epsilon=0} = \int \phi(P)(o)h(o)dP(o).$$

We denote $\pi_\epsilon(x) = E_{P_\epsilon}[A \mid X = x]$, $\mu_{a,\epsilon}(x) = E_{P_\epsilon}[Y \mid X = x, A = a]$, $S = \partial \log p_\epsilon / \partial \epsilon$, and can compute

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\Psi(P_\epsilon)\Big|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon} E_{P_\epsilon}\left[\frac{\delta(X)\pi_\epsilon(X)\mu_{1,\epsilon}(X) + \{1-\pi_\epsilon(X)\}\mu_{0,\epsilon}(X)}{\delta(X)\pi_\epsilon(X) + 1 - \pi_\epsilon(X)}\right]\Big|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon} E_P\left[(1+\epsilon S)\frac{\delta(X)\pi_\epsilon(X)\mu_{1,\epsilon}(X) + \{1-\pi_\epsilon(X)\}\mu_{0,\epsilon}(X)}{\delta(X)\pi_\epsilon(X) + 1 - \pi_\epsilon(X)}\right]\Big|_{\epsilon=0} \\
&= E_P\left[S\frac{\delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)}\right] \\
&\quad + E_P\left[\frac{1}{\delta(X)\pi(X) + 1 - \pi(X)}\left(\pi(X)\frac{\partial}{\partial \epsilon}\mu_{1,\epsilon}(X)\Big|_{\epsilon=0} + \mu_1(X)\frac{\partial}{\partial \epsilon}\pi_\epsilon(X)\Big|_{\epsilon=0}\right)\right] \\
&\quad + E_P\left[\frac{1}{\delta(X)\pi(X) + 1 - \pi(X)}\left(\{1-\pi(X)\}\frac{\partial}{\partial \epsilon}\mu_{0,\epsilon}(X)\Big|_{\epsilon=0} - \mu_0(X)\frac{\partial}{\partial \epsilon}\pi_\epsilon(X)\Big|_{\epsilon=0}\right)\right] \\
&\quad - E_P\left[\frac{\delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\{\delta(X)\pi(X) + 1 - \pi(X)\}^2}\left(\delta(X)\frac{\partial}{\partial \epsilon}\pi_\epsilon(X)\Big|_{\epsilon=0} - \frac{\partial}{\partial \epsilon}\pi_\epsilon(X)\Big|_{\epsilon=0}\right)\right].
\end{aligned}$$

Then we need to compute

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\pi_\epsilon(X)\Big|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon}\frac{\pi(X) + \epsilon E_P[SA \mid X]}{1 + \epsilon E_P[S \mid X]}\Big|_{\epsilon=0} \\
&= E_P[SA \mid X] - \pi(X)E_P[S \mid X] \\
&= E_P[S(A - \pi(X)) \mid X],
\end{aligned}$$

and for $a = 0, 1$,

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\mu_{a,\epsilon}(X)\Big|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon}\frac{\mu_a(X) + \epsilon E_P[SY \mid X, A=a]}{1 + \epsilon E_P[S \mid X, A=a]}\Big|_{\epsilon=0} \\
&= E_P[SY \mid X, A=a] - \mu_a(X)E_P[S \mid X, A=a] \\
&= E_P[S(Y - \mu_a(X)) \mid X, A=a].
\end{aligned}$$

Combining the above derivations, we obtain that

$$\begin{aligned}
\phi(P)(O) = &\frac{A\delta(X)\{Y - \mu_1(X)\} + (1-A)\{Y - \mu_0(X)\} + \delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)} \\
&+ \frac{\delta(X)\tau(X)\{A - \pi(X)\}}{\{\delta(X)\pi(X) + 1 - \pi(X)\}^2} - \Psi(P),
\end{aligned}$$

which yields the result.

C Proof of Theorem 1

We first outline the inferential strategy from semiparametric theory. Consider a statistical model $\mathcal{M}$ for distributions $\tilde{P}$, with P denoting the true distribution. Under sufficient smoothness conditions, we have the following von Mises expansion for $\Psi(\tilde{P})$:

$$\Psi(\tilde{P}) = \Psi(P) - \int \phi(\tilde{P})(o)dP(o) + \mathrm{Rem}(\tilde{P}, P),$$

where $\phi(P)$ is the influence function derived in Section B such that $\int \phi(P)(o)dP(o) = 0$, and $\mathrm{Rem}(\tilde{P}, P) = O(\|\tilde{P} - P\|^2)$ is a second-order reminder term that we will analyze later.

Let $\hat{P}$ be an estimator of P, then we obtain the following one-step estimator of $\Psi(P)$:

$$\hat{\Psi} = \Psi(\hat{P}) + \int \phi(\hat{P})(o)dP_n(o),$$

where P_n is the empirical distribution.

Next, we characterize the asymptotic properties of $\hat{\Psi}$. Note that

$$\begin{aligned}
\hat{\Psi} - \Psi(P) &= \left\{\Psi(\hat{P}) + \int \phi(\hat{P})(o) dP_n(o)\right\} - \Psi(P) \\
&= \left\{\Psi(\hat{P}) - \Psi(P)\right\} + \int \phi(\hat{P})(o) dP_n(o) \\
&= -\int \phi(\hat{P})(o) dP(o) + \text{Rem}(\hat{P}, P) + \int \phi(\hat{P})(o) dP_n(o) \\
&= \int \phi(\hat{P})(o) d\left\{P_n(o) - P(o)\right\} + \text{Rem}(\hat{P}, P) \\
&= \int \phi(P)(o) dP_n(o) + \int \left\{\phi(\hat{P})(o) - \phi(P)(o)\right\} d\left\{P_n(o) - P(o)\right\} + \text{Rem}(\hat{P}, P).
\end{aligned}$$

Therefore, $\sqrt{n}\left\{\hat{\Psi} - \Psi(P)\right\}$ is expressed as the following three terms:

$$\begin{aligned}
\sqrt{n}\left\{\hat{\Psi} - \Psi(P)\right\} = \sqrt{n}\int \phi(P)(o) dP_n(o) \\
&+ \sqrt{n}\int \left\{\phi(\hat{P})(o) - \phi(P)(o)\right\} d\left\{P_n(o) - P(o)\right\} \\
&+ \sqrt{n}\text{Rem}(\hat{P}, P).
\end{aligned}$$

By the central limit theorem, $\sqrt{n}\int \phi(P)(o) dP_n(o)$ is asymptotically normal with the asymptotic variance given by $E[\phi^2(P)(O)]$.

We assume that $\phi(P)$ belongs to a Donsker class, so we have that the centered empirical process

$$\sqrt{n}\int \left\{\phi(\hat{P})(o) - \phi(P)(o)\right\} d\left\{P_n(o) - P(o)\right\} = o_p(1).$$

Finally, we characterize the second-order remainder term:

$$\text{Rem}(\hat{P}, P) = \Psi(\hat{P}) - \Psi(P) + E_P[\phi(\hat{P})(O)].$$

We have that

$$\Psi(P) = E_P\left[\frac{\delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)}\right],$$

and

$$\begin{aligned}
&E_P[\phi(\hat{P})(O)] \\
&= E_P\left[\frac{A\delta(X)\{Y - \hat{\mu}_1(X)\} + (1-A)\{Y - \hat{\mu}_0(X)\} + \delta(X)\hat{\pi}(X)\hat{\mu}_1(X) + \{1-\hat{\pi}(X)\}\hat{\mu}_0(X)}{\delta(X)\hat{\pi}(X) + 1 - \hat{\pi}(X)}\right. \\
&\quad \left. + \frac{\delta(X)\hat{\tau}(X)\{A - \hat{\pi}(X)\}}{\{\delta(X)\hat{\pi}(X) + 1 - \hat{\pi}(X)\}^2}\right] - \Psi(\hat{P}).
\end{aligned}$$

Combining the derivations above, we have that

$$\begin{aligned}
\left|\text{Rem}(\hat{P}, P)\right| \leq \hat{C}_1 \|\hat{\mu}_1(X) - \mu_1(X)\|_{L_2} \times \|\hat{\pi}(X) - \pi(X)\|_{L_2} \\
&+ \hat{C}_2 \|\hat{\mu}_0(X) - \mu_0(X)\|_{L_2} \times \|\hat{\pi}(X) - \pi(X)\|_{L_2} \\
&+ \hat{C}_3 \|\hat{\pi}(X) - \pi(X)\|_{L_2}^2,
\end{aligned}$$

where $\hat{C}_1$, $\hat{C}_2$ and $\hat{C}_3$ are $O_p(1)$. We assume that $\|\hat{\pi}(x) - \pi(x)\|_{L_2} = o_p(n^{-1/4})$, and $\|\hat{\mu}_a - \mu_a\|_{L_2} = o_p(n^{-1/4})$ for $a = 0, 1$. Therefore, we have that $\sqrt{n}\text{Rem}(\hat{P}, P) = o_p(1)$. That is, we conclude that

$$\sqrt{n}\left\{\hat{\Psi} - \Psi(P)\right\} \to \mathcal{N}(0, E[\phi^2(P)(O)]),$$

which completes the proof.

D Proof of Theorem 2

Essentially, we need to prove that the centered empirical process is $o_p(1)$, when we avoid Donsker conditions by using the cross-fitting technique. We first review a useful lemma from Kennedy et al. (2020).

Lemma 1. *Consider two independent samples $\mathcal{O}_1 = (O_1, \ldots, O_n)$ and $\mathcal{O}_2 = (O_{n+1}, \ldots, O_N)$ drawn from the distribution $\mathbb{P}$. Let $\hat{f}(o)$ be a function estimated from $\mathcal{O}_2$, and $\mathbb{P}_n$ the empirical measure over $\mathcal{O}_1$, then we have*

$$(\mathbb{P}_n - \mathbb{P})(\hat{f} - f) = O_{\mathbb{P}}\left(\frac{\|\hat{f} - f\|}{\sqrt{n}}\right).$$

Proof. First note that by conditioning on $\mathcal{O}_2$, we obtain that

$$\mathbb{E}\left\{\mathbb{P}_n(\hat{f} - f) \,|\, \mathcal{O}_2\right\} = \mathbb{E}(\hat{f} - f \,|\, \mathcal{O}_2) = \mathbb{P}(\hat{f} - f),$$

and the conditional variance is

$$var\{(\mathbb{P}_n - \mathbb{P})(\hat{f} - f) \,|\, \mathcal{O}_2\} = var\{\mathbb{P}_n(\hat{f} - f) \,|\, \mathcal{O}_2\} = \frac{1}{n} var(\hat{f} - f \,|\, \mathcal{O}_2) \leq \|\hat{f} - f\|^2/n,$$

therefore by the Chebyshev's inequality we have that

$$\mathbb{P}\left\{\frac{|(\mathbb{P}_n - \mathbb{P})(\hat{f} - f)|}{\|\hat{f} - f\|^2/n} \geq t\right\} = \mathbb{E}\left[\mathbb{P}\left\{\frac{|(\mathbb{P}_n - \mathbb{P})(\hat{f} - f)|}{\|\hat{f} - f\|^2/n} \geq t \,\middle|\, \mathcal{O}_2\right\}\right] \leq \frac{1}{t^2},$$

thus for any $\epsilon > 0$ we can pick $t = 1/\sqrt{\epsilon}$ so that the probability above is no more than ϵ, which yields the result. □

Next, we characterize the asymptotic properties of the cross-fitted estimator $\hat{\Psi}_{\text{CF}}$. Following similar steps as Section C, we have that

$$\sqrt{n}\left\{\hat{\Psi}_{\text{CF}} - \Psi(P)\right\} = \sqrt{n}\int \phi(P)(o) dP_n(o) + \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sqrt{n_k}(R_{k,1} + R_{k,2}),$$

where $R_{k,1} = \int\left\{\phi(\hat{P}_{-k})(o) - \phi(P)(o)\right\} d\left\{P_{n,k}(o) - P(o)\right\}$, $R_{k,2} = \text{Rem}(\hat{P}_{-k}, P)$.

We note that

$$\begin{aligned} R_{k,1} &= \int\left\{\phi(\hat{P}_{-k})(o) - \phi(P)(o)\right\} d\left\{P_{n,k}(o) - P(o)\right\} \\ &= \int\left\{\xi(\hat{P}_{-k})(o) - \xi(P)(o)\right\} d\left\{P_{n,k}(o) - P(o)\right\}, \end{aligned}$$

where $\xi(P)(o) = \phi(P)(o) + \Psi(P)$, and by Lemma 1, we have that

$$\sqrt{n_k} R_{k,1} = O_p\left(\|\xi(\hat{P}_{-k}) - \xi(P)\|_{L_2}\right).$$

Note that

$$\begin{aligned} &\xi(\hat{P}_{-k})(O) - \xi(P)(O) \\ &= \frac{A\delta(X)\{Y - \mu_1(X)\} + (1-A)\{Y - \mu_0(X)\}}{\delta(X)\pi(X) + 1 - \pi(X)} - \frac{A\delta(X)\{Y - \mu_1(X)\} + (1-A)\{Y - \mu_0(X)\}}{\delta(X)\pi(X) + 1 - \pi(X)} \\ &\quad + \frac{\delta(X)\pi(X)\mu_1(X) + \{1 - \pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)} - \frac{\delta(X)\pi(X)\mu_1(X) + \{1 - \pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)} \\ &\quad + \frac{\delta(X)\tau(X)\{A - \pi(X)\}}{\{\delta(X)\pi(X) + 1 - \pi(X)\}^2} - \frac{\delta(X)\tau(X)\{A - \pi(X)\}}{\{\delta(X)\pi(X) + 1 - \pi(X)\}^2}, \end{aligned}$$

and we assume that $|Y|$ and $|\delta(X)|$ are bounded in probability. By the triangle and Cauchy-Schwarz inequalities, we have that

$$\|\xi(\hat{P}_{-k}) - \xi(P)\|_{L_2} \leq \hat{C}_{1,-k}\|\hat{\mu}_{0,-k}(X) - \mu_0(X)\|_{L_2} + \hat{C}_{2,-k}\|\hat{\mu}_{1,-k}(X) - \mu_1(X)\|_{L_2} + \hat{C}_{3,-k}\|\hat{\pi}_{-k}(X) - \pi(X)\|_{L_2}$$

where $\hat{C}_{1,-k}$, $\hat{C}_{2,-k}$ and $\hat{C}_{3,-k}$ are $O_p(1)$. We assume that $\|\hat{\pi}(x) - \pi(x)\|_{L_2} = o_p(n^{-1/4})$, and $\|\hat{\mu}_a - \mu_a\|_{L_2} = o_p(n^{-1/4})$ for $a = 0, 1$. Therefore, we have that $\sqrt{n_k}R_{k,1} = o_p(1)$.

By the same arguments as Section C, we have that $\sqrt{n_k}R_{k,2} = o_p(1)$. That is, we conclude that

$$\sqrt{n}\left\{\hat{\Psi}_{\mathrm{CF}} - \Psi(P)\right\} \to \mathcal{N}(0, E[\phi^2(P)(O)]),$$

which completes the proof.

E Proof of Theorem 3

In this section, we consider a parametric policy class $\mathcal{D}(H)$ indexed by $\eta \in H$. That is, the off-policy learning task is given by the following optimization problem:

$$\begin{aligned} \eta^* = \arg\max_{\eta \in H} \quad & V(\eta), \\ \text{subject to} \quad & c(\eta) \leq 0, \end{aligned}$$

and the estimated policy is given by

$$\begin{aligned} \hat{\eta} = \arg\max_{\eta \in H} \quad & \hat{V}(\eta), \\ \text{subject to} \quad & \hat{c}(\eta) \leq 0. \end{aligned}$$

We first review a useful lemma from Shapiro (1991).

Lemma 2. *Let H be a compact subset of $\mathbb{R}^k$. Let $C(H)$ denote the set of continuous real-valued functions on H, with $\mathcal{L} = C(H) \times \cdots \times C(H)$ the r-dimensional Cartesian product. Let $f(\eta) = (f_0, \ldots, f_r) \in \mathcal{L}$ be a vector of convex functions. Consider the quantity η^* defined as the solution to the following convex optimization program:*

$$\eta^* = \arg\min_{\eta \in H} f_0(\eta),$$
$$\text{subject to } f_j(\eta) \leq 0, j = 1, \ldots, r.$$

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

$$\hat{\eta}_n = \arg\min_{\eta \in H} \hat{f}_{n,0}(\eta),$$
$$\text{subject to } \hat{f}_{n,j}(\eta) \leq 0, j = 1, \ldots, r,$$

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

$$r(n)\left(\hat{f}_{n,0}(\eta)(\hat{\eta}_n) - f_0(\eta^*)\right) \to L,$$

for a particular random variable L. It follows that $\hat{f}_{n,0}(\eta)(\hat{\eta}_n) - f_0(\eta^*) = O_p(1/r(n))$.

By Theorem 1 or 2, we have that

$$\sqrt{n}\left(\hat{V}(\eta) - V(\eta)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_V(O_i; \eta) + o_p(1),$$

and by condition (ii), we have that

$$\sqrt{n}\left(\hat{c}(\eta) - c(\eta)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_c(O_i; \eta) + o_p(1),$$

where ϕ_V and ϕ_c are the influence functions.

By condition (i) and Lemma 2 with $r(n) = \sqrt{n}$, we obtain the conclusion (ii).

To prove conclusion (i), note that

$$V(\hat{\eta}) - V(\eta^*) = V(\hat{\eta}) - \hat{V}(\hat{\eta}) + \hat{V}(\hat{\eta}) - V(\eta^*),$$

where we have that $V(\hat{\eta}) - \hat{V}(\hat{\eta}) = O_p(n^{-1/2})$, and $\hat{V}(\hat{\eta}) - V(\eta^*) = O_p(n^{-1/2})$. Hence, we conclude that $V(\hat{\eta}) - V(\eta^*) = O_p(n^{-1/2})$, which completes the proof.

F Proof of Theorem 4

In this section, we follow similar techniques in Li et al. (2023) and consider the off-policy learning task given by the following optimization problem:

$$\begin{aligned} d^* &= \arg\max_{d \in \mathcal{D}} V(d) = \arg\max_{d \in \mathcal{D}} E[\xi(P)(O)], \\ &\text{subject to } c(d) = E[\phi_c(P)(O)] \leq 0, \end{aligned}$$

where $\mathcal{D}$ is a Glivenko–Cantelli class, and the estimated optimal policy is given by

$$\begin{aligned} \hat{d} &= \arg\max_{d \in \mathcal{D}} \hat{V}(d) = \arg\max_{d \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \xi(\hat{P})(O_i) \\ &\text{subject to } \hat{c}(d) = \frac{1}{n} \sum_{i=1}^{n} \phi_c(\hat{P})(O_i) \leq 0. \end{aligned}$$

By condition (iii) of Theorems 1 or condition (ii) of Theorems 2, we have that both $\{\xi(O; d) : d \in \mathcal{D}\}$ and $\{\phi_c(O; d) : d \in \mathcal{D}\}$ are GC classes.

To simplify the notation, let we denote $\mathcal{D}_c = \{d \in \mathcal{D} : c(d) \leq 0\}$, and $\mathcal{D}_{n,c} = \{d \in \mathcal{D} : \hat{c}(d) \leq 0\}$. First we note that the estimation error can be expressed as

$$V(d^*) - \hat{V}(\hat{d}) = V_n^{(1)} + V_n^{(2)} + V_n^{(3)},$$

where we define

$$\begin{aligned} V_n^{(1)} &= \max_{d \in \mathcal{D}_c} E[\xi(P)(O)] - \max_{d \in \mathcal{D}_c} P_n \xi(P)(O), \\ V_n^{(2)} &= \max_{d \in \mathcal{D}_c} P_n \xi(P)(O) - \max_{d \in \mathcal{D}_c} P_n \xi(\hat{P})(O), \\ V_n^{(3)} &= \max_{d \in \mathcal{D}_c} P_n \xi(\hat{P})(O) - \max_{d \in \mathcal{D}_{n,c}} P_n \xi(\hat{P})(O). \end{aligned}$$

We analyze the three terms as follows. We have that

$$\begin{aligned} V_n^{(1)} &= \max_{d \in \mathcal{D}_c} E[\xi(P)(O)] - \max_{d \in \mathcal{D}_c} P_n \xi(P)(O) \\ &\leq \max_{d \in \mathcal{D}_c} |E[\xi(P)(O)] - P_n \xi(P)(O)| \\ &= o_p(1), \end{aligned}$$

and similarly we have that

$$\begin{aligned} V_n^{(2)} &= \max_{d\in\mathcal{D}_c} P_n\xi(P)(O) - \max_{d\in\mathcal{D}_c} P_n\xi(\hat{P})(O) \\ &\leq \max_{d\in\mathcal{D}_c} \left| P_n\{\xi(P)(O) - \xi(\hat{P})(O)\} \right| \\ &= o_p(1). \end{aligned}$$

To analyze $V_n^{(3)}$, note that for any $d \in \mathcal{D}$, we have that

$$\begin{aligned} &E[\phi_c(P)(O)] - P_n\phi_c(\hat{P})(O) \\ &= \{E[\phi_c(P)(O)] - P_n\phi_c(P)(O)\} + \{P_n\phi_c(P)(O) - P_n\phi_c(\hat{P})(O)\}, \end{aligned}$$

and $E[\phi_c(P)(O)] - P_n\phi_c(P)(O)$ converges to 0 uniformly as $\{\phi_c(O;d) : d \in \mathcal{D}\}$ is a GC class, and $P_n\phi_c(P)(O) - P_n\phi_c(\hat{P})(O)$ converges to 0 uniformly by condition (ii).

Hence, $\forall\epsilon > 0$, $\exists N_1 \in \mathbb{N}$, such that for all $n > N_1$, $|E[\phi_c(P)(O)] - P_n\phi_c(\hat{P})(O)| < \epsilon$, by which we obtain that, for all $d \in \mathcal{D}_c$, i.e., $E[\phi_c(P)(O)] \leq c$, we have that $P_n\phi_c(\hat{P})(O) < c + \epsilon$. Therefore, we have that $\frac{c}{c+\epsilon}d \in \mathcal{D}_{n,c}$.

As $\xi(\hat{P})(O)$ is uniformly bounded, there exists a constant $L > 0$ such that for any d_1, d_2, we have that

$$|\xi(\hat{P})(O;d_1) - \xi(\hat{P})(O;d_2)| \leq L \sup_{x\in\mathcal{X}} |d_1(x) - d_2(x)|.$$

Thus, $\forall\epsilon > 0$, $\exists N_1 \in \mathbb{N}$, such that for all $n > N_1$,

$$\begin{aligned} V_n^{(3)} &= \max_{d\in\mathcal{D}_c} P_n\xi(\hat{P})(O) - \max_{d\in\mathcal{D}_{n,c}} P_n\xi(\hat{P})(O) \\ &\leq \max_{d\in\mathcal{D}_c} P_n\xi(\hat{P})(O) - \max_{d\in\frac{c}{c+\epsilon}\mathcal{D}_c} P_n\xi(\hat{P})(O) \\ &\leq \frac{\epsilon}{c+\epsilon}L, \end{aligned}$$

and similarly, we can obtain that $\exists N_2 \in \mathbb{N}$, such that for all $n > N_2$,

$$V_n^{(3)} \geq -\frac{\epsilon}{c+\epsilon}L,$$

which in combination implies that $V_n^{(3)} = o_p(1)$.

Next, we prove our result (ii) for the regret. Note that

$$V(d^*) - V(\hat{d}) = \{V(d^*) - \hat{V}(d^*)\} + \{\hat{V}(d^*) - \hat{V}(\hat{d})\} + \{\hat{V}(\hat{d}) - V(\hat{d})\}.$$

We analyze the three terms as follows. By the same argument for proving (i), we have that

$$\begin{aligned} V(d^*) - \hat{V}(d^*) &= E[\xi(P)(O;d^*)] - P_n\xi(\hat{P})(O;d^*) = o_p(1), \\ \hat{V}(\hat{d}) - V(\hat{d}) &= P_n\xi(\hat{P})(O;\hat{d}) - E[\xi(P)(O;\hat{d})] = o_p(1). \end{aligned}$$

Also by a similar argument, we have that for any $d \in \mathcal{D}$ and $\epsilon > 0$, $\exists N_2 \in \mathbb{N}$, for all $n > N_2$, $\frac{c}{c+\epsilon}d \in \mathcal{D}_{n,c}$, and

$$\begin{aligned} \hat{V}(d^*) - \hat{V}(\hat{d}) &= \hat{V}(d^*) - \hat{V}\left(\frac{c}{c+\epsilon}d^*\right) + \hat{V}\left(\frac{c}{c+\epsilon}d^*\right) - \hat{V}(\hat{d}) \\ &\leq \frac{\epsilon}{c+\epsilon}L, \end{aligned}$$

and also that for any $d \in \mathcal{D}$ and $\epsilon > 0$, $\exists N_3 \in \mathbb{N}$, for all $n > N_3$, $\frac{c}{c+\epsilon}\hat{d} \in \mathcal{D}_{n,c}$, and

$$V(d^*) - V(\hat{d}) \geq V\left(\frac{c}{c+\epsilon}\hat{d}\right) - V(\hat{d}) \geq -\frac{\epsilon}{c}L,$$

so we conclude that $V(d^*) - V(\hat{d}) = o_p(1)$, which completes the proof.

G Additional simulations

In this section, we present additional simulation results.

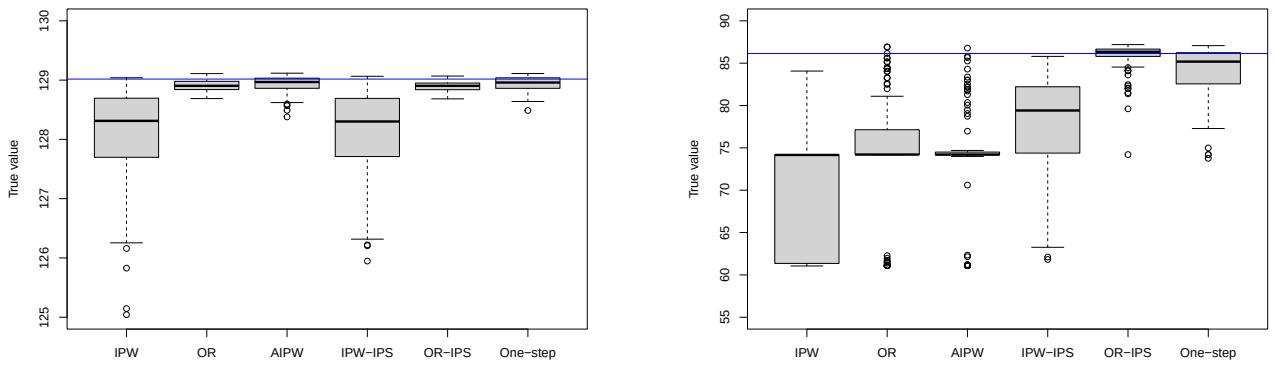
G.1 Incremental propensity score policy learning with sufficent overlap

We examine the performance of our proposed methods by comparison with standard policy learning methods, when sufficient overlap indeed holds. We consider the following data generating process:

$$
\begin{aligned}
(X_1, X_2) &\sim \text{Uniform}(0, 1), \\
(X_3, X_4) &\sim \mathcal{N}\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right\}, \\
A &\sim \text{Bernoulli}(\text{expit}(0.3 - 0.4X_1 - 0.2X_2 - 0.3X_3 + 0.1X_4)), \\
Y(0) &\sim \mathcal{N}\{20(1 + X_1 - X_2 + X_3^2 + \exp(X_2)), 20^2\}, \\
Y(1) &\sim \mathcal{N}\{20(1 + X_1 - X_2 + X_3^2 + \exp(X_2)) + 25(3 - 5X_1 + 2X_2 - 3X_3 + X_4), 20^2\}.
\end{aligned}
$$

We perform the vanilla direct policy search tasks without constraint. Hence, the optimal policy is simply $d^*(x) = I\{3 - 5X_1 + 2X_2 - 3X_3 + X_4 > 0\}$. For standard methods, we consider the policy class of linear rules $\mathcal{D}_{\text{linear}} = \{d(x) = I\{(1, x_1, x_2, x_3, x_4)\beta > 0\} : \beta \in \mathbb{R}^5, \|\beta\|_2 = 1\}$. For the incremental propensity score policies, we consider the class $\mathcal{D}_{\text{IPS}} = \{d(x) = \delta(x;\beta)\pi(x)/\{\delta(x;\beta)\pi(x) + 1 - \pi(x)\} : \beta \in \mathbb{R}^5\}$, which is indexed by $\delta(x;\beta) = \exp\{(1, x_1, x_2, x_3, x_4)\beta\}$.

We estimate the outcome regression model $\mu(x)$ and the propensity score $\pi(x)$ using the generalized random forests (Athey et al., 2019) implemented in the `R` package `grf`. The unconstrained optimization problems are solved by the genetic algorithm (Sekhon and Mebane, 1998) implemented in the `R` package `rgenoud`. The sample size is $n = 2000$. We compare the true values of the estimated optimal policies using test data with sample size $N = 10^5$. The true optimal value is approximated using the test data. Simulation results of 100 Monte Carlo repetition are reported in Figure 2a.



(a) Sufficient overlap.

(b) Parametric models.

Figure 2: Performance of optimal policies under three standard methods (IPW, OR, AIPW) and our proposed three methods (IPW-IPS, OR-IPS, One-step). The blue line is the (approximate) true optimal value.

Despite the fact that the true optimal rule is included in the standard policy class of linear rules but not in our proposed class of incremental propensity score policies, we still observe comparable performance of both classes, which exemplifies the effectiveness of our proposed methods.

G.2 Incremental propensity score policy learning with parametric models

We examine the performance of our proposed methods by comparison with standard policy learning methods, when using correctly specified parametric models.

The simulation setup is the same as in the main paper where the positivity assumption is violated, except that the sample size $n = 500$ is smaller and the outcome regression $\mu(s, x)$ and the propensity score $\pi(s, x)$ models are estimated by correctly specified parametric models. Simulation results of 100 Monte Carlo repetition are reported in Figure 2b. The standard methods IPW, OR, and AIPW have the worst performance. The IPW-IPS estimator still has large variability, and the OR-IPS and efficient one-step estimators achieve the best performance with the highest value.

H Diabetes data analysis

In this section, we provide supplementary information on our Diabetes data analysis.

The original dataset is available in the UCI Repository Diabetes 130-US hospitals for years 1999-2008 (Strack et al., 2014). The `Fairlearn` open source project (Weerts et al., 2023) provides full dataset pre-processing script in `python` on GitHub. We follow these pre-processing steps, and provide the `R` script.

The dataset contains 101766 patients, and a detailed description of the 25 variables are available at the `Fairlearn` project. Originally, the categories of race include "African American", "Asian", "Caucasian", "Hispanic", "Other", "Unknown", and the categories of age include "30 years or younger", "$30 - 60$ years", "Over 60 years". We dichotomize them, so the resultant categories of race include "Caucasian" or "Non-Caucasian", and the resultant categories of age include "30 years or younger" or "Over 30 years".

The empirical CDF of estimated propensity scores for the Diabetes data is plotted in Figure 3. Since many of the propensity scores are close to 0, we conclude that the positivity violation is severe.

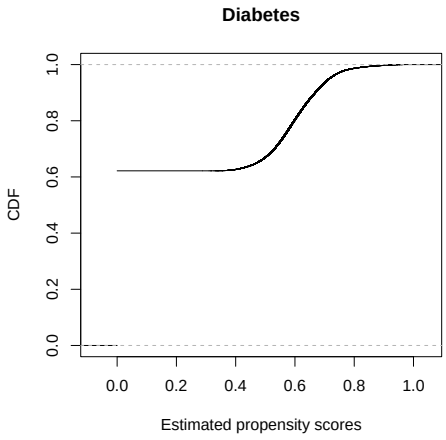


Figure 3: Empirical CDF of estimated Diabetes propensity scores.

The missing data are completed by multivariate imputation by chained equations, implemented in the `R` package `mice`.