Distributionally Robust Policy Evaluation and Learning in Offline Contextual Bandits

Nian Si[*1] **Fan Zhang**[*1] **Zhengyuan Zhou**[2] **Jose Blanchet**[1]

Abstract

Policy learning using historical observational data is an important problem that has found widespread applications. However, existing literature rests on the crucial assumption that the future environment where the learned policy will be deployed is the same as the past environment that has generated the data–an assumption that is often false or too coarse an approximation. In this paper, we lift this assumption and aim to learn a distributionally robust policy with bandit observational data. We propose a novel learning algorithm that is able to learn a robust policy to adversarial perturbations and unknown covariate shifts. We first present a policy evaluation procedure in the ambiguous environment and also give a heuristic algorithm to solve the distributionally robust policy learning problems efficiently. Additionally, we provide extensive simulations to demonstrate the robustness of our policy.

1. Introduction

The past decade has witnessed an explosion of user-specific data across a variety of application domains: electronic medical data in health care, marketing data in product recommendation and customer purchase/selection data in digital advertising (Bertsimas & Mersereau, 2007; Li et al., 2010; Chapelle, 2014; Bastani & Bayati, 2015; Schwartz et al., 2017). Such growing availability of user-specific data has ushered in an exciting era of personalized decision making, one that allows the decision maker(s) to personalize the service decisions based on each individual's distinct features. The key value added by personalized decision making is that heterogeneity across individuals, a ubiquitous phenomenon in these applications, can be intelligently exploited to achieve better outcomes - because best recommendation decisions vary across different individuals.

Rising to this opportunity, contextual bandits have emerged to be the predominant mathematical framework that is at once elegant and powerful: its three components, the contexts (representing individual characteristics), the actions (representing the recommended items), and the rewards (representing the outcomes), capture the salient aspects of the problem and provide fertile ground for developing algorithms that contribute to making quality decisions. In particular, within the broad landscape of contextual bandits, the offline[1] contextual bandits literature has precisely aimed to answer the following questions that lie at the heart of data-driven decision making: *given a historical collection of past data that consists of the three components as mentioned above, how can a new policy (mapping from contexts to actions) be evaluated accurately, and one step further, how can an effective policy be learned efficiently?*

Such questions–both policy evaluation and policy learning using historical data–have motivated a flourishing and rapidly developing line of recent work (Dudík et al., 2011; Zhang et al., 2012; Zhao et al., 2012; 2014; Swaminathan & Joachims, 2015; Vitus et al., 2015; Rakhlin & Sridharan, 2016; Kallus, 2018; Dimakopoulou et al., 2018; Zhou et al., 2017b; Jiang et al., 2018; Kitagawa & Tetenov, 2018; Kallus & Zhou, 2018; Zhou et al., 2018; Joachims et al., 2018; Chernozhukov et al., 2019) that contributed valuable insights: novel policy evaluation and policy learning algorithms have been developed; sharp minimax regret guarantees have been characterized (through a series of efforts) in many different

[*]Equal contribution [1]Department of Management Science & Engineering, Stanford University [2]IBM research and Stern School of Business, New York University. Correspondence to: Nian Si <niansi@stanford.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]Correspondingly, there has also been an extensive literature on online contextual bandits (Li et al., 2010; Rusmevichientong & Tsitsiklis, 2010; Filippi et al., 2010; Rigollet & Zeevi, 2010; Chu et al., 2011; Goldenshluger & Zeevi, 2013; Agrawal & Goyal, 2013a;b; Jun et al., 2017; Li et al., 2017; Abeille et al., 2017; Li et al., 2017; Master et al., 2017; Dimakopoulou et al., 2019; Bistritz et al., 2019; Zhou et al., 2019), whose focus is to develop online adaptive algorithms that effectively balance exploration and exploitation. This is not the focus of our paper and we simply mention them in passing here. See (Bubeck et al., 2012; Lattimore & Szepesvári, 2018; Slivkins et al., 2019) for a few articulate expositions.

settings; extensive and illuminating experimental results have been performed to offer practical advice for optimizing empirical performance.

However, a key assumption underlying the existing offline contextual bandits work mentioned above is that the future environment in which the learned policy is deployed stays the same as the past environment from which the historical data is collected (and the to-be-deployed policy is trained). In practice, such an assumption rarely holds and there are two primary sources of such "environment change":

1. **Covariate shift:** The individuals–and hence their characteristics–in a population can change, thereby resulting in a different distribution of the contexts. For instance, an original population with more young people can shift to a population with more senior people.

2. **Concept drift:** How the rewards depend on the underlying contexts and actions can also change, thereby resulting in different conditional distributions of the rewards given the contexts and the actions. For instance, individuals' preferences over products can shift over time and sometimes exhibit seasonal patterns.

As a consequence, these offline contextual bandit algorithms are **fragile**: should the future environment change, the deployed policy–having not taken into account the possible environment changes in the future–will perform poorly. This naturally leads to the following fundamental question: *Can we learn a **robust** policy that performs well in the presence of either (or both) of the above environment shifts?* Our goal in this paper is to provide a framework for thinking about this question and providing an affirmative answer therein.

1.1. Our Contributions and Related Work

Our contributions are threefold. First, we propose a distributionally robust formulation of policy evaluation and learning in offline contextual bandits, that accommodates both types of environment shifts mentioned above. Our formulation postulates that the future environment–characterized by a joint distribution on the context and all the rewards when taking different actions–is in a Kullback-Leibler neighborhood around the training environment's distribution, thereby allowing for learning a robust policy from training data that is not sensitive to the future environment being the same as the past. Despite the fact that there has been a growing literature (Bertsimas & Sim, 2004; Delage & Ye, 2010; Hu & Hong, 2013; Shafieezadeh-Abadeh et al., 2015; Bayraksan & Love, 2015; Gao & Kleywegt, 2016; Duchi et al., 2016; Staib & Jegelka, 2017; Shapiro, 2017; Lam & Zhou, 2017; Chen et al., 2018; Sinha et al., 2018; Lee & Raginsky, 2018; Nguyen et al., 2018; Yang, 2018; Mohajerin Esfahani & Kuhn, 2018; Abadeh et al., 2018; Zhao & Guan, 2018; Sinha et al., 2018; Gao et al., 2018; Chen et al., 2018; Ghosh & Lam, 2019; Blanchet & Murthy, 2019) on distributionally robust optimization (DRO)–one that shares the same philosophical underpinning on distributionally robustness as ours–the existing DRO literature has mostly focused on the statistical learning aspects, including supervised learning and feature selection type problems, rather than the decision making aspects. The idea of using DRO to solve "environment change" problems is not new (Duchi & Namkoong, 2018; Duchi et al., 2019). However, Applying distributionally robust formulation under bandit feedback is not trivial. We notice that (Faury et al., 2019) considers a similar setting. However, the policy value estimator proposed in their paper is inconsistent in the situation that the future environment is not the same as the past environment. To the best of our knowledge, we provide the first distributionally robust formulation for policy evaluation and learning under bandit feedback in the "environment change" setting.

Second, we provide a novel scheme for distributionally robust policy evaluation (Algorithm 1) that estimates the robust value of any given policy using historical data. We do so by drawing from duality theory and transforming the primal robust value estimation problem–an infinitely-dimensional problem–into a dual problem that is 1-dimensional and convex, hence admitting an efficiently computable solution. Further, we study the efficiency of this distributionally robust estimator and establish, in the form of a central limit theorem, that the proposed estimator converges to the true value at an $O_p\left(n^{-1/2}\right)$ rate (n is the number of data points).

Third, we build upon this distributionally robust policy evaluation scheme and propose a distributionally robust learning scheme, which we call Stable Distributionally Robust Policy Learning (Algorithm 3). In our distributionally robust policy learning setting, naively picking a policy that maximizes the estimated distributionally robust value (Algorithm 2) can be extremely unstable and yields poor performance. We thus provide a stabilized version of the distributionally robust estimator that, despite having similar performance when evaluating a single, fixed policy as the non-stabilized version, achieves much better performance in the learning phase. We provide extensive experimental results to demonstrate the efficiency and effectiveness of the proposed distributionally robust policy evaluation and learning schemes.

2. A Distributionally Robust Formulation

Let $\mathcal{A}$ be the set of d actions: $\mathcal{A} = \{a^1, a^2, \ldots, a^d\}$ and let $\mathcal{X}$ be the set of contexts (typically a subset of $\mathbf{R}^p$). Following the standard contextual bandits model, we posit the existence of a fixed underlying data-generating distribution on $(X, Y(a^1), Y(a^2), \ldots, Y(a^d)) \in \mathcal{X} \times \prod_{j=1}^{d} \mathcal{Y}_j$, where $X \in \mathcal{X}$ denotes the context vector, and each $Y(a^j) \in \mathcal{Y}_j \subset \mathbb{R}$ denotes the random reward obtained when action a^j is

selected under context X.

Let $\{(X_i, A_i, Y_i)\}_{i=1}^n$ be n independent and identically distributed (i.i.d.) observed triples that comprise of the training data, where $(X_i, Y_i(a^1), \ldots, Y_i(a^d))$ are drawn i.i.d. from the fixed underlying distribution described above, and we denote this underlying distribution by $\mathbf{P}_0$. Further, for the i-th data point (X_i, A_i, Y_i), A_i denotes the action selected and $Y_i = Y_i(A_i)$. In other words, Y_i in the i-th data point is the observed reward under the context X_i and action A_i.

We assume the actions in the training data are selected by some fixed underlying policy π_0 that is known to the decision-maker, where $\pi_0(a \mid x)$ gives the probability of selecting action a when the context is x. In other words, for each context X_i, a random action A_i is selected according to the distribution $\pi_0(\cdot \mid X_i)$, after which the reward $Y_i(A_i)$ is then observed. Finally, we use $\mathbf{P}_0 * \pi_0$ to denote the product distribution of $(X, Y(a^1), Y(a^2), \ldots, Y(a^d), A)$ on space $\mathcal{X} \times \prod_{j=1}^d \mathcal{Y}_j \times \mathcal{A}$. We make the following assumptions on the data-generating process.

Assumption 1. The joint distribution $(X, Y(a^1), Y(a^2), \ldots, Y(a^d), A)$ satisfies:

1. Unconfoundedness: $((Y(a^1), Y(a^2), \ldots, Y(a^d))$ is independent with A conditional on X, i.e.,
$$((Y(a^1), Y(a^2), \ldots, Y(a^d)) \perp\!\!\!\perp A | X.$$
2. Overlapping: There exists some $\eta > 0$, $\pi_0(a \mid x) \geq \eta$, $\forall (x, a) \in \mathcal{X} \times \mathcal{A}$.
3. Bounded reward support: $0 \leq Y(a^i) \leq M$ for $i = 1, 2, \ldots, d$.
4. Positive density: For any $i = 1, 2, \ldots, d$, $Y(a^i)|X$ has conditional density $f_i(y_i|x)$ and $f_i(y_i|x)$ has a uniform non-zero lower bound i.e., $f_i(y_i|x) \geq \underline{b} > 0$ over the interval $[0, M]$ with $2\underline{b}M \leq 1$.

The overlap assumption ensures that some minimum positive probability is guaranteed no matter what the context is. This ensures sufficient exploration in collecting the training data. The above assumptions are standard and commonly adopted in both the estimation literature (Rosenbaum & Rubin, 1983; Imbens, 2004; Imbens & Rubin, 2015) and the policy learning literature (Zhang et al., 2012; Zhao et al., 2012; Swaminathan & Joachims, 2015; Zhou et al., 2017a; Kitagawa & Tetenov, 2018).

2.1. Standard Policy Learning

In the standard contextual bandits terminology, $\mu_a(x) = \mathbf{E}_{\mathbf{P}_0}[Y(a) \mid X = x]$ is known as the mean reward function (for action a). Depending on whether one assumes a parametric form of $\mu_a(x)$ or not, one needs to employ different statistical methodologies. In particular, when $\mu_a(x)$ is a linear function of x, this is known as linear contextual bandits, an important and most extensively studied subclass of contextual bandits. In this paper, we do not make any structural assumption on $\mu_a(x)$: we are in the non-parametric contextual bandits regime and work with a general underlying data-generating distributions $\mathbf{P}_0$.

With the above setup, the goal is to learn a good policy from a fixed deterministic policy class Π using the training data. This is often known as the batch contextual bandits problem (in contrast to online contextual bandits), because all the data has already been collected at once before the decision maker aims to learn a policy. A policy $\pi : \mathcal{X} \to \mathcal{A}$ is a function that maps a context vector x to an action and the performance of π is measured by the expected reward this policy generates, as characterized by the policy value function:

Definition 2.1. The policy value function $Q : \Pi \to \mathbf{R}$ is defined as: $Q(\pi) = \mathbf{E}_{\mathbf{P}_0}[Y(\pi(X))]$, where the expectation is taken with respect to the randomness in the underlying joint distribution $\mathbf{P}_0$ of $(X, Y(a^1), Y(a^2), \ldots, Y(a^d))$.

With this definition, the optimal policy is a policy that maximizes the policy value function. The objective in the standard policy learning context is to learn a policy π that has the policy value as large as possible.

2.2. Distributionally Robust Policy Learning

Using the policy value function $Q(\cdot)$ (as defined in Definition 2.1) to measure the quality of a policy brings out an implicit assumption that the decision maker is imposing: the environment that generated the training data is the same as the environment where the policy will be deployed. This is manifested in that the expectation in $Q(\cdot)$ is taken with respect to the same underlying distribution $\mathbf{P}_0$. However, the underlying data-generating distribution may be different for the training environment and the test environment. In such cases, the policy learned with the goal to maximize the value under $\mathbf{P}_0$ may not work well under the new test environment.

To address this issue, we propose a distributionally robust formulation for policy learning, where we explicitly incorporate into the learning phase the consideration that the test distribution may not be the same as the training distribution $\mathbf{P}_0$. To that end, we start by introducing some terminology. First, the Kullback -Leibler (KL) divergence between two probability measures $\mathbf{P}$ and $\mathbf{P}_0$, denoted by $D(\mathbf{P}||\mathbf{P}_0)$, is defined as $D(\mathbf{P}||\mathbf{P}_0) = \int \log\left(\frac{d\mathbf{P}}{d\mathbf{P}_0}\right) d\mathbf{P}$. With KL-divergence, we can define a class of neighborhood distributions around a given distribution. Specifically, the distributionally uncertainty set $\mathcal{U}_{\mathbf{P}_0}(\delta)$ of size δ is defined as $\mathcal{U}_{\mathbf{P}_0}(\delta) = \{\mathbf{P} \ll \mathbf{P}_0 \mid D(\mathbf{P}||\mathbf{P}_0) \leq \delta\}$. When it is

clear from the context what the uncertainty radius δ is, we sometimes suppress δ for notational simplicity and write $\mathcal{U}_{\mathbf{P}_0}$ instead.

Definition 2.2. For a given $\delta > 0$, the distributionally robust value function $Q_{\mathrm{DRO}} : \Pi \to \mathbf{R}$ is defined as: $Q_{\mathrm{DRO}}(\pi) = \inf_{\mathbf{P}\in\mathcal{U}_{\mathbf{P}_0(\delta)}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]$.

In other words, $Q_{\mathrm{DRO}}(\pi)$ measures the performance of a policy π by evaluating how it performs in the worst possible environment among the set of all environments that are δ-close to $\mathbf{P}_0$. To be robust to the changes between the test environment and the training environment, our goal is to learn a policy such that its distributionally robust policy value is as large as possible.

3. Distributionally Robust Policy Evaluation

In order to learn a distributionally robust policy–one that maximizes $Q_{\mathrm{DRO}}(\pi)$–a key step lies in accurately estimating the given policy π's distributionally robust value. We devote this section to tackling this problem.

3.1. Algorithm

In this subsection, we provide an algorithm for evaluating $\inf_{\mathbf{P}\in\mathcal{U}_{\mathbf{P}_0(\delta)}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]$, the distributionally robust policy value for a given policy $\pi \in \Pi$. Our algorithm hinges on the strong duality associated with the distributional robust policy value, as formally characterized next.

Lemma 3.1 (Strong Duality)**.** *For any policy $\pi \in \Pi$, we have*

$$\begin{aligned}
&\inf_{\mathbf{P}\in\mathcal{U}_{\mathbf{P}_0(\delta)}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))] \\
&= \sup_{\alpha\geq 0}\{-\alpha\log\mathbf{E}_{\mathbf{P}_0}[\exp(-Y(\pi(X))/\alpha)] - \alpha\delta\} \\
&= \sup_{\alpha\geq 0}\left\{-\alpha\log\mathbf{E}_{\mathbf{P}_0*\pi_0}\left[\frac{\exp(-Y(A)/\alpha)\mathbf{1}\{\pi(X)=A\}}{\pi_0(A\mid X)}\right]\right. \\
&\quad \left.-\alpha\delta\right\}. \qquad (3.1)
\end{aligned}$$

Proof. The first equality follows from Theorem 1 of (Hu & Hong, 2013). The second equality holds, because for any (Borel measurable) function $f : \mathbf{R} \to \mathbf{R}$ and any policy $\pi \in \Pi$, we have

$$\begin{aligned}
&\mathbf{E}_{\mathbf{P}}[f(Y(\pi(X)))] \\
=\ &\mathbf{E}_{\mathbf{P}*\pi_0}\left[\frac{f(Y(\pi(X)))\mathbf{1}\{\pi(X)=A\}}{\pi_0(A\mid X)}\right] \\
=\ &\mathbf{E}_{\mathbf{P}*\pi_0}\left[\frac{f(Y(A))\mathbf{1}\{\pi(X)=A\}}{\pi_0(A\mid X)}\right].
\end{aligned}$$

Plugging in $f(x) = \exp(-\frac{x}{\alpha})$ yields the result. □

Remark 3.1. When $\alpha = 0$, by Proposition 2 of (Hu & Hong, 2013), we can define

$$-\alpha\log\mathbf{E}_{\mathbf{P}_0}[\exp(-Y(\pi(X))/\alpha)] - \alpha\delta|_{\alpha=0} = \operatorname{ess\,inf}\{Y\},$$

where ess inf denotes the essential infimum. Furthermore, $-\alpha\log\mathbf{E}_{\mathbf{P}_0}[\exp(-Y(\pi(X))/\alpha)] - \alpha\delta$ is right continuous at zero. In fact, in Lemma A3 in Appendix A.3, we show that the optimal value is not attained at $\alpha = 0$.

The above strong duality allows us to transform the original problem of evaluating $\inf_{\mathbf{P}\in\mathcal{U}_{\mathbf{P}_0(\delta)}} \mathbf{E}_{\mathbf{P}}[Y(\pi(X))]$, where the (primal) variable is a distribution $\mathbf{P}$ into a simpler problem, where the (dual) variable is a positive scalar α. Note that in the dual problem, the expectation is taken with respect to the same underlying distribution $\mathbf{P}_0$. This then allows us to use an easily-computable plug-in estimator of the distributionally robust policy value, $\hat{Q}_{\mathrm{DRO}}(\pi)$ to approximate $Q_{\mathrm{DRO}}(\pi)$.

Finally, for ease of reference in the subsequent analysis of Algorithm 1, we capture the important terms in the following definition.

Definition 3.2. Let $\{(X_i, A_i, Y_i)\}_{i=1}^n$ be a given dataset. Define $W_i(\pi,\alpha) = \frac{\mathbf{1}\{\pi(X_i)=A_i\}}{\pi_0(A_i|X_i)}\exp(-Y_i(A_i)/\alpha)$ and $\hat{W}_n(\pi,\alpha) = \frac{1}{n}\sum_{i=1}^n W_i(\pi,\alpha)$. We also define the dual objective function and the empirical dual objective function as

$$\phi(\pi,\alpha) = -\alpha\log\mathbf{E}_{\mathbf{P}_0}[\exp(-Y(\pi(X))/\alpha)] - \alpha\delta,$$

and

$$\hat{\phi}_n(\pi,\alpha) = -\alpha\log\hat{W}_n(\pi,\alpha) - \alpha\delta,$$

respectively.

Then, we define the distributionally robust value estimators and the optimal dual variable using the notation above.

1. The distributionally robust value estimator $\hat{Q}_{\mathrm{DRO}} : \Pi \to \mathbf{R}$ is defined by $\hat{Q}_{\mathrm{DRO}}(\pi) = \sup_{\alpha\geq 0}\left\{\hat{\phi}_n(\pi,\alpha)\right\}$.
2. The optimal dual variable α^* is defined by $\alpha^* = \arg\max_{\alpha\geq 0}\{\phi(\pi,\alpha)\}$.

The upper and lower bound of α^* in Appendix A.3 establish the validity of the definitions α^*, namely, α^* is attainable.

In the last step of Algorithm 1, one needs to solve an optimization problem to obtain the distributionally robust estimate of the policy π. As the following indicates, this optimization problem is easy to solve.

Lemma 3.3. *The empirical dual objective function $\hat{\phi}_n(\pi,\alpha)$ is concave in α and its partial derivative admits*

the expression

$$\begin{aligned}\frac{\partial}{\partial\alpha}\hat{\phi}_n(\pi,\alpha) &= -\frac{\sum_{i=1}^n Y_i(A_i)W_i(\pi,\alpha)}{\alpha n\hat{W}_n(\pi,\alpha)} \\ &\quad -\log\hat{W}_n(\pi,\alpha)-\delta, \\ \frac{\partial^2}{\partial\alpha^2}\hat{\phi}_n(\pi,\alpha) &= \frac{(\sum_{i=1}^n Y_i(A_i)W_i(\pi,\alpha))^2}{\alpha^3 n^2(\hat{W}_n(\pi,\alpha))^2} \\ &\quad -\frac{\sum_{i=1}^n Y_i^2(A_i)W_i(\pi,\alpha)}{\alpha^3 n\hat{W}_n(\pi,\alpha)}.\end{aligned}$$

Further, if the array $\{Y_i(A_i)\mathbf{1}\{\pi(X_i)=A_i\}\}_{i=1}^n$ *has at least two different non-zero entries, then* $\hat{\phi}_n(\pi,\alpha)$ *is strictly-concave in* α.

The proof of Lemma 3.3 is in Appendix A.2.

Since the optimization problem $\max_{\alpha\geq 0}\left\{\hat{\phi}_n(\pi,\alpha)\right\}$ is maximizing a concave function, it can be computed by the Newton–Raphson method. We make it precise in Algorithm 1.

Algorithm 1 Distributionally Robust Policy Evaluation

1: **Input:** Dataset $\{(X_i,A_i,Y_i)\}_{i=1}^n$, data-collecting policy π_0, policy $\pi\in\Pi$, and initial value of dual variable α.
2: **Output:** Estimator of the distributionally robust policy value $\hat{Q}_{\mathrm{DRO}}(\pi)$.
3: **repeat**
4: Let $W_i(\pi,\alpha)\leftarrow\frac{\mathbf{1}\{\pi(X_i)=A_i\}}{\pi_0(A_i|X_i)}\exp(-Y_i(A_i)/\alpha)$.
5: Compute $\hat{W}_n(\pi,\alpha)\leftarrow\frac{1}{n}\sum_{i=1}^n W_i(\pi,\alpha)$.
6: Update $\alpha\leftarrow\alpha-(\frac{\partial}{\partial\alpha}\hat{\phi}_n)/(\frac{\partial^2}{\partial\alpha^2}\hat{\phi}_n)$.
7: **until** α converge.
8: **Return** $\hat{Q}_{\mathrm{DRO}}(\pi)\leftarrow\hat{\phi}_n(\pi,\alpha)$.

3.2. Theoretical Guarantees of Distributionally Robust Policy Evaluation

In the next theorem, we will demonstrate that the approximation error for policy evaluation function $\hat{Q}_{\mathrm{DRO}}(\pi)$ is $O_p(n^{-1/2})$ for a fixed policy π.

Theorem 3.4. *Let* $\sigma^2(\alpha)=\alpha^2\frac{\mathbf{Var}[W_i(\pi,\alpha)]}{(\mathbf{E}[W_i(\pi,\alpha)])^2}$. *Suppose Assumption 1 is enforced, for any policy* $\pi\in\Pi$, *we have*

$$\sqrt{n}\left(\hat{Q}_{\mathrm{DRO}}(\pi)-Q_{\mathrm{DRO}}(\pi)\right)\Rightarrow N\left(0,\sigma^2(\alpha^*)\right), \quad (3.2)$$

where α^* *is defined in Definition 3.2 and* $\Rightarrow$ *denotes convergence in distribution.*

Sketch of Proof. The proof is based on the functional central limit theorem and the Delta theorem.

We first show for every $\alpha\in(0,\infty)$,

$$\sqrt{n}\left(\hat{W}_n(\pi,\alpha)-\mathbf{E}[W_i(\pi,\alpha)]\right)\Rightarrow Z(\alpha),$$

where $Z(\alpha)=N\left(0,\mathbf{Var}\left[W_i(\pi,\alpha)\right]\right)$. Let $\mathcal{C}([\underline{\alpha}/2,2\overline{\alpha}])$ be the space of continuous functions supported on $[\underline{\alpha}/2,2\overline{\alpha}]$, equipped with the supremum norm, where $0<\underline{\alpha}\leq\alpha^*\leq\overline{\alpha}<\infty$. Then, by the functional central limit theorem (see, for example, Corollary 7.17 in (Araujo & Giné, 1980)), we have

$$\sqrt{n}\left(\hat{W}_n(\pi,\cdot)-\mathbf{E}[W_i(\pi,\cdot)]\right)\Rightarrow Z\left(\cdot\right),$$

in the Banach space $\mathcal{C}([\underline{\alpha}/2,2\overline{\alpha}])$.

We next apply the Delta theorem (Theorem 7.59 in (Shapiro et al., 2009)), which generalizes the Delta method to the infinity-dimensional space, and obtain that

$$\sqrt{n}\left(V(\hat{W}_n(\pi,\cdot))-V(\mathbf{E}[W_i(\pi,\cdot)])\right)\Rightarrow V'_{\mathbf{E}[W_i(\pi,\cdot)]}\left(Z\right),$$

where V is a functional defined by

$$V(\psi)=\inf_{\alpha\in[\underline{\alpha}/2,2\overline{\alpha}]}\alpha\log\left(\psi(\alpha)\right)+\alpha\delta,$$

and $V'_\mu(\nu)$ is the directional derivative of $V(\cdot)$ at μ in the direction of ν, which is computable by the Danskin theorem (Theorem 4.13 in (Bonnans & Shapiro, 2000)). Finally, we show $\mathbf{P}(\hat{Q}_{\mathrm{DRO}}(\pi)\neq -V(\hat{W}_n(\pi,\alpha)))\to 0$ as $n\to\infty$, and complete the proof by Slutsky's lemma (Theorem 1.8.10 in (Lehmann & Casella, 2006)). □

The detailed proof of Theorem 3.4 is in Appendix A.3. Furthermore, Lemma 3.5 gives an approximation of the value of the optimal dual variable α^*, when δ is small.

Lemma 3.5. *When* δ *is closed to zero, we have*

$$\alpha^*=\sqrt{\mathbf{Var}(Y(\pi(X))/\delta)}+o(1/\sqrt{\delta}).$$

The proof of Lemma 3.5 is in Appendix A.4. (Faury et al., 2019) also considers a similar problem. Their scheme amounts to perturbing the data-collecting policy π_0 as well. The estimator they proposed, which is equivalent to (in our notation)

$$\sup_{\alpha>0}\left\{-\alpha\log\left(\frac{1}{n}\sum_{i=1}^n\exp\left(-\frac{\mathbf{1}\{\pi(X_i)=A_i\}Y_i(A_i)}{\alpha\pi_0(A_i\mid X_i)}\right)\right)-\alpha\delta\right\},$$

however, is not a consistent estimator in our setting.

4. Distributionally Robust Policy Learning

We now harness the distributionally robust policy evaluation scheme to design policy learning algorithms.

4.1. Algorithm

In this section, we provide an algorithm for computing the distributionally robust optimal policy within a given policy class Π. We denote by $\hat{\pi}_{\rm DRO}$ the optimal policy that maximizes the value of $\hat{Q}_{\rm DRO}$, i.e.,

$$\begin{aligned}\hat{\pi}_{\rm DRO} &= \arg\max_{\pi\in\Pi} \hat{Q}_{\rm DRO}(\pi) \qquad (4.1)\\ &= \arg\max_{\pi\in\Pi} \sup_{\alpha\geq 0}\left\{-\alpha\log\hat{W}_n(\pi,\alpha)-\alpha\delta\right\}.\end{aligned}$$

In Algorithm 2, we provide a numerical scheme for learning $\hat{\pi}_{\rm DRO}$ by alternatively updating π and α.

Algorithm 2 Distributionally Robust Policy Learning

1: **Input:** Dataset $\{(X_i,A_i,Y_i)\}_{i=1}^n$, data-collecting policy π_0, and initial value of dual variable α.
2: **Output:** Distributionally robust optimal policy $\hat{\pi}_{\rm DRO}$.
3: **repeat**
4: Let $W_i(\pi,\alpha)\leftarrow \frac{\mathbf{1}\{\pi(X_i)=A_i\}}{\pi_0(A_i|X_i)}\exp(-Y_i(A_i)/\alpha)$.
5: Form $\hat{W}_n(\pi,\alpha)\leftarrow \frac{1}{n}\sum_{i=1}^n W_i(\pi,\alpha)$.
6: Update $\pi\leftarrow \arg\min_{\pi\in\Pi}\hat{W}_n(\pi,\alpha)$.
7: Update $\alpha\leftarrow \alpha-(\frac{\partial}{\partial\alpha}\hat{\phi}_n)/(\frac{\partial^2}{\partial\alpha^2}\hat{\phi}_n)$.
8: **until** α converge.
9: **Return** $\hat{\pi}_{\rm DRO}\leftarrow\pi$.

In Section 5, we show the empirical performance of the policy $\hat{\pi}_{\rm DRO}$ on the synthetic data.

4.2. Stable Policy Learning

In the numerical experiment, we found that the policy learned from (4.1) is unstable and has huge bias. The reason is that the algorithm attempts to learn a policy that matches $\pi(X_i)=A_i$ as little as possible. Therefore, we propose the following stable evaluation formula with normalization,

$$\hat{W}_n^{\rm stable}(\pi,\alpha)=\frac{1}{nS_n^\pi}\sum_{i=1}^n W_i(\pi,\alpha),$$

where $S_n^\pi=\frac{1}{n}\sum_{i=1}^n\frac{\mathbf{1}\{\pi(X_i)=A_i\}}{\pi_0(A_i|X_i)}$. Accordingly, we define

$$\hat{\phi}_n^{\rm stable}(\pi,\alpha)=-\alpha\log\hat{W}_n^{\rm stable}(\pi,\alpha)-\alpha\delta,$$

and

$$\hat{Q}_{\rm DRO}^{\rm stable}(\pi)=\sup_{\alpha\geq 0}\left\{\hat{\phi}_n^{\rm stable}(\pi,\alpha)\right\}.$$

In order to implement the Newton–Raphson method to update α, we compute the closed form expression of $\frac{\partial}{\partial\alpha}\hat{\phi}_n^{\rm stable}(\pi,\alpha)$ and $\frac{\partial^2}{\partial\alpha^2}\hat{\phi}_n^{\rm stable}(\pi,\alpha)$ as follows

$$\begin{aligned}\frac{\partial}{\partial\alpha}\hat{\phi}_n^{\rm stable}(\pi,\alpha) &= -\frac{\sum_{i=1}^n Y_i(A_i)W_i(\pi,\alpha)}{\alpha nS_n^\pi\hat{W}_n^{\rm stable}(\pi,\alpha)}\\ &\quad -\log\hat{W}_n^{\rm stable}(\pi,\alpha)-\delta,\\ \frac{\partial^2}{\partial\alpha^2}\hat{\phi}_n^{\rm stable}(\pi,\alpha) &= \frac{(\sum_{i=1}^n Y_i(A_i)W_i(\pi,\alpha))^2}{\alpha^3(nS_n^\pi)^2(\hat{W}_n^{\rm stable}(\pi,\alpha))^2}\\ &\quad -\frac{\sum_{i=1}^n Y_i^2(A_i)W_i(\pi,\alpha)}{\alpha^3 nS_n^\pi\hat{W}_n^{\rm stable}(\pi,\alpha)}.\end{aligned}$$

This evaluation scheme then allows us for designing a better distributionally robust policy learning algorithm, which is formally given in Algorithm 3.

Algorithm 3 Stable Distributionally Robust Policy Learning

1: **Input:** Dataset $\{(X_i,A_i,Y_i)\}_{i=1}^n$, data-collecting policy π_0, and initial value of dual variable α.
2: **Output:** Distributionally robust optimal policy $\hat{\pi}_{\rm DRO}^{\rm stable}$.
3: **repeat**
4: Let $W_i(\pi,\alpha)\leftarrow \frac{\mathbf{1}\{\pi(X_i)=A_i\}}{\pi_0(A_i|X_i)}\exp(-Y_i(A_i)/\alpha)$.
5: Compute $S_n^\pi\leftarrow\frac{1}{n}\sum_{i=1}^n\frac{\mathbf{1}\{\pi(X_i)=A_i\}}{\pi_0(A_i|X_i)}$.
6: Compute $\hat{W}_n^{\rm stable}(\pi,\alpha)\leftarrow\frac{1}{nS_n^\pi}\sum_{i=1}^n W_i(\pi,\alpha)$.
7: Update $\pi\leftarrow\arg\min_{\pi\in\Pi}\hat{W}_n^{\rm stable}(\pi,\alpha)$.
8: Update $\alpha\leftarrow\alpha-(\frac{\partial}{\partial\alpha}\hat{\phi}_n^{\rm stable})/(\frac{\partial^2}{\partial\alpha^2}\hat{\phi}_n^{\rm stable})$.
9: **until** α converge.
10: **Return** $\hat{\pi}_{\rm DRO}^{\rm stable}\leftarrow\pi$.

This normalization trick, although only present in the context of distributionally robust bandit learning problems in this paper, we believe, is applicable to many other bandit learning settings.

In fact, Proposition 4.1 shows that for a fix policy π, the difference between $\hat{Q}_{\rm DRO}^{\rm stable}(\pi)$ and $\hat{Q}_{\rm DRO}(\pi)$ is still at a canonical statistical $O(n^{-1/2})$ rate. This means in terms of policy evaluation (in contrast to policy learning), $\hat{Q}_{\rm DRO}^{\rm stable}(\pi)$ and $\hat{Q}_{\rm DRO}(\pi)$ give the similar performance.

Proposition 4.1. *Suppose Assumption 1 is enforced, when* $n>\frac{1}{2}\left(\frac{\log(2/\epsilon)}{(1-\max\{1/4,\exp(-\delta/2)\})\eta}\right)^2$, *with probability at least* $1-\epsilon$, *we have*

$$\left|\hat{Q}_{\rm DRO}^{\rm stable}(\pi)-\hat{Q}_{\rm DRO}(\pi)\right|\leq\frac{2\sqrt{2}M}{\delta\eta}\frac{\log(2/\epsilon)}{\sqrt{n}}.$$

The proof of Proposition 4.1 is in Appendix A.5. From Proposition 4.1, we observe that $\hat{Q}_{\rm DRO}^{\rm stable}$ and $\hat{Q}_{\rm DRO}(\pi)$ are relatively closed if δ is relatively large. In fact, we can also derive the central limit theorem for $\hat{Q}_{\rm DRO}^{\rm stable}$ using the same techniques as the proof of Theorem 3.4:

$$\sqrt{n}\left(\hat{Q}_{\rm DRO}^{\rm stable}(\pi)-Q_{\rm DRO}(\pi)\right)\Rightarrow N\left(0,\sigma_{\rm stable}^2(\alpha^*)\right),$$

where

$$\sigma_{\rm stable}^2(\alpha)$$

	Region 0	Region 1	Region 2	Region 3
Action 0	0.4	0.2	0.2	0.2
Action 1	0.2	0.3	0.3	0.2
Action 2	0.2	0.3	0.3	0.2
Action 3	0.2	0.2	0.2	0.4

Table 1: The probabilities of selecting an action based on π_0.

$$= \frac{\alpha^2}{\mathbf{E}\left[W_i(\pi,\alpha)\right]^2}\mathbf{E}\left[\frac{1}{\pi_0\left(\pi(X)|X\right)}\left(\exp\left(-Y(\pi(X))/\alpha\right)\right.\right.$$
$$\left.\left. - \mathbf{E}\left[\exp\left(-Y(\pi(X))/\alpha\right)\right]\right)^2\right].$$

When δ is small, we have $\sigma^2(\alpha^*)$ is $O(1/\delta)$, since $\mathbf{Var}\left[W_i(\pi,\alpha)\right] = O(1)$; while, on the other hand, $\sigma^2_{\text{stable}}(\alpha^*)$ is $O(1)$. Furthermore, direct calculation gives us if the propensity score is sufficiently uniform, namely

$$\frac{2}{\pi_0(\pi(X)|X)} \geq 1 + \mathbf{E}\left[\frac{1}{\pi_0(\pi(X)|X)}\right] \text{ almost surely,}$$

we have $\sigma^2_{\text{stable}}(\alpha^*) > \sigma^2(\alpha^*)$. Therefore, in general, $\hat{Q}^{\text{stable}}_{\text{DRO}}(\pi)$ is expected to have a smaller variance than $\hat{Q}_{\text{DRO}}(\pi)$.

5. Numerical Experiments

5.1. Experiment Setup

We first describe the simulation environment. The feature vectors $X_i \in \mathbf{R}^{10}$ are independently and uniformly drawn from $[0,1]^{10}$. We denote the 10 dimensions by $x_0, \ldots, x_9$. The action set $\mathcal{A}$ consists of four actions, i.e., $\mathcal{A} = \{0,1,2,3\}$. Given X_i, each reward vector $Y_i = (Y_i(0), Y_i(1), Y_i(2), Y_i(3))$ is are drawn i.i.d. from multivariate normal distribution such that the entries are mutually independent with variance σ_i and $\mathbf{E}[Y_i(a)|X_i] = \mu_a(X_i)$ for $a = 0,1,2,3$, where

$$\begin{aligned}
&\mu_0(x) = 0, && \sigma_0 = 0.1,\\
&\mu_1(x) = 1 - \max\left\{\tfrac{|x_7-1|}{0.65}, \tfrac{|x_5|}{0.6}\right\}, && \sigma_1 = 0.3,\\
&\mu_2(x) = 1 - \tfrac{x_5^2}{0.6^2} - \tfrac{x_7^2}{0.35^2}, && \sigma_2 = 0.2,\\
&\mu_3(x) = 1 - \tfrac{(x_5-1)^2}{0.4^2} - \tfrac{(x_7-1)^2}{0.35^2}, && \sigma_3 = 0.1.
\end{aligned}$$

The feature space $[0,1]^{10}$ is partitioned into four mutually disjoint regions such that the optimal action in each region coincides with the region number, of which a graphical illustration is provided in Figure 1.

Given X_i, the action A_i is drawn according to the underlying data collection policy π_0, which is described in Table 1.

5.2. Decision Trees and Greedy Tree Search

In this section, we introduce a celebrated policy class called decision trees (Breiman et al., 1984). A depth-L tree has L layers in total: the first $L-1$ layers consist of branch nodes, while the L-th layer consists of leaf nodes. By traversing a particular path determined by features x from the root node to a leaf node, an action a is uniquely determined. Each branch node is specified by two quantities: the dimension to be split on and the threshold b. If for a particular branching node, the splitting dimension is the i-th dimension and $x_i < b$, then the left child of the node is followed; otherwise the right child is followed. Every path terminates at a leaf node, each of which is assigned a unique label corresponding to one of the possible actions in $\mathcal{A}$.

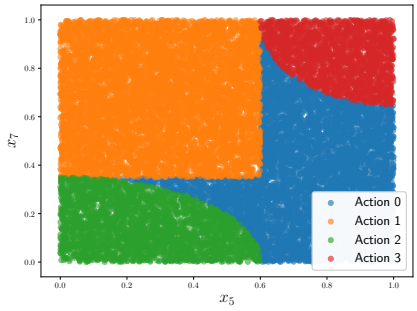

Figure 1: A pictorial illustration of the setup of the data-generating distribution. The figure provides a two-dimensional slice of the entire 10-dimensional feature space.

The algorithm for decision tree learning need to be computationally efficient, since algorithm will be iteratively executed in Line 6 of Algorithm 2. As finding an optimal classification tree is generally intractable (Bertsimas & Dunn, 2017), here we introduce an heuristic algorithm called Greedy Tree Search (GTS). The procedure of GTS can be inductively defined. First, to learn a depth-2 tree, GTS will brute force search all the possible spliting choices of the branch node, as well as all the possible actions of the leaf nodes. Suppose that the learning procedure for depth-$(L-1)$ tree has been defined. To learn a depth-L tree, we first learn a depth-2 tree with the optimal branching node, which partitions all the training data into two disjoint groups associated with two leaf nodes. Then each leaf node is replaced by the depth-$(L-1)$ tree trained using the data in the associated group.

5.3. Policy Evaluation

In this section, we test the efficacy of $\hat{Q}_{\text{DRO}}(\pi)$ and $\hat{Q}^{\text{stable}}_{\text{DRO}}(\pi)$. The goal is to validate that for each $\delta > 0$, both estimators converge to $Q_{\text{DRO}}(\pi)$ when size of training set is increasing to infinity. Due to the fact that the true value of $Q_{\text{DRO}}(\pi)$ is unknown, we define the following benchmark estimator $\hat{Q}^{\text{full}}_{\text{DRO}}(\pi)$ which utilizes all the entries of

Y_i:

$$\hat{Q}^{\text{full}}_{\text{DRO}}(\pi) = \sup_{\alpha \geq 0} \left\{ -\alpha \log \left(\frac{1}{n} \sum_{i=1}^{n} \exp(-Y_i(\pi(X_i))/\alpha) \right) - \alpha\delta \right\}.$$

It follows from the first equality of Lemma 3.1 that $\hat{Q}^{\text{full}}_{\text{DRO}}(\pi)$ is a consistent estimator of $Q_{\text{DRO}}(\pi)$. However, due to the fact that $Y_i(a)$ is in practice not observable when $a \neq A_i$, the estimator $\hat{Q}^{\text{full}}_{\text{DRO}}(\pi)$ is designed only for the sake of comparison in the simulation environment, where all the entries of Y_i are simulated.

We first test the convergence of different estimators for $\delta = 0.2$ and three different sizes of dataset: $n = 10^3, 10^4, 10^5$. The data $\{X_i, Y_i, A_i\}_{i=1}^n$ is generated using the procedure described in Section 5.1. We also choose a fixed depth-3 decision tree as shown in Figure 2. We report in Table 2 the mean and standard error of the estimators computed using 100 i.i.d. experiments. The numerical result demonstrates that all the estimators are converging to the same value and the scaling rate of standard errors is consistent with the $O(n^{-1/2})$ rate suggested by Theorem 3.4.

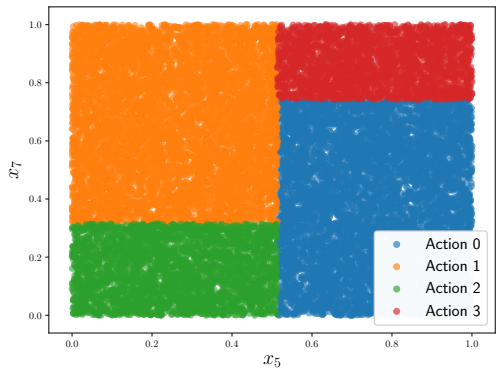

Figure 2: Visualization of decision tree used in policy evaluation.

Estimator	$n = 10^3$	$n = 10^4$	$n = 10^5$
$\hat{Q}_{\text{DRO}}(\pi)$	$0.0294 \pm .0312$	$0.0258 \pm .0101$	$0.0247 \pm .0030$
$\hat{Q}^{\text{full}}_{\text{DRO}}(\pi)$	$0.0267 \pm .0099$	$0.0254 \pm .0034$	$0.0247 \pm .0010$
$\hat{Q}^{\text{stable}}_{\text{DRO}}(\pi)$	$0.0256 \pm .0207$	$0.0255 \pm .0066$	$0.0245 \pm .0021$

Table 2: Policy evaluation for different size of dataset.

We also test the performance of different estimators with different levels of robustness. Since Table 2 suggests that $n = 10^5$ is large enough to provide stable estimation, here we pick $n = 10^5$ and only conduct the experiment once. The numerical result given in table 3 suggests that our estimator achieves uniformly satisfactory performance for different δ, which is consistent with Proposition 4.1.

Estimator	$\delta = 0.0$	$\delta = 0.2$	$\delta = 0.4$	$\delta = 0.6$	$\delta = 0.8$
$\hat{Q}_{\text{DRO}}(\pi)$	0.2489	0.0243	-0.0587	-0.1207	-0.1723
$\hat{Q}^{\text{full}}_{\text{DRO}}(\pi)$	0.2492	0.0253	-0.0588	-0.1215	-0.1741
$\hat{Q}^{\text{stable}}_{\text{DRO}}(\pi)$	0.2470	0.0259	-0.0576	-0.1198	-0.1716

Table 3: Policy evaluation for different δ.

5.4. Policy Learning

In this section we report the performance of distributionally robust policy learning. We consider three different distributionally robust policies: $\hat{\pi}^{\text{full}}_{\text{DRO}}, \hat{\pi}_{\text{DRO}}, \hat{\pi}^{\text{stable}}_{\text{DRO}}$ that maximize $\hat{Q}^{\text{full}}_{\text{DRO}}, \hat{Q}_{\text{DRO}}, \hat{Q}^{\text{stable}}_{\text{DRO}}$, respectively. As a comparison, we also report the performance of two non-robust policies. We define two non-robust policy evaluation estimators:

$$\hat{Q}_{\text{emp}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i(A_i)\mathbf{1}\{\pi(X_i) = A_i\}}{\pi_0(A_i \mid X_i)},$$

$$\hat{Q}^{\text{full}}_{\text{emp}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} Y_i(\pi(X_i)).$$

We apply GTS to $\hat{Q}_{\text{emp}}(\pi)$ and $\hat{Q}^{\text{full}}_{\text{emp}}(\pi)$, and denote the output policy by $\hat{\pi}_{\text{emp}}$ and $\hat{\pi}^{\text{full}}_{\text{emp}}$, respectively.

We fix $\delta = 0.1$ and the size of training set is $n = 3000$, and the policy class is depth-3 trees. Algorithms 2 and 3 (also an analog for $\hat{\pi}^{\text{full}}_{\text{DRO}}$) are applied to the training set, producing policies, $\hat{\pi}_{\text{DRO}}$, $\hat{\pi}^{\text{stable}}_{\text{DRO}}$ and $\hat{\pi}^{\text{full}}_{\text{DRO}}$. To evaluate the performance of our policy, we do the following three sets of experiments and repeat each sets of experiments 100 times to obtain the mean and standard error of all the evaluation metrics. The results are reported in Table 4. We also present an instance of optimal distributionally robust decision tree in Figure 3.

1. We generate a test set with $n = 3000$ i.i.d. data points sampled from $\mathbf{P}_0$ and evaluate the performance of each policy using $\hat{Q}^{\text{full}}_{\text{emp}}$. Notice that in this experiment, the training and test environments are the same. There is no distribution shifts between the training and test data. The results are reported in the first column in Table 4.

2. We generate a test set with $n = 3000$ i.i.d. data points sampled from $\mathbf{P}_0$ and evaluate the worst case performance of each policy using $\hat{Q}^{\text{full}}_{\text{DRO}}$. The results are reported in the second column in Table 4.

3. We first generate $M = 100$ independent test sets, where each test set consists of $n = 3000$ i.i.d. data points sampled from $\mathbf{P}_0$. We denote them by $\left\{ \left\{ \left(X_i^{(j)}, Y_i^{(j)}(a^1), \ldots, Y_i^{(j)}(a^d) \right) \right\}_{i=1}^{n} \right\}_{j=1}^{M}$. Then, we randomly sample a new dataset around each dataset, i.e., $\left(\tilde{X}_i^{(j)}, \tilde{Y}_i^{(j)}(a^1), \ldots, \tilde{Y}_i^{(j)}(a^d) \right)$ is sampled in the

KL-ball centered at $\left(X_i^{(j)}, Y_i^{(j)}(a^1), \ldots, Y_i^{(j)}(a^d)\right)$. Then, we evaluate each policy using $\hat{Q}^{\text{full}}_{\min}$, defined by

$$\hat{Q}^{\text{full}}_{\min}(\pi) = \min_{1\le j\le M} \left\{ \frac{1}{n}\sum_{i=1}^{n} \tilde{Y}_i^{(j)} \left(\pi\left(\tilde{X}_i^{(j)}\right)\right) \right\}.$$

The results are reported in the third column in Table 4.

The third set of experiment mimic the distribution shifts in the real world, since the worst case performance is usually too extreme.

Policy	$\hat{Q}^{\text{full}}_{\text{emp}}$	$\hat{Q}^{\text{full}}_{\text{DRO}}$	$\hat{Q}^{\text{full}}_{\min}$
$\hat{\pi}_{\text{emp}}$	$0.243 \pm .023$	$0.059 \pm .029$	$0.191 \pm .024$
$\hat{\pi}^{\text{full}}_{\text{emp}}$	$0.246 \pm .013$	$0.067 \pm .018$	$0.195 \pm .019$
$\hat{\pi}_{\text{DRO}}$	$0.240 \pm .012$	$0.067 \pm .018$	$0.195 \pm .018$
$\hat{\pi}^{\text{full}}_{\text{DRO}}$	$0.246 \pm .011$	$0.078 \pm .015$	$0.203 \pm .016$
$\hat{\pi}^{\text{stable}}_{\text{DRO}}$	$0.243 \pm .013$	$0.075 \pm .016$	$0.203 \pm .017$

Table 4: Comparison of different policies.

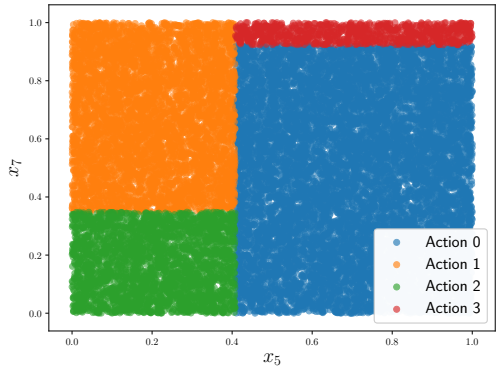

Figure 3: An instance of optimal distributionally robust decision tree $\hat{\pi}^{\text{stable}}_{\text{DRO}}$.

From Table 4, we find that $\hat{\pi}^{\text{stable}}_{\text{DRO}}$ performs consistently better than $\hat{\pi}_{\text{DRO}}$. Further, the performance of $\hat{\pi}^{\text{stable}}_{\text{DRO}}$ and $\hat{\pi}_{\text{emp}}$ are comparable in the non-robust experiment (the first column in Table 4) and $\hat{\pi}^{\text{stable}}_{\text{DRO}}$ is indeed more robust in terms of the worst performance and the minimum performance of 100 test sets sampled from the uncertainty ball. Despite our distributionally robust policies are trained from metric $\hat{Q}_{\text{DRO}}$, they also exhibit robustness under the other robustness measure $\hat{Q}^{\text{full}}_{\min}$.

6. Conclusion and Future Work

We have provided a distributionally robust formulation for policy evaluation and policy learning in offline contextual bandits. Our work only provides a preliminary result and there are many interesting subsequent directions worthy pursuing. For example, how to develop a robust policy when confounding factors are presented. we believe it is possible to incorporate the instrument variables in our framework. Another interesting direction would be to extend the algorithm and results to the Wasserstein distance case for batch contextual bandits, which has a fundamental difference with our KL-divergence framework. We leave them for future work.

Acknowledgements

Support is acknowledged from NSF grants 1915967, 1820942, 1838576 and AFOSR MURI 19RT1056 and the China Merchants Bank.

References

Abadeh, S. S., Nguyen, V. A., Kuhn, D., and Esfahani, P. M. Wasserstein distributionally robust kalman filtering. In *Advances in Neural Information Processing Systems*, pp. 8483–8492, 2018.

Abeille, M., Lazaric, A., et al. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

Agrawal, S. and Goyal, N. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pp. 99–107, 2013a.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135, 2013b.

Araujo, A. and Giné, E. *The central limit theorem for real and Banach valued random variables*. John Wiley & Sons, 1980.

Bastani, H. and Bayati, M. Online decision-making with high-dimensional covariates. 2015.

Bayraksan, G. and Love, D. K. Data-driven stochastic programming using phi-divergences. In *The Operations Research Revolution*, pp. 1–19. Catonsville: Institute for Operations Research and the Management Sciences, 2015.

Bertsimas, D. and Dunn, J. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.

Bertsimas, D. and Mersereau, A. J. A learning approach for interactive marketing to a customer segment. *Operations Research*, 55(6):1120–1135, 2007.

Bertsimas, D. and Sim, M. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

Bistritz, I., Zhou, Z., Chen, X., Bambos, N., and Blanchet, J. Online exp3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, pp. 11345–11354, 2019.

Blanchet, J. and Murthy, K. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44 (2):565–600, 2019. doi: 10.1287/moor.2018.0936.

Bonnans, J. F. and Shapiro, A. Perturbation analysis of optimization problems. 2000.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and Regression Trees*. CRC press, 1984.

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Chapelle, O. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1097–1105. ACM, 2014.

Chen, Z., Kuhn, D., and Wiesemann, W. Data-driven chance constrained programs over wasserstein balls. *arXiv preprint arXiv:1809.00210*, 2018.

Chernozhukov, V., Demirer, M., Lewis, G., and Syrgkanis, V. Semi-parametric efficient policy learning with continuous actions. In *Advances in Neural Information Processing Systems*, pp. 15039–15049, 2019.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

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

Dimakopoulou, M., Zhou, Z., Athey, S., Imbens, G., et al. Estimation considerations in contextual bandits. Technical report, 2018.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. Balanced linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3445–3453, 2019.

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Duchi, J., Glynn, P., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

Duchi, J. C., Hashimoto, T., and Namkoong, H. Distributionally robust losses against mixture covariate shifts. *Under review*, 2019.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011.

Faury, L., Tanielian, U., Vasile, F., Smirnova, E., and Dohmatob, E. Distributionally robust counterfactual risk minimization. *arXiv preprint arXiv:1906.06211*, 2019.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pp. 586–594, 2010.

Gao, R. and Kleywegt, A. J. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

Gao, R., Xie, L., Xie, Y., and Xu, H. Robust hypothesis testing using wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, pp. 7902–7912, 2018.

Ghosh, S. and Lam, H. Robust analysis in stochastic simulation: Computation and performance guarantees. *Operations Research*, 2019.

Goldenshluger, A. and Zeevi, A. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

Hu, Z. and Hong, L. J. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

Imbens, G. and Rubin, D. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, 2015. ISBN 9780521885881.

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

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2304–2313, 2018.

Joachims, T., Swaminathan, A., and Rijke, M. d. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*, May 2018.

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, pp. 99–109, 2017.

Kallus, N. Balanced policy evaluation and learning. *Advances in Neural Information Processing Systems*, 2018.

Kallus, N. and Zhou, A. Confounding-robust policy improvement. *arXiv preprint arXiv:1805.08593*, 2018.

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

Lam, H. and Zhou, E. The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters*, 45(4):301–307, 2017.

Lattimore, T. and Szepesvári, C. Bandit algorithms. *preprint*, pp. 28, 2018.

Lee, J. and Raginsky, M. Minimax statistical learning with wasserstein distances. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 2692–2701, USA, 2018. Curran Associates Inc.

Lehmann, E. L. and Casella, G. *Theory of point estimation*. Springer Science & Business Media, 2006.

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

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2071–2080. JMLR. org, 2017.

Master, N., Zhou, Z., Miller, D., Scheinker, D., Bambos, N., and Glynn, P. Improving predictions of pediatric surgical durations with supervised learning. *International Journal of Data Science and Analytics*, 4(1):35–52, 2017.

Mohajerin Esfahani, P. and Kuhn, D. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, Sep 2018. ISSN 1436-4646. doi: 10.1007/s10107-017-1172-1.

Nguyen, V. A., Kuhn, D., and Esfahani, P. M. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *arXiv preprint arXiv:1805.07194*, 2018.

Rakhlin, A. and Sridharan, K. BISTRO: An efficient relaxation-based method for contextual bandits. In *Proceedings of the International Conference on Machine Learning*, pp. 1977–1985, 2016.

Rigollet, P. and Zeevi, A. Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*, 2010.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

Shafieezadeh-Abadeh, S., Esfahani, P., and Kuhn, D. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pp. 1576–1584. 2015.

Shapiro, A. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

Sinha, A., Namkoong, H., and Duchi, J. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Slivkins, A. et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Staib, M. and Jegelka, S. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, 2017.

Swaminathan, A. and Joachims, T. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16:1731–1755, 2015.

Vitus, M. P., Zhou, Z., and Tomlin, C. J. Stochastic control with uncertain parameters via chance constrained control. *IEEE Transactions on Automatic Control*, 61(10):2892–2905, 2015.

Yang, I. Wasserstein distributionally robust stochastic control: A data-driven approach. *arXiv preprint arXiv:1812.09808*, 2018.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.

Zhao, C. and Guan, Y. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262 – 267, 2018. ISSN 0167-6377. doi: https://doi.org/10.1016/j.orl.2018.01.011.

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

Zhao, Y.-Q., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1): 151–168, 2014.

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

Zhou, Z., Bloem, M., and Bambos, N. Infinite time horizon maximum causal entropy inverse reinforcement learning. *IEEE Transactions on Automatic Control*, 63(9):2787–2802, 2017b.

Zhou, Z., Athey, S., and Wager, S. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*, 2018.

Zhou, Z., Xu, R., and Blanchet, J. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*, pp. 5198–5209, 2019.