Distributionally Robust Policy Evaluation and Learning for Continuous Treatment with Observational Data

Cheuk Hang Leung[1*], **Yiyan Huang**[2*], **Yijun Li**[1], **Qi Wu**[1†]
[1]City University of Hong Kong
[2]The Hong Kong Polytechnic University
chleung87@cityu.edu.hk, yiyhuang3-c@my.cityu.edu.hk, yijunli5-c@my.cityu.edu.hk, qi.wu@cityu.edu.hk

Abstract

Using offline observational data for policy evaluation and learning allows decision-makers to evaluate and learn a policy that connects characteristics and interventions. Most existing literature has focused on either discrete treatment spaces or assumed no difference in the distributions between the policy-learning and policy-deployed environments. These restrict applications in many real-world scenarios where distribution shifts are present with continuous treatment. To overcome these challenges, this paper focuses on developing a distributionally robust policy under a continuous treatment setting. The proposed distributionally robust estimators are established using the Inverse Probability Weighting (IPW) method extended from the discrete one for policy evaluation and learning under continuous treatments. Specifically, we introduce a kernel function into the proposed IPW estimator to mitigate the exclusion of observations that can occur in the standard IPW method to continuous treatments. We then provide finite-sample analysis that guarantees the convergence of the proposed distributionally robust policy evaluation and learning estimators. The comprehensive experiments further verify the effectiveness of our approach when distribution shifts are present.

Introduction

Most decision-making problems necessitate learning an effective personalized policy based on individual features from observational data. This process, commonly referred to as offline policy evaluation/learning, has diverse applications across various domains, including healthcare (Tang and Wiens 2021), recommendation (Li et al. 2010), and finance (Qin et al. 2022). Many studies have investigated offline policy evaluation/learning in discrete treatment settings which assume that the deployment environment is identical to the environment generating the training data, i.e., that there are no distributional shifts. This assumption, however, is often unrealistic in many real-world applications (Huang et al. 2023). For instance, an investment firm has developed an automated investment strategy for the US stock market based on extensive historical trading data. When attempting to apply this strategy directly to the UK stock market, it may lose the predictive power due to the substantial difference between financial market environments. Similarly, a pharmaceutical company has developed a strategy for individualized Warfarin dosage adjustment according to their recent research on older adults. This strategy may perform well in the original clinical trial population, yet it may falter when applied to a new population, such as young adults, due to significant differences in physical conditions.

To address the challenge of distribution shifts in policy evaluation and learning, the problem can be formulated as a Distributionally Robust Optimization (DRO) problem. In the DRO framework, the goal is to find the worst-case solution within a set of distributions under certain degrees of model uncertainties. The uncertainty set is assumed to contain the distributions due to potential distribution shifts, and it can be characterized by constraining certain moments of order (Delage and Ye 2010; Zymler, Kuhn, and Rustem 2013) or by using divergence measures (Hu and Hong 2013; Kuhn et al. 2019; Chen, Sim, and Xu 2019; Gao, Chen, and Kleywegt 2022) to define appropriate deviations from a nominal distribution. The resulting solution provides robust, reliable, and conservative guarantees which can cope with the most adverse situations.

Furthermore, the objective function of the formulated DRO utilizes an inverse probability weighting (IPW) estimator (Wooldridge 2007) to estimate the expected potential reward under continuous treatment. Specifically, we extend the existing IPW estimator designed for discrete treatment settings to accommodate continuous treatments. Generally, the discrete-based IPW approach cannot be directly applied in continuous treatment settings, as it would reject most observed data. Moreover, although discretizing continuous treatments into categories is an intuitive and simple solution, it can lead to information loss and may fail to produce inferences that vary continuously with the treatment. We introduce a modified IPW approach incorporating a scaled kernel function with a bandwidth parameter, serving as a smooth nonparametric extension for computing histogram "buckets". The proposed IPW estimator enables the distributionally robust policy evaluation and learning using observational datasets.

In summary, our framework addresses the challenges of policy evaluation and learning in continuous treatment settings in the presence of distribution shifts. The key contribu-

[*]The co-first authors contribute equally.
[†]Corresponding author.

tions of our paper are threefold:

1. We formulate the DRO problem with an IPW approach for policy evaluation/learning under the continuous treatment setting, and convert it to its equivalent dual form. As the standard IPW approach is not directly applicable in this context, we develop a tractable kernel-based form to approximate the dual problem.
2. We establish estimators for policy evaluation/learning and investigate their asymptotic properties. Specifically, the established estimators of distributionally robust values exhibit asymptotic normality, and the finite-sample regret decays to zero asymptotically.
3. Through simulated and empirical studies, we demonstrate that the policy learned using our method provides robustness to distribution shifts compared to standard nonrobust policy learning methods.

Literature Review

Considerable research has focused on causality in discrete treatment settings. However, exploring causality under continuous treatment remains limited in many real-world applications. Existing research on continuous treatment settings primarily focuses on directly modeling the relationship among response, treatment, and covariates. Notable contributions include (Schwab et al. 2020), who construct a multi-head neural network for this purpose; (Bica, Jordon, and van der Schaar 2020), who propose an end-to-end neural network based on generative adversarial networks (GANs); and (Bahadori, Tchetgen, and Heckerman 2022), who introduce a novel algorithm within the entropy balancing framework to optimize accuracy through end-to-end optimization. Another approach modifies IPW-based and Doubly Robust-based estimators (e.g., (Chernozhukov et al. 2018; Huang et al. 2022)) from discrete treatment settings by incorporating kernel functions to mitigate the direct rejection of observed data, as demonstrated by (Su, Ura, and Zhang 2019) and (Colangelo and Lee 2019).

Recent studies have focused on offline policy evaluation and learning. (Kitagawa and Tetenov 2018) establish finite sample regret bounds with a rate of $O_P\left(1/\sqrt{N}\right)$ for policy learning over a policy class with finite VC dimension. (Athey and Wager 2021) extend this analysis to examine regret bounds from an asymptotic perspective. (Zhao et al. 2012) and (Zhou et al. 2017) propose algorithms for policy learning and explore the statistical properties of learned policies and associated regret bounds. (Dudík, Langford, and Li 2011) utilize classic estimators for policy evaluation. (Kallus 2018) proposes a balance-based approach to reweight historical data and mimic datasets generated by evaluated or learned policies. (Zhou, Athey, and Wager 2023) exploit a cross-fitted approach for policy learning. The aforementioned studies primarily assume discrete treatment and do not account for distributional shifts. Nevertheless, distributional shifts are common since the studies are often conducted in different environments, highlighting the significance of studying distributionally robust policies. For instance, (Yang et al. 2023), (Shen, Xu, and Zavlanos 2024), (Mo, Qi, and Liu 2021), and (Faury et al. 2020) primarily focus on shifts in covariates, whereas (Si et al. 2023) and (Kallus et al. 2022) address shifts in the joint distribution of responses, features, and treatments. Notably, to the best of our knowledge, studying policy evaluation and learning in the presence of distribution shifts under continuous treatment settings is still an open problem.

Background

Notations and Assumptions

Throughout the paper, we denote $A \in \mathcal{A} \subset \mathbb{R}$, $X \in \mathcal{X} \subset \mathbb{R}^d$, and $Y \in \mathcal{Y} \subset \mathbb{R}$ as the continuous treatment (also known as action or intervention), the covariates, and the continuous response (also known as outcome), respectively. We write $Y(A)$ to mean the potential response variable under the treatment A. We also assume that Y and $Y(A)$ are non-negative bounded variables, i.e., there exists $M > 0$ such that $0 \leq Y(a),\ Y \leq M$. Finally, we let $(X_i, A_i, Y_i)_{i=1}^N$ be N independent and identically distributed (i.i.d.) triples from a fixed underlying distribution, and the probability measure of the underlying distribution is denoted as $\mathbb{P}_0$.

Further, we adopt the Rubin potential outcome framework (e.g., (Rubin 1974; Imbens 2004; Imbens and Rubin 2015; Huang et al. 2021, 2024; Li et al. 2024)). Throughout the paper, we impose the following (*causal*) assumptions that are standard in the causal inference literature:

Assumption 1 (*Consistency*)**.** *If $A = a$, we have $Y = Y(a)$.*

Assumption 2 (*Unconfoundedness*)**.** $Y(a) \perp\!\!\!\perp A|X,\ \forall a$.

Assumption 3 (*Positivity*)**.** *There exists a positive constant $\epsilon > 0$ such that $\inf_{a\in\mathcal{A}} \operatorname{ess\,inf}_{x\in\mathcal{X}} f_0(a|x) \geq \epsilon > 0$.*

Additionally, we follow (Kallus and Zhou 2018; Colangelo and Lee 2019) and impose differentiability assumptions on the probability density functions $f_0(y|a,x)$ and $f_0(a|x)$: $f_0(y|a,x)$ and $f_0(a|x)$ are three-times differentiable w.r.t. a and bounded uniformly on $(y,a,x) \in (\mathcal{Y},\mathcal{A},\mathcal{X})$. All proofs of the theorems are in the Appendix[1].

Problem Setup

Our objective is to find a policy π^* that maps $\mathcal{X}$ to $\mathcal{A}$ within a policy class Π that maximizes the expected outcomes , i.e.,

$$\pi^* = \arg\max_{\pi\in\Pi} Q(\pi) \coloneqq \arg\max_{\pi\in\Pi} \mathbb{E}_{\mathbb{P}_0}[Y(\pi(X))]. \tag{1}$$

The learned policy obtained in Eqn. (1) may not generalize well to a new environment with a distribution that differs from $\mathbb{P}_0$. As such, we can consider distributionally robust formulation of Eqn. (1):

$$\begin{aligned} \pi^*_{\mathrm{DRO}} &= \arg\max_{\pi\in\Pi} Q_{\mathrm{DRO}}(\pi), \quad \text{where} \\ Q_{\mathrm{DRO}}(\pi) &= \inf_{\mathbb{P}\in\mathcal{U}_{\mathbb{P}_0}(\eta)} \mathbb{E}_{\mathbb{P}}[Y(\pi(X))], \\ \mathcal{U}_{\mathbb{P}_0}(\eta) &= \{\mathbb{P} : \mathbb{D}(\mathbb{P}\|\mathbb{P}_0) \leq \eta\}. \end{aligned} \tag{2}$$

Here, $\mathbb{D}(\cdot\|\cdot)$ denotes the distribution discrepancy. Throughout the paper, we choose it as the Kullback–Leibler (KL)

[1] Available in the "proof" file of the Supplementary Material.

divergence (Kullback 1959; Kullback and Leibler 1951)[2]. $\mathcal{U}_{\mathbb{P}_0}(\eta)$ is the *ambiguity set* (also known as the uncertainty set) with an *ambiguity radius* η. The ambiguity set contains all the possible distributions $\mathbb{P}$ such that the discrepancy of $\mathbb{P}$ relative to $\mathbb{P}_0$ is at most η.

Distributionally Robust Policy Evaluation

The Estimation of $Q_{\text{DRO}}(\pi)$

As proven in (Hu and Hong 2013), Eqn. (2) is equivalent to solving its Lagrangian dual, which is given as follows:

$$\begin{aligned} &-\min_{\alpha\geq 0}\left\{\alpha\log\mathbb{E}\left[e^{\frac{-Y(\pi(X))}{\alpha}}\right]+\alpha\eta\right\} \\ &=\max_{\alpha\geq 0}\left\{-\alpha\log\mathbb{E}\left[e^{\frac{-Y(\pi(X))}{\alpha}}\right]-\alpha\eta\right\}\coloneqq\max_{\alpha\geq 0}\phi(\pi,\alpha). \end{aligned} \tag{3}$$

Since $Y(\pi(X))$ in Eqn. (3) is inaccessible, we reformulate $\mathbb{E}[e^{\frac{-Y(\pi(X))}{\alpha}}]$ to an IPW form similar to that in (Horvitz and Thompson 1952). The result is given in Lemma 1.

Lemma 1. *Under Assumptions 1 - 3, we have*

$$\mathbb{E}\left[e^{\frac{-Y(\pi(X))}{\alpha}}\right]=\mathbb{E}\left[\frac{\delta(\pi(X)-A)}{f_0(A|X)}e^{\frac{-Y}{\alpha}}\right] \tag{4}$$

for any $\alpha\geq 0$, where $\delta(\cdot)$ is the Dirac Delta function[3].

Using Lemma 1, the expectation in Eqn. (3) can be replaced according to Eqn. (4). Note that the Dirac function $\delta(\cdot)$ is a theoretical generalized function and is often approximated by the scaled kernel function $K_h(\cdot)$[4]. As a result, we can consider the following approximated form:

$$Q^h_{\text{DRO}}(\pi)=\sup_{\alpha\geq 0}\left\{-\alpha\log\mathbb{E}\left[\frac{e^{-\frac{Y}{\alpha}}K_h(\pi(X)-A)}{f_0(A|X)}\right]-\alpha\eta\right\}. \tag{5}$$

The above two quantities, $Q_{\text{DRO}}(\pi)$ and $Q^h_{\text{DRO}}(\pi)$, bring two important insights: (1) The optimal solutions of $Q_{\text{DRO}}(\pi)$ and $Q^h_{\text{DRO}}(\pi)$ are obtained by solving Eqns. (3) and (5) which are attainable for positive α due to the causal assumption. Further, the optimal solutions are finite for any π (see Auxiliary Result 2 in Appendix for details); (2) $Q^h_{\text{DRO}}(\pi)\to Q_{\text{DRO}}(\pi)$ as $h\to 0$. These two insights, consequently, guarantee that the optimal solutions of $Q^h_{\text{DRO}}(\pi)$ also converge to the optimal solutions of $Q_{\text{DRO}}(\pi)$. Therefore, we can construct an estimator of the IPW-based distributionally robust value $Q^h_{\text{DRO}}(\pi)$ to study the original distributionally robust value $Q_{\text{DRO}}(\pi)$ in Eqn. (2). We define

$$\bar{W}^h_N(\pi,\alpha)=\frac{1}{N}\sum_{i=1}^{N}\frac{K_h(\pi(X_i)-A_i)}{f_0(A_i|X_i)}e^{\frac{-Y_i}{\alpha}}. \tag{6a}$$

$$\hat{W}^h_N(\pi,\alpha)=\frac{\bar{W}^h_N(\pi,\alpha)}{S^h_N}=\frac{\bar{W}^h_N(\pi,\alpha)}{\frac{1}{N}\sum_{i=1}^{N}\frac{K_h(\pi(X_i)-A_i)}{f_0(A_i|X_i)}}. \tag{6b}$$

It is known that the IPW-based estimator $\bar{W}^h_N(\pi,\alpha)$ in Eqn. (6a) suffers from high-variance (Swaminathan and Joachims 2015; Khan and Ugander 2023). To address this challenge, we can use a normalized estimator $\hat{W}^h_N(\pi,\alpha)$ with a normalization factor S^h_N in Eqn. (6b) to approximate $\bar{W}^h_N(\pi,\alpha)$. Note that $\mathbb{E}[S^h_N]=1$ and $S^h_N\to 1$ almost surely (see Auxiliary Result 1 in Appedix). Thus, $\hat{W}^h_N(\pi,\alpha)$ is asymptotically equivalent to $\bar{W}^h_N(\pi,\alpha)$. Consequently, we use the following $\hat{Q}^h_{\text{DRO}}(\pi)$ as the estimator of $Q^h_{\text{DRO}}(\pi)$ in Eqn. (5):

$$\begin{aligned}\hat{Q}^h_{\text{DRO}}(\pi)&=\max_{\alpha\geq 0}\hat{\phi}^h_N(\pi,\alpha)\\ &\coloneqq\max_{\alpha\geq 0}\big\{-\alpha\log\hat{W}^h_N(\pi,\alpha)-\alpha\eta\big\}.\end{aligned} \tag{7}$$

$\hat{Q}^h_{\text{DRO}}(\pi)$ can be used to evaluate distributional robustness of a policy π. To summarize, we present the specific steps of obtaining $\hat{Q}^h_{\text{DRO}}(\pi)$ in the following Algorithm 1.

Algorithm 1: Distributionally robust policy evaluation

Input observed dataset $(X_i,A_i,Y_i)_{i=1}^N$, h, policy $\pi\in\Pi$. Initialize: $\alpha\in\mathbb{R}^+\cup 0$.

1: **repeat**
2: Compute $\hat{W}^h_N(\pi,\alpha)$ given in Eqn. (6b).
3: Update α: $\alpha\leftarrow\alpha-\frac{\frac{\partial\hat{\phi}^h_N}{\partial\alpha}}{\frac{\partial^2\hat{\phi}^h_N}{\partial\alpha^2}}$, where

$$\frac{\partial\hat{\phi}^h_N}{\partial\alpha}=-\eta-\log\hat{W}^h_N-\frac{\alpha\frac{\partial\hat{W}^h_N}{\partial\alpha}}{\hat{W}^h_N}$$

$$\frac{\partial^2\hat{\phi}^h_N}{\partial\alpha^2}=-\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{K_h(\pi(X_i)-A_i)}{f_0(A_i|X_i)}Y_i^2e^{-\frac{Y_i}{\alpha}}}{\alpha^3S^h_N\hat{W}^h_N}+\frac{\alpha\big(\frac{\partial\hat{W}^h_N}{\partial\alpha}\big)^2}{(\hat{W}^h_N)^2},$$

$$\frac{\partial\hat{W}^h_N}{\partial\alpha}=\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{K_h(\pi(X_i)-A_i)}{f_0(A_i|X_i)}Y_ie^{-\frac{Y_i}{\alpha}}}{\alpha^2S^h_N}$$

4: **until** α converges
5: **Return** $\hat{Q}^h_{\text{DRO}}(\pi)\leftarrow\hat{\pi}^h_N(\pi,\alpha)$

[2]Other measures such as the Wasserstein metric or other ϕ-divergence measures can be utilized (e.g., see (Kuhn et al. 2019) and (Husain, Nguyen, and van den Hengel 2023)). However, these approaches typically involve solving multi-level optimization problems which can be challenging to analyze.

[3]$\delta(x)=\begin{cases}\infty & x=0\\ 0 & \text{otherwise}\end{cases}$ such that i) $\int_{\mathbb{R}}\delta(x)dx=1$ and ii) $\int_{\mathbb{R}}\delta(x)f(x)dx=f(0)$ for any arbitrary f defined on $\mathbb{R}$.

[4]A bounded differentiable function $K(\cdot)$ (i.e., $|K(\cdot)|\leq M_K$) is said to be a second-order kernel function if it satisfies *i)* $K(\cdot)$ is a symmetric function; *ii)* $\int_{-\infty}^{\infty}uK(u)du=0$; *iii)* $\int_{-\infty}^{\infty}K(u)du=1$. The scaled kernel function $K_h(\cdot)$ is defined such that $K_h(x)=\frac{1}{h}K\big(\frac{x}{h}\big)$, where h is termed as the *bandwidth* parameter. Note that $K_h(x)\xrightarrow{w}\delta(x)$ when $h\to 0$. Examples of kernels include Gaussian kernels or the Epanechnikov kernel.

The Statistical Property of $\hat{Q}^h_{\mathrm{DRO}}(\pi)$

As $\hat{Q}^h_{\mathrm{DRO}}(\pi)$ is an estimator established using observed empirical samples, it is important to delve into the finite-sample statistical performance guarantee for the estimator $\hat{Q}^h_{\mathrm{DRO}}(\pi)$. To achieve this, we first discuss the theoretical property of $\hat{W}^h_N$ in Theorem 2.

Theorem 2. *Suppose that $N \to \infty$, $h \to 0$ such that $Nh \to \infty$ and $Nh^5 \to C \in [0, \infty)$. Then we have*

$$\sqrt{Nh}\left(\hat{W}^h_N - \mathbb{E}[e^{-\frac{Y(\pi(X))}{\alpha}}] - B_\pi(\alpha)h^2\right) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}_\pi(\alpha)),$$

$$\text{where} \quad B_\pi(\alpha) = \frac{\left(\int u^2 K(u)du\right)}{2} \times \mathbb{E}\left[\mathbb{E}\left[e^{\frac{-Y}{\alpha}} \frac{\partial^2_{aa} f_0(Y|\pi(X), X)}{f_0(Y|\pi(X), X)} \middle| A = \pi(X), X\right]\right], \tag{8}$$

$$\mathbb{V}_\pi(\alpha) = \left(\int K(u)^2 du\right) \times \left\{\mathbb{E}\left[\mathbb{E}\left[\frac{e^{-\frac{2Y}{\alpha}}}{f_0(\pi(X)|X)} \middle| A = \pi(X), X\right]\right] + \mathbb{E}\left[\frac{1}{f_0(\pi(X)|X)}\right](\mathbb{E}[e^{-\frac{Y(\pi(X))}{\alpha}}])^2 - 2\mathbb{E}\left[\mathbb{E}\left[\frac{e^{-\frac{Y}{\alpha}}}{f_0(\pi(X)|X)} \middle| A = \pi(X), X\right]\right]\mathbb{E}[e^{-\frac{Y(\pi(X))}{\alpha}}]\right\}. \tag{9}$$

The estimator $\hat{W}^h_N$ is the key component of $\hat{Q}^h_{\mathrm{DRO}}(\pi)$, as shown in Eqn. (7). Consequently, based on the statistical property of $\hat{W}^h_N$, we can derive the asymptotic normality of $\hat{Q}^h_{\mathrm{DRO}}(\pi)$ in Theorem 3.

Theorem 3. *Suppose that $N \to \infty$, $h \to 0$ such that $Nh \to \infty$ and $Nh^5 \to C \in [0, \infty)$. Further, denote $\alpha_*(\pi)$ s.t. $\phi(\pi, \alpha_*(\pi)) \geq \phi(\pi, \alpha)\ \forall\ \alpha \geq 0$. Then we have*

$$\sqrt{Nh}\left(\hat{Q}^h_{\mathrm{DRO}}(\pi) - Q_{\mathrm{DRO}}(\pi) + \frac{\alpha_*(\pi)B_\pi(\alpha_*(\pi))}{\mathbb{E}\left[e^{-\frac{Y(\pi(X))}{\alpha_*(\pi)}}\right]}h^2\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\alpha^2_*(\pi)\mathbb{V}_\pi(\alpha_*(\pi))}{\left(\mathbb{E}\left[e^{-\frac{Y(\pi(X))}{\alpha_*(\pi)}}\right]\right)^2}\right).$$

A good choice of bandwidth is essential for effective policy learning and evaluation. We can use a rule-of-thumb bandwidth (see e.g., (Su, Ura, and Zhang 2019)), or select h^* by minimizing the asymptotic mean squared error (AMSE) (e.g., (Kallus and Zhou 2018)) of $\hat{Q}^h_{\mathrm{DRO}}(\pi)$:

$$h^* := \arg\min\left[B_\pi(\alpha_*(\pi))^2 h^4 + \frac{\mathbb{V}_\pi(\alpha_*(\pi))}{Nh}\right] \Rightarrow h^* = \left(\frac{\mathbb{V}_\pi(\alpha_*(\pi))}{4NB_\pi(\alpha_*(\pi))}\right)^{\frac{1}{5}} = \Theta(N^{-\frac{1}{5}}). \tag{10}$$

Empirically, we would follow the notions presented in (Kallus and Zhou 2018), of which we choose the optimal bandwidth via a plug-in estimator.

Distributionally Robust Policy Learning

The Estimation of π^*_{DRO}

In the preceding section, we have established $\hat{Q}^h_{\mathrm{DRO}}(\pi)$ as an estimator for $Q_{\mathrm{DRO}}(\pi)$. Next, we aim to construct an estimator for the optimal policy π^*_{DRO}. Specifically, we derive $\hat{\pi}^h_{\mathrm{DRO}}$ from $\hat{Q}^h_{\mathrm{DRO}}(\pi)$ such that

$$\begin{aligned}\hat{\pi}^h_{\mathrm{DRO}} &= \arg\max_{\pi\in\Pi} \hat{Q}^h_{\mathrm{DRO}}(\pi) \\ &= \arg\max_{\pi\in\Pi} \max_{\alpha\geq 0} \left\{ -\alpha \log \hat{W}^h_N(\pi, \alpha) - \alpha\eta \right\}.\end{aligned}$$

$\hat{\pi}^h_{\mathrm{DRO}}$ is the distributionally robust policy learned from $\hat{Q}^h_{\mathrm{DRO}}(\pi)$. To summarize, we present the specific steps of obtaining $\hat{\pi}^h_{\mathrm{DRO}}$ in Algorithm 2.

Algorithm 2: Distributionally robust policy learning

Input observed dataset $(X_i, A_i, Y_i)_{i=1}^N$, h. Initialize: $\pi \in \Pi$ and $\alpha \in \mathbb{R}^+ \cup 0$.

1: **repeat**
2: Compute $\hat{W}^h_N(\pi, \alpha)$ given in Eqn. (6b).
3: Solve $\min_{\pi\in\Pi} \hat{W}^h_N(\pi, \alpha)$ using any numerical methods. Update π: $\pi \leftarrow \arg\min_{\pi\in\Pi} \hat{W}^h_N(\pi, \alpha)$.
4: Solve $\max_{\alpha\geq 0} \hat{\phi}^h_N(\pi, \alpha)$ using any numerical methods where $\hat{\phi}^h_N(\pi, \alpha)$ is given in Eqn. (7). Update α: $\alpha \leftarrow \arg\max_{\alpha\geq 0} \hat{\phi}^h_N(\pi, \alpha)$.
5: **until** α converges
6: **Return** $\hat{\pi}^h_{\mathrm{DRO}} \leftarrow \pi$

The Statistical Property of $\hat{\pi}^h_{\mathrm{DRO}}$

An essential aspect of our study is examining the statistical performance guarantee of $\hat{\pi}^h_{\mathrm{DRO}}$, which enables researchers to assess the gap between the learned policy $\hat{\pi}^h_{\mathrm{DRO}}$ and the optimal distributionally robust policy $\pi^*_{\mathrm{DRO}} = \max_{\tilde{\pi}\in\Pi} Q_{\mathrm{DRO}}(\tilde{\pi})$. To achieve this, we use the distributionally robust regret defined in Definition 4 as the evaluation metric.

Definition 4. *Let the optimal distributionally robust policy be $\pi^*_{\mathrm{DRO}} = \arg\max_{\tilde{\pi}\in\Pi} Q_{\mathrm{DRO}}(\tilde{\pi})$. The distributionally robust regret of a policy $\pi \in \Pi$, denoted by $R_{\mathrm{DRO}}(\pi)$, is then defined as*

$$\begin{aligned}&R_{\mathrm{DRO}}(\pi) \\ &= \max_{\tilde{\pi}\in\Pi} \inf_{\mathbb{P}\in\mathcal{U}_{\mathbb{P}_0}(\eta)} \mathbb{E}_{\mathbb{P}}[Y(\tilde{\pi}(X))] - \inf_{\mathbb{P}\in\mathcal{U}_{\mathbb{P}_0}(\eta)} \mathbb{E}_{\mathbb{P}}[Y(\pi(X))] \\ &= \max_{\tilde{\pi}\in\Pi} Q_{\mathrm{DRO}}(\tilde{\pi}) - Q_{\mathrm{DRO}}(\pi) = Q_{\mathrm{DRO}}(\pi^*_{\mathrm{DRO}}) - Q_{\mathrm{DRO}}(\pi).\end{aligned}$$

Before studying $R_{\mathrm{DRO}}(\hat{\pi}^h_{\mathrm{DRO}})$, we will now introduce the required notions of the Rademacher complexity and the covering number of a functional class (Shalev-Shwartz and Ben-David 2014; Mohri, Rostamizadeh, and Talwalkar 2018; Wainwright 2019), which are stated in Definition 5.

Definition 5. *Let $\mathcal{F}$ be a family of real-valued functions f where $f : \mathcal{Z} \to \mathbb{R}$. Given $Z_1, \cdots, Z_N \in \mathcal{Z}$, the Rademacher complexity of $\mathcal{F}$ is defined as $\mathcal{R}_N(\mathcal{F})$ such that*

$$\mathcal{R}_N(\mathcal{F}) = \mathbb{E}_Z[\hat{\mathcal{R}}_N(\mathcal{F})] = \mathbb{E}_{Z,\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i f(Z_i)\right|\right],$$

$$\hat{\mathcal{R}}_N(\mathcal{F}) \coloneqq \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i f(Z_i)\right|\middle| Z_1, \cdots, Z_N\right].$$

Here, $\sigma_1, \cdots, \sigma_N$ are i.i.d. with the distribution $\mathbb{P}\{\sigma_i = 1\} = \mathbb{P}\{\sigma_i = -1\} = \frac{1}{2}$. Additionally, consider a set $\{X_1, \cdots, X_N\}$ in a metric space with metric $\|\cdot\|$. A set $\mathcal{A}_{\{X_1,\cdots,X_N\}} \subset \mathcal{F}$ is said to be a t-covering of $\mathcal{F}$ if, for any $f \in \mathcal{F}$, there exists $\tilde{f} \in \mathcal{A}_{\{X_1,\cdots,X_N\}}$ such that $\|(f(X_1), \cdots, f(X_N)) - (\tilde{f}(X_1), \cdots, \tilde{f}(X_N))\| \le t$. The size of the smallest t-covering, denoted by $\mathfrak{N}\big(t, \mathcal{F}(\{X_1, \cdots, X_N\}), \|\cdot\|\big)$, is the t-covering number.

With Definition 5, the regret $R_{\mathrm{DRO}}(\pi)$ can be generally upper bounded as the following Theorem 6.

Theorem 6. *Suppose that the kernel function $K(x)$ is bounded where $|K(x)| \le M_K$. Given $\delta > 0$, $h > 0$, and a policy class Π, denote*

$$\mathcal{F}_\Pi \coloneqq \left\{\frac{K_h(\pi(X) - A)}{f_0(A|X)} : \pi \in \Pi\right\},$$

$$\mathcal{F}_{\Pi,x} \coloneqq \left\{\frac{K_h(\pi(X) - A)\mathbf{1}_{\{Y(\pi(X))\le x\}}}{f_0(A|X)} : \pi \in \Pi,\ x \in [0, M]\right\}.$$

Then, with probability $1 - \delta$, we have

$$\begin{aligned} R_{\mathrm{DRO}}(\hat{\pi}^h_{\mathrm{DRO}}) \le & \frac{4}{\epsilon}\mathcal{R}_N(\mathcal{F}_{\Pi,x}) + \frac{4}{\epsilon}\mathcal{R}_N(\mathcal{F}_\Pi) \\ & + \frac{4\sqrt{2}M_K\sqrt{\ln\left(\frac{2}{\delta}\right)}}{h\epsilon^2\sqrt{N}} + O(h^2). \end{aligned} \tag{11}$$

The Rademacher complexities in Eqn. (11) can be further bounded using covering numbers (see, for instance, (Shalev-Shwartz and Ben-David 2014)). Under certain conditions, such as when the square root of the metric entropy (i.e., the logarithm of the covering number) is summable, we can bound $\mathcal{R}_N(\mathcal{F}_{\Pi,x})$ and $\mathcal{R}_N(\mathcal{F}_\Pi)$ by the covering number of Π. This result is presented in detail in Corollary 7.

Corollary 7. *If the kernel function $K(x)$ is Lipschitz continuous with constant $L_K > 0$ (i.e., $|K(x) - K(y)| \le L_K|x - y|$) and there exists a finite value κ which equals*

$$\mathbb{E}\left[\int_0^{\frac{2M_Kh}{L_K}}\sqrt{\log\ \mathfrak{N}\big(t, \Pi(\{X_1, \cdots, X_N\}), \|\cdot\|_{\mathcal{L}_2(\mathbb{P}_N)}\big)}dt\right].$$

Then, for some constant $\mathcal{K}$, Eqn. (11) *becomes*

$$\begin{aligned} & R_{\mathrm{DRO}}(\hat{\pi}^h_{\mathrm{DRO}}) \\ & \le \frac{288L_K\kappa}{\sqrt{N}h^2\epsilon^2} + \frac{192M_K(\sqrt{\log\mathcal{K}} + 2\sqrt{2})}{\sqrt{N}h\epsilon^2} \\ & + \frac{4M_K\sqrt{2\log\left(\frac{2}{\delta}\right)}}{\sqrt{N}h\epsilon^2} + O(h^2). \end{aligned} \tag{12}$$

Note that the distributionally robust regret is independent of η, as it is unaffected by the expectation term in the dual problem. In conjunction with Eqn. (10), selecting $h = O(N^{-\frac{1}{5}})$ in Corollary 7 ensures consistent learning of the optimal linear policy, as the distributionally robust regret $R_{\mathrm{DRO}}(\hat{\pi}^h_{\mathrm{DRO}})$ converges to zero when N tends to infinity.

To conclude this section, we discuss the covering numbers of various policy classes. A common policy class is the linear policy class, defined as $\Pi = \{\pi : \mathcal{X} \to \mathcal{A} | \pi(X) = w^\top X,\ \|w\|_p \le a,\ \|X\|_q \le b,\ w,\ X \in \mathbb{R}^d\}$. For instance, when $p = \infty$ and $q = 1$, we can demonstrate that

$$\mathfrak{N}\big(t, \Pi(X_1, \cdots, X_N), \|\cdot\|_{\mathcal{L}_\infty(\mathbb{P}_N)}\big) \le \left(\frac{\max_{1\le i\le N}\|X_i\|_1}{t} + 2\right)^d.$$

Consequently, κ in Eqn. (12) is bounded above by $\sqrt{d}\left\{\frac{2\sqrt{2}M_Kh}{L_K} + 2\sqrt{\frac{2M_Kh}{L_K}}\mathbb{E}\big[\max_{1\le i\le N}\|X_i\|_1\big]\right\}$, as per its definition. (Zhang 2002) provide the covering number of linear policy class for $2 \le p < \infty$.

We can extend the study from linear policy classes to classes containing non-linear policies such as neural networks or support vector machines (SVMs). For example, shallow neural networks can be represented as linear functions composed with Lipschitz activations. The covering number for the class can be bounded by the Lipschitz constant and the linear class (Zhang 2002; Anthony et al. 1999). Covering numbers for other classes can be found in sources such as (Bartlett, Foster, and Telgarsky 2017).

Experiments

In this section, we mainly investigate the robustness of the proposed policy π^h_{DRO} against distribution shift. Our analysis includes two parts: simulation and empirical studies. First, in the "Simulation Experiment" subsection, we compare results under continuous treatments with those under discretized treatments, as well as outcomes with and without robustness. We evaluate these results in a distributionally robust manner to assess the policy's performance under varying conditions. Following this, in the "Empirical Experiment" subsection, the experiments on Warfarin dataset compare the robustness performance of the robust and nonrobust policies. All experiments are run on a Dell 3640 with an Intel Xeon W-1290P 3.60GHz CPU[5].

Simulation Experiment

Continuous v.s. Discrete. We begin by comparing our distributionally robust policy $\hat{\pi}^h_{\mathrm{DRO}}$ under continuous treatment with the distributionally robust policy $\hat{\pi}^{dis\text{-}k}_{\mathrm{DRO}}$, where the continuous treatment is discretized using the method proposed in (Si et al. 2023) into k bins ($k \in 2, 3, 4$) based on the discretized strategy in (Zhou et al. 2017). To enable a fair comparison between these two forms, we consider a simple data generating process with known optimal values. Specifically, we assume: $X \sim \mathrm{Uniform}(0, 1)$, $A|X \sim \mathrm{Uniform}(X, X+1)$, $Y = 5 + X/A + \epsilon$, $\epsilon \sim \mathrm{Uniform}(0, 1)$. We define the policy class Π as $\{\beta X : 1 \le \beta \le 3\}$, and

[5]Available in the "code" file of the Supplementary Material.

set the ambiguity radius $\eta = 0.05$. With these specifications, we compute the optimal distributionally robust value $Q^* = \max_{\pi\in\Pi} \inf_{\mathbb{P}\in\mathcal{U}_{\mathbb{P}_0}(\eta)} \mathbb{E}_{\mathbb{P}}[Y(\pi(X))]$, which evaluates to 6.41 using numerical approaches. For the bandwidth parameter h, we follow the approach of selecting the bandwidth as given by (Kallus and Zhou 2018) using a plug-in estimator based on Eqn. (10). We generate 100 different datasets, each consisting of 2500 training samples and 2500 test samples.

For policy learning on the training data, both $\hat{\pi}^h_{\mathrm{DRO}}$ and $\hat{\pi}^{dis\text{-}k}_{\mathrm{DRO}}$ are learned using $\eta^{train} = 0.05$. We then compute $\hat{Q}^h_{\mathrm{DRO}}(\hat{\pi}^h_{\mathrm{DRO}})$ and $\hat{Q}^{dis}_{\mathrm{DRO}}(\hat{\pi}^{dis\text{-}k}_{\mathrm{DRO}})$ on the test data and compare the results. For various $\eta^{test} \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$, $\hat{Q}^h_{\mathrm{DRO}}(\hat{\pi}^h_{\mathrm{DRO}})$ is estimated according to Algorithm 1, while $\hat{Q}^{dis}_{\mathrm{DRO}}(\hat{\pi}^{dis\text{-}k}_{\mathrm{DRO}})$ is estimated by solving $\max_{\alpha\geq 0}\{-\alpha \log \hat{W}_N(\hat{\pi}^{dis\text{-}k}_{\mathrm{DRO}}, \alpha) - \alpha\eta\}$. Here $\hat{W}_N(\pi,\alpha) = \left(\sum_{i=1}^N \frac{\mathbf{1}_{\{\pi(X_i)=A_i\}} e^{-\frac{Y_i}{\alpha}}}{\hat{p}_0(A_i|X_i)}\right) / \left(\sum_{j=1}^N \frac{\mathbf{1}_{\{\pi(X_j)=A_j\}}}{\hat{p}_0(A_j|X_j)}\right)$ and $\hat{p}_0(A|X)$ is the estimated probability of receiving treatment A conditioning on X. The results given in Table 1 indicate that the learned policy $\hat{\pi}^h_{\mathrm{DRO}}$ achieves the best robust performance when evaluated using the $\hat{Q}^h_{\mathrm{DRO}}$ metric (see the first row of Table 1). Further, the mean value 6.24 exhibits a significantly smaller gap with the optimal distributionally robust value of 6.41 compared to the discrete-treatment policies evaluated using $\hat{Q}^{dis}_{\mathrm{DRO}}$.

η^{test}				
0.05	0.1	0.2	0.3	0.4
6.24±0.32	6.19±0.33	6.11±0.36	6.04±0.38	5.99±0.40
5.88±0.15	5.81±0.15	5.71±0.15	5.64±0.15	5.58±0.15
5.85±0.12	5.79±0.12	5.70±0.12	5.63±0.12	5.58±0.12
5.83±0.12	5.77±0.12	5.68±0.12	5.61±0.12	5.56±0.12

Table 1: Comparison of robustness performance (continuous v.s. discrete) with $\eta^{train} = 0.05$ for policy learning and various η^{test} for policy evaluation. When $\eta^{train} = \eta^{test} = 0.05$, the optimal distributionally robust value is $Q^* = 6.41$. The reported Mean ± Standard Error (the Standard Error is in %) is computed over 100 runs. The first/second/third/fourth row records values produced by $\hat{Q}^h_{\mathrm{DRO}}(\hat{\pi}^h_{\mathrm{DRO}})$/$\hat{Q}^{dis}_{\mathrm{DRO}}(\hat{\pi}^{dis\text{-}2}_{\mathrm{DRO}})$/$\hat{Q}^{dis}_{\mathrm{DRO}}(\hat{\pi}^{dis\text{-}3}_{\mathrm{DRO}})$/$\hat{Q}^{dis}_{\mathrm{DRO}}(\hat{\pi}^{dis\text{-}4}_{\mathrm{DRO}})$.

Robust v.s. Nonrobust. We then compare our distributionally robust policy $\hat{\pi}_{\mathrm{DRO}}$ with the non-robust policy $\hat{\pi}_{\mathrm{NRO}} \in \arg\max_{\pi\in\Pi} \hat{W}^h_N(\pi)$, where $\hat{W}^h_N(\pi) = \frac{\bar{W}^h_N(\pi)}{S^h_N}$ and $\bar{W}^h_N(\pi) = \frac{1}{N}\sum_{i=1}^N \frac{K_h(\pi(X_i)-A_i)}{f_0(A_i|X_i)} Y_i$, as given in (Kallus and Zhou 2018), with $\Pi = \{\zeta^\top X : \|\zeta\|_\infty \leq 2\}$. We follow (Kallus and Zhou 2018) to simulate i.i.d. data as follows: $X_k \sim \mathrm{Uniform}(-0.2, 0.2)$ for $k = 1$ to 10, $A|X \sim \mathcal{N}(\theta^\top X, 0.1) + X_1 + 2X_2 - 3X_3$, and $Y = 5 + \beta_1^\top X + \beta_2^\top XA + \beta_3 A$. Here, $\theta,\ \beta_1,\ \beta_2 \in \mathbb{R}^{10}$ such that $\theta^\top = \beta_1^\top = \beta_2^\top = \mathbf{1}^{10} := [1, \cdots, 1]^\top$, $\beta_3 = 1$. To induce sparsity, we randomly set three dimensions of the coefficients $\beta_1^\top$ and $\beta_2^\top$ and two dimensions of $\theta^\top$ to zero. For the bandwidth parameter h, we follow the approach of selecting the bandwidth as given by (Kallus and Zhou 2018) using a plug-in estimator based on Eqn. (10). We repeat the data generating process to create 100 different datasets, each consisting of N_{train} ($N_{train} \in 500, 1000, 1500, 2000, 2500$) training samples and $N_{test} = 2000$ test samples.

η^{test}				
0.05	0.1	0.2	0.3	0.4
5.66±12.06	5.60±11.97	5.52±11.85	5.45±11.77	5.40±11.70
5.05±8.68	4.99±8.60	4.91±8.50	4.85±8.44	4.80±8.39
5.48±15.46	5.47±15.46	5.46±15.46	5.45±15.47	5.44±15.47
5.02±10.37	5.01±10.36	5.00±10.34	4.99±10.33	4.98±10.32

Table 2: Comparison of robustness performance (robust v.s. nonrobust) with $\eta^{train} = 0.2$ and $N_{train} = 2000$ for policy learning and various η^{test} for policy evaluation. The reported Mean ± Standard Error (the Standard Error is in %) is computed over 100 runs. The first/second/third/fourth row records values produced by $\hat{Q}^h_{\mathrm{DRO}}(\hat{\pi}^{h,\eta=0.2}_{\mathrm{DRO}})$/$\hat{Q}^h_{\mathrm{DRO}}(\hat{\pi}_{\mathrm{NRO}})$/$\hat{Q}_{pert}(\hat{\pi}^{h,\eta=0.2}_{\mathrm{DRO}})$/$\hat{Q}_{pert}(\hat{\pi}_{\mathrm{NRO}})$.

For policy learning on the training data, both $\hat{\pi}^h_{\mathrm{DRO}}$ and $\hat{\pi}_{\mathrm{NRO}}$ are learned within a linear policy class, and $\hat{\pi}_{\mathrm{DRO}}$ is learned with $\eta^{train} = 0.2$, denoted by $\hat{\pi}^{h,\eta=0.2}_{\mathrm{DRO}}$. For policy evaluation on the test data, in addition to the evaluation metric $\hat{Q}^h_{\mathrm{DRO}}(\pi)$ (Eqn. (7)), we also introduce another metric $\hat{Q}_{pert}(\pi)$ based on a data perturbation strategy. For each of the total 100 original datasets, we perturb each original test dataset $(X_i, A_i, Y_i)_{i=1}^{N_{test}}$ to obtain a new dataset $(\tilde{X}_i, \tilde{A}_i, \tilde{Y}_i)_{i=1}^{N_{test}}$ such that the new dataset lies within a KL-ball centred at the original test dataset with a radius η^{test}, introducing a distribution shift in the new dataset relative to the original test dataset. Then we can evaluate each policy using $\hat{Q}_{pert}(\pi) = \min_{1\leq j\leq 100}\{\frac{1}{N_{\mathrm{test}}}\sum_{i=1}^{N_{test}} \tilde{Y}_i^{(j)}(\pi(\tilde{X}_i^{(j)}))\}$.

The results presented in Table 2 and Table 3 demonstrate that $\hat{\pi}^{h,\eta=0.2}_{\mathrm{DRO}}$ exhibits superior robustness compared to the non-robust policy $\hat{\pi}_{\mathrm{NRO}}$. Specifically, $\hat{\pi}^{h,\eta=0.2}_{\mathrm{DRO}}$ showcases significantly lower sensitivity to data perturbations than $\hat{\pi}_{\mathrm{NRO}}$, consistently achieving higher reward in most cases. Moreover, in Table 2, as the level of data perturbation η^{test} increases from 0.05 to 0.4, $\hat{\pi}^{h,\eta=0.2}_{\mathrm{DRO}}$ shows more stable performance than $\hat{\pi}_{\mathrm{NRO}}$. Notably, in Table 3, even with an increase in training sample size, $\hat{\pi}_{\mathrm{NRO}}$ shows no improvement when faced with a distribution shift η^{test}. In contrast, $\hat{\pi}^{h,\eta=0.2}_{\mathrm{DRO}}$ demonstrates significant improvement as the number of training samples increases.

Empirical Experiment - The Warfarin Case Study

Description. We follow (Kallus and Zhou 2018) to conduct a semi-synthetic study using the Warfarin dataset (Consortium 2009). The dataset contains 5528 patients' medical records, including personal information (e.g., age, gender,

N_{train}				
500	1000	1500	2000	2500
5.19±11.45	5.32±11.55	5.43±14.31	5.48±12.04	5.52±11.85
4.85±8.19	4.79±8.10	4.83±7.95	4.84±7.86	4.91±8.50
4.94±15.64	5.12±16.70	5.19±15.82	5.21±16.05	5.46±15.46
4.95±10.46	4.99±10.16	5.02±10.34	5.00±10.35	5.00±10.34

Table 3: Comparison of robustness performance (robust vs. nonrobust) for various N_{train}. η is chosen as 0.2 for both policy learning and evaluation. The reported Mean $\pm$ Standard Error (the Standard Error is in $\%$) is computed over 100 runs. The first/second/third/fourth row records values due to $\hat{Q}^h_{\text{DRO}}(\hat{\pi}^{h,\eta=0.2}_{\text{DRO}})$/$\hat{Q}^h_{\text{DRO}}(\hat{\pi}_{\text{NRO}})$/$\hat{Q}_{pert}(\hat{\pi}^{h,\eta=0.2}_{\text{DRO}})$/$\hat{Q}_{pert}(\hat{\pi}_{\text{NRO}})$.

race, height, weight), medical problems (e.g., comorbidities and diabetes), medical medication history (e.g., aspirin, atorvastatin, etc.), and their genotypes. The dataset also provides the suggested treatment dose (therapeutic dose).

Setting. We employ a random forest regressor on the therapeutic dose and select 41-dimensional covariates based on the feature importance ranking. There are 3306 samples after dropping those with missing values. The observed dataset is generated as follows: $A|X \sim \mathcal{N}(\theta^\top X + 1, 0.1)$ and $Y = 5 + \beta_1^\top X + \beta_2^\top XA + \epsilon$, where $\beta_1,\ \beta_2,\ \theta \in \mathbb{R}^p$ (with $p = 41$ in our setting), $\beta_1^\top = 0.2 \cdot \mathbf{1}^p$, $\beta_2^\top = 0.1 \cdot \mathbf{1}^p$, $\theta^\top = 0.1 \cdot \mathbf{1}^p$. We also assume $\Pi = \{\zeta^\top X : \|\zeta\|_\infty \leq 2\}$. We again follow the approach of selecting the bandwidth as given by (Kallus and Zhou 2018) using a plug-in estimator based on Eqn. (10). To create distribution shifts, we split the training and test data based on patients' age information. We select 1983 patients aged 10-69 as the training set and 1323 patients older than 70 as the test set. We repeat this process 1000 times to create a total of 1000 semi-synthetic Warfarin datasets.

Results. We learn a non-robust policy $\hat{\pi}_{\text{NRO}}$ and robust policies $\hat{\pi}^{h,\eta}_{\text{DRO}}$ for $\eta \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ on the training set, and we evaluate the six policies on test set based on the sample averaged potential outcome: $\hat{Q}_{mean}(\pi) = \frac{1}{N}\sum_{i=1}^N Y_i(\pi(X_i))$. Consequently, we obtain 1000 values of $\hat{Q}_{\mathcal{M}}(\pi)$ w.r.t. each of the six policies. We then report the mean, standard error, and the $5^{\text{th}}/10^{\text{th}}/15^{\text{th}}/20^{\text{th}}/25^{\text{th}}/30^{\text{th}}$ percentile of the total 1000 values in Table 4.

Table 4 demonstrates four important insights: (1) $\hat{Q}_{mean}(\hat{\pi}^{h,\eta=0.3}_{\text{DRO}})$ exhibits a comparable mean value to $\hat{Q}_{mean}(\hat{\pi}_{\text{NRO}})$. (2) The expected reward initially increases, reaching the optimal (e.g., when $\eta^{train} \in \{0.4, 0.5\}$), and then decreases with larger η. This trend is reasonable, as a very small η neglects the robustness effect and results in a relatively aggressive policy, while a very large η results in an overly conservative policy[6]. (3) The standard error of all robust policies is smaller than that of the non-robust policy.

[6]Determining the optimal η is beyond the scope of this study and is left for future work. Some useful guidances are provided. For example, (Pardo 2018) show that the distance $\mathbb{D}(\cdot|\cdot)$ is asymptotically χ^2 distributed which enables us to select proper η.

		percentile					
Mean	SE	5^{th}	10^{th}	15^{th}	20^{th}	25^{th}	30^{th}
6.377	4.5	4.058	4.525	4.887	5.114	5.350	5.605
6.372	4.3	4.054	4.648	5.020	5.298	5.533	5.703
6.454	4.3	4.244	4.653	5.086	5.306	5.523	5.737
6.409	4.5	4.112	4.552	4.884	5.212	5.449	5.671
6.355	4.5	4.085	4.536	4.774	5.106	5.344	5.548
6.350	4.4	4.092	4.613	4.908	5.240	5.434	5.620

Table 4: Comparison of the rewards of robust and nonrobust policies in Warfarin study. The reported result are computed over 1000 runs. Note that the SE in the table represents the Standard Error which is reported in $\%$. The first/second/third/fourth/fifth/sixth row records values produced by $\hat{Q}_{mean}(\hat{\pi}_{\text{NRO}})$/$\hat{Q}_{mean}(\hat{\pi}^{h,\eta=0.3}_{\text{DRO}})$/$\hat{Q}_{mean}(\hat{\pi}^{h,\eta=0.4}_{\text{DRO}})$/$\hat{Q}_{mean}(\hat{\pi}^{h,\eta=0.5}_{\text{DRO}})$/$\hat{Q}_{mean}(\hat{\pi}^{h,\eta=0.6}_{\text{DRO}})$/$\hat{Q}_{mean}(\hat{\pi}^{h,\eta=0.7}_{\text{DRO}})$.

(4) From the percentile results, most robust policies outperform the non-robust policy in "bad" scenarios, underscoring the robustness of the proposed $\hat{\pi}^h_{\text{DRO}}$.

Conclusion, Limitation and Future Work

Conclusion. We investigate offline policy evaluation and learning under continuous treatment in the distributionally robust optimization (DRO) setting. We propose an estimator, $\hat{Q}^h_{\text{DRO}}(\pi)$, for offline policy evaluation and obtain a distributionally robust policy, $\hat{\pi}^h_{\text{DRO}}$, based on $\hat{Q}^h_{\text{DRO}}(\pi)$. We study the asymptotic distribution and the statistical guarantee of $\hat{Q}^h_{\text{DRO}}(\pi)$ and $\hat{\pi}^h_{\text{DRO}}$. Experimental results demonstrate the superior performance of our approach.

Future Work and Limitations. The proposed framework can be applied in various fields where distribution shifts occur in the context of continuous-valued treatments. For instance, doctors may seek to determine a robust dosage that minimizes potential disease risks for target patients, while policymakers might aim to establish a robust credit-increasing strategy that maximizes potential consumption for target customers. Thus, applying our framework to real-world scenarios represents a significant next step. Additionally, several potential technical investigations can be further explored. First, selecting the divergence measures and determining the ambiguity radius for the distributional ambiguity set pose significant challenges in both the Operations Research and Machine Learning communities. Future research would benefit from establishing statistical guarantees for other metrics (e.g., Wasserstein metric) and offering guidance on setting the radius for policy evaluation and learning. Second, exploring methods for the generalized propensity score when it is unknown would be interesting. Third, expanding our framework to include the doubly robust estimator might improve the convergence rate of policy learning. Lastly, strictly limiting the policy class to linear functions may fail to capture complex relationships between covariates and treatment, leading to suboptimal results. Considering broader policy classes (e.g., nonlinear policy classes with infinite VC dimensions) is therefore essential.

Acknowledgements

Qi WU acknowledges the support from The CityU-JD Digits Joint Laboratory in Financial Technology and Engineering and The Hong Kong Research Grants Council [General Research Fund 11219420/9043008]. The work described in this paper was partially supported by the InnoHK initiative, the Government of the HKSAR, and the Laboratory for AI-Powered Financial Technologies.

References

Anthony, M.; Bartlett, P. L.; Bartlett, P. L.; et al. 1999. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge.

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

Bahadori, T.; Tchetgen, E. T.; and Heckerman, D. 2022. End-to-end balancing for causal continuous treatment-effect estimation. In *International Conference on Machine Learning*, 1313–1326. PMLR.

Bartlett, P. L.; Foster, D. J.; and Telgarsky, M. J. 2017. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30.

Bica, I.; Jordon, J.; and van der Schaar, M. 2020. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33: 16434–16445.

Chen, Z.; Sim, M.; and Xu, H. 2019. Distributionally robust optimization with infinitely constrained ambiguity sets. *Operations Research*, 67(5): 1328–1344.

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

Colangelo, K.; and Lee, Y.-Y. 2019. Double debiased machine learning nonparametric inference with continuous treatments. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Consortium, I. W. P. 2009. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8): 753–764.

Delage, E.; and Ye, Y. 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3): 595–612.

Dudík, M.; Langford, J.; and Li, L. 2011. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, 1097–1104. Madison, WI, USA: Omnipress. ISBN 9781450306195.

Faury, L.; Tanielian, U.; Dohmatob, E.; Smirnova, E.; and Vasile, F. 2020. Distributionally Robust Counterfactual Risk Minimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 3850–3857.

Gao, R.; Chen, X.; and Kleywegt, A. J. 2022. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*.

Horvitz, D. G.; and Thompson, D. J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260): 663–685.

Hu, Z.; and Hong, L. J. 2013. Kullback-Leibler divergence constrained distributionally robust optimization. *Optimization Online*, 1(2): 9.

Huang, Y.; Leung, C. H.; Ma, S.; Yuan, Z.; Wu, Q.; Wang, S.; Wang, D.; and Huang, Z. 2023. Towards Balanced Representation Learning for Credit Policy Evaluation. In *International Conference on Artificial Intelligence and Statistics*, 3677–3692. PMLR.

Huang, Y.; Leung, C. H.; Wu, Q.; Yan, X.; Ma, S.; Yuan, Z.; Wang, D.; and Huang, Z. 2022. Robust causal learning for the estimation of average treatment effects. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–9. IEEE.

Huang, Y.; Leung, C. H.; Yan, X.; Wu, Q.; Peng, N.; Wang, D.; and Huang, Z. 2021. The Causal Learning of Retail Delinquency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1): 204–212.

Huang, Y.; Siyi, W.; Leung, C. H.; Qi, W.; Dongdong, W.; and Huang, Z. 2024. Dignet: Learning decomposed patterns in representation balancing for treatment effect estimation. *Transactions on Machine Learning Research*.

Husain, H.; Nguyen, V.; and van den Hengel, A. 2023. Distributionally robust Bayesian optimization with ϕ-divergences.

Imbens, G. W. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1): 4–29.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Kallus, N. 2018. Balanced policy evaluation and learning. *Advances in neural information processing systems*, 31.

Kallus, N.; Mao, X.; Wang, K.; and Zhou, Z. 2022. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, 10598–10632. PMLR.

Kallus, N.; and Zhou, A. 2018. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, 1243–1251. PMLR.

Khan, S.; and Ugander, J. 2023. Adaptive normalization for IPW estimation. *Journal of Causal Inference*, 11(1): 20220019.

Kitagawa, T.; and Tetenov, A. 2018. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2): 591–616.

Kuhn, D.; Esfahani, P. M.; Nguyen, V. A.; and Shafieezadeh-Abadeh, S. 2019. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, 130–166. Informs.

Kullback, S. 1959. *Information Theory and Statistics*. New York: Wiley.

Kullback, S.; and Leibler, R. A. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, 661–670.

Li, Y.; Leung, C. H.; Sun, X.; Wang, C.; Huang, Y.; Yan, X.; Wu, Q.; Wang, D.; and Huang, Z. 2024. The Causal Impact of Credit Lines on Spending Distributions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1): 180–187.

Mo, W.; Qi, Z.; and Liu, Y. 2021. Learning optimal distributionally robust individualized treatment rules. *Journal of the American Statistical Association*, 116(534): 659–674.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2018. *Foundations of machine learning*. MIT press.

Pardo, L. 2018. *Statistical inference based on divergence measures*. Chapman and Hall/CRC.

Qin, R.-J.; Zhang, X.; Gao, S.; Chen, X.-H.; Li, Z.; Zhang, W.; and Yu, Y. 2022. NeoRL: A near real-world benchmark for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 24753–24765.

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

Schwab, P.; Linhardt, L.; Bauer, S.; Buhmann, J. M.; and Karlen, W. 2020. Learning Counterfactual Representations for Estimating Individual Dose-Response Curves. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 5612–5619.

Shalev-Shwartz, S.; and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shen, Y.; Xu, P.; and Zavlanos, M. 2024. Wasserstein Distributionally Robust Policy Evaluation and Learning for Contextual Bandits. *Transactions on Machine Learning Research*. Featured Certification.

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

Su, L.; Ura, T.; and Zhang, Y. 2019. Non-separable models with high-dimensional data. *Journal of Econometrics*, 212(2): 646–677.

Swaminathan, A.; and Joachims, T. 2015. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28.

Tang, S.; and Wiens, J. 2021. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, 2–35. PMLR.

Wainwright, M. J. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wooldridge, J. M. 2007. Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, 141(2): 1281–1301.

Yang, Z.; Guo, Y.; Xu, P.; Liu, A.; and Anandkumar, A. 2023. Distributionally robust policy gradient for offline contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, 6443–6462. PMLR.

Zhang, T. 2002. Covering Number Bounds of Certain Regularized Linear Function Classes. *J. Mach. Learn. Res.*, 2: 527–550.

Zhao, Y.; Zeng, D.; Rush, A. J.; and Kosorok, M. R. 2012. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499): 1106–1118.

Zhou, X.; Mayer-Hamblett, N.; Khan, U.; and Kosorok, M. R. 2017. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517): 169–187.

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

Zymler, S.; Kuhn, D.; and Rustem, B. 2013. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137: 167–198.