Finite-Sample Regret Bound for Distributionally Robust Policy Learning in Tabular RL

Zhengqing Zhou
Stanford University

Zhengyuan Zhou
NYU Stern

Qinxun Bai
Horizon Robotics

Linhai Qiu
Google

Jose Blanchet
Stanford University

Peter Glynn
Stanford University

Abstract

While reinforcement learning has witnessed tremendous success recently in a wide range of domains, robustness–or the lack thereof–remains an important issue that has not been fully explored. In this paper, we provide a distributionally robust formulation of offline learning policy in tabular RL that aims to learn a policy from historical data (collected by some other behavior policy) that is robust to the future environment that can deviate from the training environment. We first develop a novel policy evaluation scheme that accurately estimates the robust value (i.e. how robust it is in a perturbed environment) of any given policy and establish its finite-sample estimation error. Building on this, we then develop a novel and minimax-optimal distributionally robust learning algorithm that achieves $O_P\left(1/\sqrt{n}\right)$ regret, meaning that with high probability, the policy learned from using n training data points will be $O\left(1/\sqrt{n}\right)$ close to the optimal distributionally robust policy. Finally, our simulation results demonstrate the superiority of our distributionally robust approach compared to non-robust RL algorithms.

1 Introduction

Reinforcement learning (RL) has emerged to be an important and active research area that has found widespread applications–including robotics (Kober et al., 2013; Gu et al., 2017), computer vision (Sadeghi and Levine, 2016; Huang et al., 2017), finance (Li et al., 2009; Choi et al., 2009; Deng et al., 2017) and game-playing (Silver et al., 2016, 2018)–and has witnessed great recent empirical success (Bertsekas and Tsitsiklis, 1996; Powell, 2007; Bertsekas, 2011; Szepesvári, 2010; Sutton and Barto, 2018).

Within RL, offline policy learning–learning an optimal policy from historical data collected from some other (behavior) policy–is an important subarea that has been extensively studied (Dudík et al., 2011; Zhang et al., 2012; Lazaric et al., 2012; Mahmood et al., 2014; Jiang and Li, 2016; Munos et al., 2016; Athey and Wager, 2017; Zhou et al., 2018; Kallus and Zhou, 2018; Fujimoto et al., 2019; Chen and Jiang, 2019). The primary focus in this line of work has been to develop efficient estimation schemes that are able to perform accurate policy evaluation for an arbitrary policy (using the historical data), which is then used in the policy optimization step to select the best policy. Further, the key quantity that is used to assess the effectiveness of a proposed policy learning algorithm is regret, which measures the discrepancy of the value of the learned policy and that of the optimal policy when n training samples are available (note that regret can be equivalently understood as sample complexity). This quantity has been well-studied and it is well-known that the optimal regret scales as $\Theta_P\left(1/\sqrt{n}\right)$ in tabular and certain parametric settings, and different algorithms exist that achieve this bound (although some are superior to others in terms of constants in the $\Theta_P(\cdot)$).

Despite the significant advances provided by this rich literature, the existing works in this area are missing an important aspect: robustness (i.e., a learned policy is insensitive to and continues to perform well in a new environment). This is because offline policy learning in RL–as it currently stands–makes the crucial as-

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

sumption that may not always hold in practice: the past environment from which the historical data has been collected is the same as the future environment in which the learned policy will be deployed. Under this assumption, learning a good policy means learning a policy that yields high value (sum of discounted future rewards) in the old environment, and because the new environment is identical to the old environment, this policy will be equally effective in the new environment.

In practice, however, this assumption often does not hold, and a change in the future environment would often render the learned policy ineffective. For instance, in financial trading (an area where RL has seen a growing applied presence in the industry (Li et al., 2009; Deng et al., 2017; Nevmyvaka et al., 2006; Choi et al., 2009)), the markets are often volatile and non-stationary, and the assumption that the past financial markets data faithfully represent the future markets is simply invalid. Another area where robustness is highly desired is robotics, where a robot trained to perform certain maneuvers (such as walking Schulman et al. (2013) or folding laundry (Maitin-Shepard et al., 2010)) in an environment can fail catastrophically (Drew, 2015) in a slightly different environment, where the terrain landscape (in walking) is slightly altered or the laundry object (in laundry folding) is positioned differently. This indicates that policy learning assuming future environments are the same will often yield fragile policies that are not useful except in highly controlled "testing" environment.

As such, learning robust policies that account for potentially varied future environments is of enormous importance and practical utility. However, traditional robust RL approaches (Başar and Bernhard, 2008; Xie and de Souza, 1990; Ugrinovskii, 1998; Morimoto and Doya, 2001; Petersen et al., 2000; Dupuis et al., 2000)–which stem from the H_∞-control perspective–studies policies that are robust to the worst possible deterministic environment. This framework–also adopted in the early robust control and robust optimization literature–is often overly conservative, and does not yield effective policies for many commonly occurring environment-shift scenarios. Consequently, developing practically useful policies that are robust to commonplace shifts in environments calls for a new paradigm.

Recently, distributional robustness (Duchi and Namkoong, 2018; Blanchet and Kang, 2020; Duchi et al., 2016; Duchi and Namkoong, 2019; Blanchet et al., 2019; Bertsimas et al., 2018; Gao and Kleywegt, 2016; Esfahani and Kuhn, 2018) has emerged to be an effective metric for capturing robust-yet-not-overly-conservative learning performance in the supervised learning setting. At a high level, distributional robustness posits that the future environment is characterized by a distribution (as opposed to a deterministic quantity) that lies in a neighborhood around the old-environment distribution. The robustness is ensured by optimizing (model parameters) over the worst possible distribution in this neighborhood (under certain distance metric on probability distributions). It is worth noting that there are two types of distributional robust philosophies one can pursue: one is that the environment itself does not change and the learner wants add robustness in its algorithm so as to not be misled by the lack of data. In such cases, δ (the robustness parameter) needs to go to 0 (at an appropriate rate). The other, which is what we are pursuing here, is that the environment has an intrinsic shift. In such cases, δ must remain constant. Our results make this adaptation formal by characterizing a clear dependency on δ in the regret bound.

While the recently flourishing line of work has studied distributional robust learning–particularly the sample complexity guarantees and algorithms for achieving them–in supervised learning, distributional robust policy learning in RL has not been explored, except in the highly special and limiting case where the entire learning horizon is one (Si et al., 2020), a setting also known as contextual bandits. Consequently, our goal in this paper is to fill in this gap and bring distributional robustness–and particularly a rigorous understanding of finite-sample guarantees thereof–into the offline RL setting. In chartering an initial path in this unexplored landscape, we focus our attention on tabular RL in the infinite-horizon discounted setting for concreteness, where we aim to understand how many samples are sufficient for learning the optimal distributionally robust policy.

1.1 Our Contributions

We have two core contributions.

First, we borrow the distributional robustness concept from supervised learning (Duchi and Namkoong, 2018; Blanchet and Kang, 2020; Duchi et al., 2016; Duchi and Namkoong, 2019; Blanchet et al., 2019; Bertsimas et al., 2018; Gao and Kleywegt, 2016; Esfahani and Kuhn, 2018) and formulate a offline policy learning problem where the goal is to learn a distributionally robust policy using as few samples as possible. We measure sample complexity using distributional robust regret (analogue of the standard regret for offline policy learning), which is the discrepancy between the value of the policy learned by the algorithm and that of the optimal distributionally robust policy. With regards to this formulation, we point out that there has been a line of work on distributionally robust MDPs (Iyengar, 2005; Xu and Mannor, 2010; Wolff et al., 2012; Wiesemann et al., 2013) from the operations research

community. Despite the seeming similarity, there are two crucial differences that clearly separate the two problems. First, that line of work assumes the underlying MDP is known, hence there is no learning that occurs and it's simply an optimization problem. Second, that line of work only focuses on developing optimization algorithms (which are not applicable here as they require the true MDP to be known) and has no finite-sample guarantees.

Second, we provide a novel policy learning algorithm (Algorithm 2) that takes advantage of the simple structure of the dual space of the distributions that characterize the possible environments. We then establish in Theorem 2 that this algorithm achieves the finite-sample regret of $O_P\left(1/\sqrt{n}\right)$, meaning that with high probability, the policy learned from using n training data points will be $O\left(1/\sqrt{n}\right)$ close to the optimal distributionally robust policy. Since $\Omega_P\left(1/\sqrt{n}\right)$ is a lower bound (even in standard offline policy learning), our algorithm is minimax-optimal. In obtaining this finite-sample regret guarantee, we also study distributionally robust policy evaluation–both algorithmically (Algorithm 1) and theoretically (Theorem 1)– which is not only an important intermediate step, but also of independent interest on its own.

Additionally, we provide two sets of simulations–one in the gambler's problem and the other in options trading–that demonstrate the superiority of our distributionally robust approach. In each setting, we show that the policy learned from our distributionally robust approach is robust and performs much better in altered future environments than that learned through the standard policy learning approach. Taken together, our results provide the first inroad into the broad landscape of distributionally robust RL that aims to understand how to learn an effective and robust policy from data.

2 A Distributionally Robust Formulation of Offline Policy Learning in Tabular RL

2.1 Standard Offline Policy Learning in RL

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ be a tabular RL environment, where $\mathcal{S}$ and $\mathcal{A}$ are finite state space and action space respectively, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{P}(\mathbb{R}_{\geq 0})$ (the set of random variables that are supported on $\mathbb{R}_{\geq 0}$) is the randomized reward function, $\mathcal{P} = \{p_{s,a}(\cdot)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ is the transition model, and $\gamma \in (0,1)$ is the discount factor. We assume that the transition is Markovian, i.e., at each state $s \in \mathcal{S}$, if action $a \in \mathcal{A}$ is chosen, then the subsequent state is determined by the conditional distribution $p_{s,a}(\cdot) = p(\cdot|s,a)$. The decision maker will therefore receive a randomized reward $r(s,a)$. A policy π is a mapping from $\mathcal{S}$ to $\mathcal{A}$. Π denotes the class of deterministic policies.

In the world of offline policy learning, agent are only allowed to utilize historical data, without additional online interaction with the environment. To be specific, we assume the agent only has access to a batch dataset $\{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ from the original environment, where the actions $\{a_i\}$ are generated by some known policy π_0, r_i is a realization of the random reward $r(s_i, a_i)$, and s_i' is determined by the transition kernel $p^0_{s_i,a_i}(\cdot)$. We may also assume that the data are sampled uniformly across the state space $\mathcal{S}$. Note that π_0 is the policy that we use to generate the observable training data, and it could be any randomized policy

Let $(s_t, a_t)_{t\in\mathcal{T}}$ be the stochastic process that is induced by the transition model $\mathcal{P}$ and the policy π, the standard value function can be defined by

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \,\middle|\, s_1 = s\right].$$

The standard goal is to learn a policy π such that its value function is as large as possible (there exists a deterministic policy π^* that maximize $V^\pi(s)$ for all s), or equivalently, a policy that minimize the regret

$$\|V^*(s) - V^\pi(s)\|_\infty,$$

where $V^*(s) := \max_\pi V^\pi(s)$ and $\|\cdot\|_\infty$ is the L^∞ norm of a function.

Throughout this paper, we impose the following assumptions on the rewards and the data generating process.

Assumption 1. *Bounded reward with density. For any $(s,a) \in \mathcal{S}\times\mathcal{A}$, $r(s,a) \in [0, R_{\max}]$. Moreover, $r(s,a)$ has a density $f_{s,a}$.*

Assumption 2. *The training data are sampled uniformly across the state space $\mathcal{S}$.*

Assumption 3. *Overlapping. There exists some $\eta > 0$, such that $\mathbb{P}(\pi_0(s) = a) \geq \eta$, $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$.*

By Assumption 1, we know immediately that $V^\pi(s) \leq \frac{R_{\max}}{1-\gamma}$ for all the policy π and $s \in \mathcal{S}$. The Assumption 2 and 3 guarantees sufficient exploration in training data collection (in Assumption 3, π_0 is the policy that we use to generate the observable training data). Both the bounded reward and the overlapping assumptions are fairly standard and can be found in some policy learning literature (Si et al., 2020; Kallus, 2018; Zhao et al., 2012). The Assumption 2 is also standard in the theoretical analysis of batch (offline) RL, and can be founded in (Chen and Jiang, 2019). Lastly, the density assumption on reward is a technical assumption that ensures the $O_P(1/\sqrt{n})$ rate of convergence in Theorem 1 and 2.

2.2 Distributionally Robust Offline Policy Learning in RL

In practice, the transition model $\mathcal{P}$ and rewards $\mathcal{R}$ in the original environment $\mathcal{M}$ are subject to change over time, and we may not be able to update our policy by interacting with the environment. This practical challenge motivates us to learn a offline policy that is robust to certain perturbations in the environment. In particular, we consider the setting of distributionally robust offline RL, where both the transition probabilities and rewards are perturbed, based on the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(P\|Q) = \int \log\left(\frac{dP}{dQ}\right) dP$ whenever $P \ll Q$ (P is absolutely continuous with respect to Q).

In the original environment, let $\mathcal{P}^0 = \{p^0_{s,a}\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ be the transition probabilities, ν^0 be the joint distribution of $(r(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, where $r(s,a) \sim \nu^0_{s,a}$ (marginal distribution with respect to (s,a)). We may assume that $\mathcal{P}^0$ and ν^0 are independent.

To construct the distributional uncertainty set, for each $(s,a) \in \mathcal{S}\times\mathcal{A}$, we define robust KL balls that are centered at $p^0_{s,a}$ and $\nu^0_{s,a}$ by

$$\mathcal{P}_{s,a}(\delta) := \left\{p_{s,a} \in \Delta_{|\mathcal{S}|} : D_{\mathrm{KL}}\left(p_{s,a}\|p^0_{s,a}\right) \le \delta\right\}$$

and

$$\mathcal{R}_{s,a}(\delta) := \left\{r_{s,a} \sim \nu_{s,a} : D_{\mathrm{KL}}(\nu_{s,a}\|\nu^0_{s,a}) \le \delta\right\}$$

respectively. Here $\delta > 0$ is the level of distributional robustness, and $\Delta_{|\mathcal{S}|}$ stands for the $|\mathcal{S}|-1$ dimensional probability simplex. Now we are able to build the uncertainty set $\mathcal{P}(\delta)$ by the Cartesian product of $\mathcal{P}_{s,a}(\delta)$ for each $(s,a) \in \mathcal{S}\times\mathcal{A}$. This type of uncertainty set is called $(s,a)-$**rectangular** set in standard literature (Wiesemann et al., 2013). Similarly we define $\mathcal{R}(\delta)$ by the Cartesian product of $\mathcal{R}_{s,a}(\delta)$ for each $(s,a) \in \mathcal{S}\times\mathcal{A}$. In the distributionally robust framework, the adversarial player is assumed to pick the worst-case transition model and rewards that minimize the expected cumulative discounted reward. To be clear, we define the distributionally robust value function as follows.

Definition 1. *Given $\delta > 0$ and policy $\pi \in \Pi$, the distributionally robust value function $V^{\mathrm{rob},\pi}_\delta$ is defined as:*

$$V^{\mathrm{rob},\pi}_\delta(s) := \inf_{\mathbf{p}\in\mathcal{P}(\delta),\mathbf{r}\in\mathcal{R}(\delta)} \mathbb{E}_{\mathbf{p},\mathbf{r}}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r(s_t,a_t)\middle| s_1 = s\right]. \tag{1}$$

Follows from the definition, $V^{\mathrm{rob},\pi}$ measures the quality of a policy π by computing its performance in the worst-case environment among the set of all possible environments that perturb the original transition $\mathcal{P}^0$ and reward distribution ν^0 under a δ-KL ball.The optimal distributionally robust value function is therefore defined by

$$V^{\mathrm{rob},*}_\delta(s) := \max_{\pi\in\Pi} V^{\mathrm{rob},\pi}_\delta(s), \quad \forall s \in \mathcal{S},$$

and the optimal policy $\pi^{\mathrm{rob},*}_\delta$ is the deterministic policy that maximizes the distributionally robust value function, i.e.,

$$\pi^{\mathrm{rob},*}_\delta \in \arg\max_{\pi\in\Pi} V^{\mathrm{rob},\pi}_\delta. \tag{2}$$

Remark 1. *It sufficies to define $V^{\mathrm{rob},*}_\delta$ by taking the maximum over the class of deterministic policies Π. Suppose that $V^{\mathrm{rob},*}_\delta$ is defined by the maximum over all possible randomized policies .Then, by using the same techniques in Theorem 3.1 and 3.2 in (Iyengar, 2005), such optimal value function can be achieved by a deterministic policy. Thus, one can restrict the decision maker to Π without affecting the optimal distributionally robust value function.*

Our goal is to learn a robust policy π such that its distributionally robust value is as large as possible. In other words, we want the distributionally robust value of π is as close to the optimal distributionally robust policy value as possible. We formalize this discrepancy by the notion of regret.

Definition 2. *The distributionally robust regret $R^{\mathrm{rob}}(\pi)$ of a policy $\pi \in \Pi$ is defined as:*

$$R^{\mathrm{rob}}(\pi) := \|V^{\mathrm{rob},*}_\delta - V^{\mathrm{rob},\pi}_\delta\|_\infty.$$

Note that in our offline policy learning framework, we learn the policy $\widehat{\pi}$ from pre-collected data set $\{(s_i,a_i,r_i,s_i')\}_{i=1}^n$, hence both $\widehat{\pi}$ and $R^{\mathrm{rob}}(\widehat{\pi})$ are random variables.

In the rest of this paper, we aim to solve the following two main problems:

Q1. For any fixed policy $\pi \in \Pi$, how to approximately compute the distributionally robust value function $V^{\mathrm{rob},\pi}_\delta$ given the observational data? What is the rate of convergence of such an approximation?

Q2. How to learn a good distributionally robust policy $\widehat{\pi}$ given the observational data, where the performance is quantified by how the distributionally robust regret scales with respect to the size of the batch dataset?

3 Policy Learning in Distributionally Robust RL

To answer the main problems that are listed in section 2.2 by order, we first propose a policy evaluation algorithm (Algorithm 1 in section 3.2) that combine the idea of distributionally robust optimization with value-based RL. Next, we develop a policy learning algorithm that base on distributionally robust value iteration (Algorithm 2 in section 3.3), which output the robust policy that is based on the accessible data. Lastly, theoretical guarantees are discussed in section 3.4.

Before state the main algorithms and theoretical results, we need to discuss the optimization technique that serve as the cornerstone of our analysis.

3.1 Strong Duality

Follow from the well known results in (Iyengar, 2005; Xu and Mannor, 2010), we can write down the distributionally robust dynamical programing for the distributionally robust value function $V_\delta^{\mathrm{rob},\pi}$ in equation 1 as follows:

$$V_\delta^{\mathrm{rob},\pi}(s) = \inf_{\substack{p_{s,\pi(s)}\in\mathcal{P}_{s,\pi(s)}(\delta),\\ r\in\mathcal{R}_{s,\pi(s)}(\delta)}} \left\{ \mathbb{E}[r(s,\pi(s))] + \gamma \sum_{s'\in\mathcal{S}} p_{s,\pi(s)}(s') V_\delta^{\mathrm{rob},\pi}(s') \right\}. \tag{3}$$

Note that equation 3 is in general computationally intractable since it involves infinite dimensional optimization. To address this issue, we introduce the following strong duality lemma from distributionally robust optimization under KL-perturbation.

Lemma 1 (Hu and Hong (2013), Theorem 1)**.** *Suppose $H(X)$ has finite moment generating function in the neighborhood of zero. Then for any $\delta > 0$,*

$$\sup_{P:D_{\mathrm{KL}}(P\|P_0)\le\delta} \mathbb{E}_P\left[H(X)\right] = \inf_{\alpha\ge 0}\left\{\alpha\log\left(\mathbb{E}_{P_0}\left[e^{H(X)/\alpha}\right]\right) + \alpha\delta\right\}.$$

By Lemma 1, we can transform the equation 3 to the following equation.

$$V_\delta^{\mathrm{rob},\pi}(s) = \underbrace{\sup_{\alpha\ge 0}\left\{-\alpha\log\left(\mathbb{E}_{\nu^0_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right]\right) - \alpha\delta\right\}}_{R_\delta^{\mathrm{rob}}(s,\pi(s))} + \underbrace{\gamma\sup_{\beta\ge 0}\left\{-\beta\log\left(\sum_{s'\in\mathcal{S}} p^0_{s,\pi(s)}(s') e^{-V_\delta^{\mathrm{rob},\pi}(s')/\beta}\right) - \beta\delta\right\}}_{T_\delta^{\mathrm{rob},\pi}(s,\pi(s))}. \tag{4}$$

In the above dual formulation, the optimization problems that presented in $R_\delta^{\mathrm{rob},\pi}$ and $T_\delta^{\mathrm{rob},\pi}$ are both one dimensional concave optimization problems, and therefore computationally feasible.

As a direct consequence of the equation 4 (note that the size of Π is finite in the tabular setting, see Theorem 3.2 in Iyengar (2005) for a standard proof), the optimal distributionally robust value function $V_\delta^{\mathrm{rob},*}$ in fact satisfies the following distributionally robust Bellman's equation.

$$V_\delta^{\mathrm{rob},*}(s) = \max_{a\in\mathcal{A}}\left\{\sup_{\alpha\ge 0}\left\{-\alpha\log\left(\mathbb{E}_{\nu^0_{s,a}}\left[e^{-r(s,a)/\alpha}\right]\right) - \alpha\delta\right\} + \gamma\cdot\sup_{\beta\ge 0}\left\{-\beta\log\left(\sum_{s'\in\mathcal{S}} p^0_{s,a}(s') e^{-V_\delta^{\mathrm{rob},*}(s')/\beta}\right) - \beta\delta\right\}\right\}. \tag{5}$$

3.2 Distributionally Robust Policy Evaluation

In this section, we aim to develop a policy evaluation algorithm that address the question **Q1** in section 2.2. Given the batch dataset $\{(s_i,a_i,r_i,s_i')\}_{i=1}^n$, let $n(s,a) := |\{i\in[n] : (s_i,a_i) = (s,a)\}|$, for $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$, we can define the empirical estimators of transitions and rewards via

$$\widehat{p}_{s,a}(\cdot) = \frac{|\{i : (s_i,a_i,s_i') = (s,a,\cdot)\}|}{n(s,a)},$$

$$\widehat{\nu}_{s,a}(dx) = \frac{1}{n(s,a)}\sum_{j\in\{i:(s_i,a_i)=(s,a)\}} \mathbb{1}(r_j\in dx).$$

By the strong duality result in section 3.1, we transform the primal infinite-dimensional robust value optimization problem into its one-dimensional concave dual, and the expectation in the dual formulation (equation 4) is taken with respect to p^0 and ν^0, i.e., the transition probabilities and rewards under the original environment. Therefore, we can approximate the distributionally robust value function $V_\delta^{\mathrm{rob},\pi}$ by a plug-in estimation $\widehat{V}_\delta^{\mathrm{rob},\pi}$ that satisfies the following equation:

$$\widehat{V}_\delta^{\mathrm{rob},\pi}(s) = \underbrace{\sup_{\alpha\ge 0}\left\{-\alpha\log\left(\mathbb{E}_{\widehat{\nu}_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right]\right) - \alpha\delta\right\}}_{\widehat{R}_\delta^{\mathrm{rob}}(s,\pi(s))} + \underbrace{\gamma\sup_{\beta\ge 0}\left\{-\beta\log\left(\sum_{s'\in\mathcal{S}} \widehat{p}_{s,\pi(s)}(s') e^{-\widehat{V}_\delta^{\mathrm{rob},\pi}(s')/\beta}\right) - \beta\delta\right\}}_{\widehat{T}_\delta^{\mathrm{rob},\pi}(s,\pi(s))}. \tag{6}$$

The empirical distributionally robust optimal value function under the estimated model is therefore defined by $\widehat{V}_\delta^{\mathrm{rob},*} = \max_{\pi\in\Pi} \widehat{V}_\delta^{\mathrm{rob},\pi}$. Moreover $\widehat{\pi}_\delta^{\mathrm{rob},*} \in \arg\max_\pi \widehat{V}_\delta^{\mathrm{rob},\pi}$ stands for the empirical distributionally robust optimal policy. Note that the right hand side of equation 6 is a linear combination of two simple concave optimization problems. Those one dimensional concave optimization problems can be resolved efficiently by Newton-Raphson method with convergence guarantee (Chapter 8, Luenberger et al. (1984)). We can therefore compute the plug-in estimation $\widehat{V}_\delta^{\mathrm{rob},\pi}$ efficiently by combining value iteration with concave optimization. The formal description of our distributionally robust policy evaluation algorithm is stated in Algorithm 1.

Algorithm 1: Distributionally Robust Policy Evaluation

Input: Dataset $\{(s_i, a_i, r_i, s_i')\}_{i=1}^n$, policy $\pi \in \Pi$, uncertainty radius $\delta > 0$.

Output: Estimator of the empirical distributionally robust value function $\widehat{V}_\delta^{\mathrm{rob},\pi}$.

initialization: $\widehat{V}_\delta^{\mathrm{rob},\pi} \equiv 0$, compute $\{\widehat{p}_{s,a}\}$ and $\{\widehat{\nu}_{s,a}\}$ base on the data set,

$\widehat{R}_\delta^{\mathrm{rob}}(s,\pi(s)) \leftarrow$
$\max_{\alpha\geq 0}\left\{-\alpha\log\left(\mathbb{E}_{\widehat{\nu}_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right]\right) - \alpha\delta\right\}$.

repeat

$\widehat{T}_\delta^{\mathrm{rob}}(s,\pi(s)) \leftarrow$
$\max_{\beta\geq 0}\left\{-\beta\log\left(\sum_{s'\in\mathcal{S}} \widehat{p}_{s,\pi(s)}(s') e^{-\widehat{V}_\delta^{\mathrm{rob},\pi}(s')/\beta}\right) - \beta\delta\right\}$.

$\widehat{V}_\delta^{\mathrm{rob},\pi}(s) \leftarrow \widehat{R}_\delta^{\mathrm{rob}}(s,\pi(s)) + \gamma\cdot\widehat{T}_\delta^{\mathrm{rob}}(s,\pi(s))$.

until $\widehat{V}_\delta^{\mathrm{rob},\pi}$ *converges (with respect to the* L^∞ *norm)*;

Output: $\widehat{V}_\delta^{\mathrm{rob},\pi}$.

3.3 Policy Learning via Distributionally Robust Value Iteration

In this section, with the purpose of answering the question **Q2** in section 2.2, we investigate the distributionally robust policy learning by proposing efficient algorithm to compute the distributionally robust policy that is learned from the accessible data. In principle, by the empirical version of distributionally robust Bellman's equation (5), the empirical optimal robust policy can be computed by:

$$\begin{aligned}&\widehat{\pi}_\delta^{\mathrm{rob},*}(s)\\ &\in \arg\max_{a\in\mathcal{A}}\left\{\sup_{\alpha\geq 0}\left\{-\alpha\log\left(\mathbb{E}_{\widehat{\nu}_{s,a}}\left[e^{-r(s,a)/\alpha}\right]\right) - \alpha\delta\right\}\right.\\ &\quad\left.+\gamma\sup_{\beta\geq 0}\left\{-\beta\log\left(\sum_{s'\in\mathcal{S}}\widehat{p}_{s,a}(s')e^{-\widehat{V}_\delta^{\mathrm{rob},*}(s')/\beta}\right) - \beta\delta\right\}\right\}.\end{aligned}$$

However, it is intractable to compute the above problem exactly as it is in general non-concave. To overcome this difficulty, we propose a novel approximation scheme (Algorithm 2) that alternatively maximize the dual variables and iterate the value function. Note that the optimization problem with respect to the dual variables is a one dimensional concave problem, and our numerical experiments suggest that only a few Newton-Raphson iterations are sufficient for the convergence of the dual variables.

Algorithm 2: Policy Learning via Distributionally Robust Value Iteration

Input: Dataset $\{(s_i, a_i, r_i, s_i')\}_{i=1}^n$, uncertainty radius $\delta > 0$.

Output: Distributionally Robust empirical optimal policy $\widehat{\pi}_\delta$.

initialization: $\widehat{V}_\delta^{\mathrm{rob}} \equiv 0$, compute $\{\widehat{p}_{s,a}\}$ and $\{\widehat{\nu}_{s,a}\}$ based on dataset.

$\widehat{R}_\delta^{\mathrm{rob}}(s,a) \leftarrow$
$\max_{\alpha\geq 0}\left\{-\alpha\log\left(\mathbb{E}_{\widehat{\nu}_{s,a}}\left[e^{-r(s,a)/\alpha}\right]\right) - \alpha\delta\right\}$.

repeat

$\widehat{T}_\delta^{\mathrm{rob}}(s,a) \leftarrow$
$\max_{\beta\geq 0}\left\{-\beta\log\left(\sum_{s'\in\mathcal{S}}\widehat{p}_{s,a}(s')e^{-\widehat{V}_\delta^{\mathrm{rob}}(s')/\beta}\right) - \beta\delta\right\}$.

$\widehat{V}_\delta^{\mathrm{rob}}(s) \leftarrow \max_{a\in\mathcal{A}}\left\{\widehat{R}_\delta^{\mathrm{rob}}(s,a) + \gamma\cdot\widehat{T}_\delta^{\mathrm{rob}}(s,a)\right\}$.

until $\widehat{V}_\delta^{\mathrm{rob}}$ *converges (with respect to the* L^∞ *norm)*;

Output:

$\widehat{\pi}_\delta(s) = \arg\max_{a\in\mathcal{A}}\left\{\widehat{R}_\delta^{\mathrm{rob}}(s,a) + \gamma\cdot\widehat{T}_\delta^{\mathrm{rob}}(s,a)\right\}$.

3.4 Theoretical Guarantee

First of all, the consistency of the output in Algorithm 1 is shown in the following proposition.

Proposition 1. *Let* $\widehat{V}_{\delta,k}^{\mathrm{rob},\pi}$ *denotes the distributionally robust value function after* k *iterations in Algorithm 1. For any initialization* $\widehat{V}_{\delta,0}^{\mathrm{rob},\pi}$*, we have* $\widehat{V}_{\delta,k}^{\mathrm{rob},\pi}(s) \to \widehat{V}_\delta^{\mathrm{rob},\pi}(s)$*, for all* $s\in\mathcal{S}$.

Proof. Note that in Algorithm 1, we have

$$\widehat{V}_{\delta,k+1}^{\mathrm{rob},\pi}(s) = \inf_{\substack{p_{s,\pi(s)}\in\widehat{\mathcal{P}}_{s,\pi(s)}(\delta),\\ r\in\widehat{\mathcal{R}}_{s,\pi(s)}(\delta)}}\left\{\mathbb{E}[r(s,\pi(s))] + \gamma\sum_{s'\in\mathcal{S}} p_{s,\pi(s)}(s')\widehat{V}_{\delta,k}^{\mathrm{rob},\pi}(s')\right\}.$$

Since the infimum operator is 1-Lipschitz, we have $\|\widehat{V}_{\delta,k+1}^{\mathrm{rob},\pi} - \widehat{V}_\delta^{\mathrm{rob},\pi}\|_\infty \leq \gamma\|\widehat{V}_{\delta,k}^{\mathrm{rob},\pi} - \widehat{V}_\delta^{\mathrm{rob},\pi}\|_\infty$. Hence $\|\widehat{V}_{\delta,k}^{\mathrm{rob},\pi} - \widehat{V}_\delta^{\mathrm{rob},\pi}\|_\infty \leq \gamma^k\|\widehat{V}_{\delta,0}^{\mathrm{rob},\pi} - \widehat{V}_\delta^{\mathrm{rob},\pi}\|_\infty \to 0$ as k goes to infinity. □

Next, we confirm that for any fixed policy $\pi \in \Pi$, the gap between the optimal distributionally robust value function $V_\delta^{\mathrm{rob},\pi}$ (equation 3) and empirical distributionally robust value function $\widehat{V}_\delta^{\mathrm{rob},\pi}$ (equation 6) is $O_P(1/\sqrt{n})$, which together with Algorithm 1 that complete the answer of **Q1** in section 2.2.

Theorem 1. *Given $\delta > 0$ and any fixed policy $\pi \in \Pi$. If Assumption 1, 2 and 3 hold, there exists a constant $N_\pi := N\left(\varepsilon, \delta, \nu^0, \mathcal{P}^0, \pi\right)$, such that for $n \geq N_\pi$, with probability at least $1-\varepsilon$,*

$$\|\widehat{V}_\delta^{\mathrm{rob},\pi} - V_\delta^{\mathrm{rob},\pi}\|_\infty \leq \frac{C_\pi \sqrt{\frac{|\mathcal{S}|}{\eta} \log\left(\frac{|\mathcal{S}|}{\varepsilon}\right)}}{(1-\gamma)\delta} \frac{1}{\sqrt{n}},$$

where C_π is a constant depending on $\nu^0, \mathcal{P}^0$ (defined in section 2.2) and the policy π.

The proof details of Theorem 1 are deferred to the supplementary material.

Finally, We are able to establish the finite-sample guarantee for the distributionally robust empirical optimal policy $\widehat{\pi}_\delta^*$ that is learned from Algorithm 2. To be precise, we utilize the idea of distributionally robust regret (see Definition 2) to quantify how well is our policy $\widehat{\pi}_\delta^*$ compare to the best possible distributionally robust policy (see equation 2). The following high probability bound illustrates that the regret scales at the rate of $O_P\left(1/\sqrt{n}\right)$, which answer the question **Q2** in section 2.2.

Theorem 2. *Given $\delta > 0$. If Assumption 1, 2 and 3 hold, there exists a constant $N := N\left(\varepsilon, \delta, \nu^0, \mathcal{P}^0\right)$, such that for $n \geq N$, with probability at least $1-\varepsilon$,*

$$R^{\mathrm{rob}}\left(\widehat{\pi}_\delta^{\mathrm{rob},*}\right) \leq \frac{C\sqrt{\frac{|\mathcal{S}|^2}{\eta} \log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon}\right)}}{(1-\gamma)\delta} \frac{1}{\sqrt{n}},$$

where the constant C is a constant depending on ν^0 and $\mathcal{P}^0$ (defined in section 2.2).

The proof details of Theorem 2 are deferred to the supplementary material.

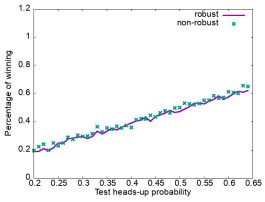
4 Numerical Results

To numerically evaluate the performance of our proposed policy learning algorithm, in the following subsections, we first evaluate it on a variant of the Gambler's problem, as a sanity check. Then we further evaluate our algorithm on a more realistic simulated problem of American options. In all experiments, we compare our proposed distributionally robust Algorithm 2 with its non-robust counterpart, i.e., standard value iteration algorithm.

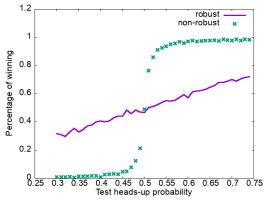
4.1 Gambler's Problem

Gambler's Problem (Sutton and Barto, 2018) A gambler makes bets on the outcomes of a sequence of coin flips, winning his stake with heads and losing with tails. The game ends when the gambler reaches \$100 or losses all the money. A gambler's policy outputs on each flip a portion of his capital to stake, in integer number of dollars. This problem can be formulated as an undiscounted, episodic, finite MDP. The state is the gambler's remaining capital, $s \in \{1, 2, \ldots, 99\}$ and actions are stakes, $a \in \{0, 1, \ldots, \min(s, 100-s)\}$. The reward is zero on all transitions except those reaching the \$100 goal, when it is +1. The state-value function then gives the probability of winning from each state. The optimal policy maximizes the probability of reaching the goal. When the probability p_h of the coin coming up heads is known, then the optimal policy can be solved by value iteration.

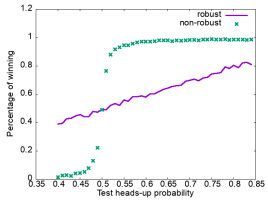
To incorporate distributionally robust setup, we consider a variant of the Gambler's problem where the actual p_h for testing the learned policy lies in a KL-ball of radius 0.1 centered at the p_h^0 provided to the policy learning algorithms. For $p_h^0 = 0.4, 0.5, 0.6$ respectively, we run both standard value iteration and our distributionally robust value iteration, then test the obtained policies under different perturbed p_h's.



(a) $p_h^0 = 0.4$



(b) $p_h^0 = 0.5$



(c) $p_h^0 = 0.6$

Figure 1: Performance of robust v.s. regular policies in the Gambler's problem. The x-axis is the heads up probability of the testing games, and the y-axis is the actual winning percentage obtained by playing $1,000$ games under the same p_h.

As shown in Figure 1, both robust and non-robust policies perform similarly when $p_h^0 = 0.4$ is provided, this is because when true p_h is less than 0.5, there is a family of optimal policies which lead to a winning probability equal to p_h. Given the provided p_h^0 is less

than 0.5, both standard and distributionally robust value iterations find such an optimal policy. When $p_h^0 \geq 0.5$, however, the performance of robust and non-robust policies are different. Note that when true p_h is greater or equal to 0.5, there is an unique optimal policy: put \$1 stake for each flip. Then it is as expected that the standard value iteration always learns this "optimal" policy when $p_h^0 = 0.5, 0.6$ is provided, which performs poorly whenever the actual p_h is lower than 0.5. On the other hand, our distributionally robust value iteration outputs the worst case (within the perturbed ball of given p_h^0) optimal policy. Given the preset radius of perturbations, when $p_h^0 = 0.5, 0.6$ is provided, our algorithm still outputs one of the optimal policies for $p_h < 0.5$ as a worst case optimum. As a result, the non-robust policy performs optimally when $p_h \geq 0.5$ but poorly when $p_h < 0.5$, while our robust policy performs more conservatively when $p_h \geq 0.5$ but much better in worst cases when $p_h < 0.5$. This example demonstrates the robustness of our proposed distributionally robust formulations in perturbed testing environment.

We further illustrate how distributionally robust regret scales when the sample size increases for the Gambler's problem. For each sample size n, we repeat the experiment 10 times and compute the average distributionally robust regret. See Figure 2.

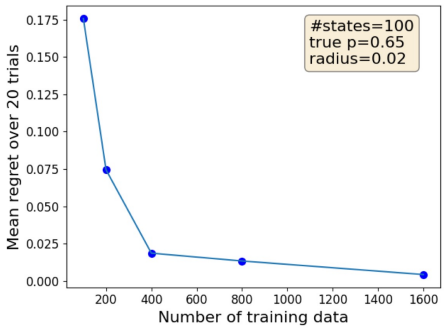


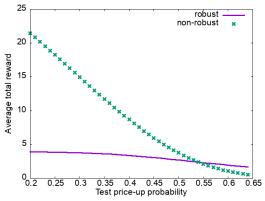
Figure 2: Regret v.s. sample size for the Gambler's problem.

4.2 American Options

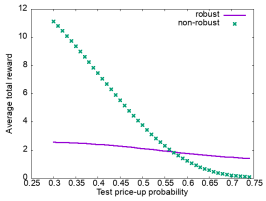
We also evaluate our algorithm in a simulated American put option problem, where the price fluctuation model follows the following Bernoulli distribution,

$$s_{t+1} = \begin{cases} c_u s_t, & \text{w.p. } p_u \\ c_d s_t, & \text{w.p. } 1 - p_u \end{cases}$$

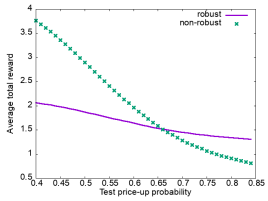
where c_u and c_d are constant price up and down factors, and p_u is the probability that the price goes up. At each time step, one can take an action of either exercising or not exercising the option. When not exercising at time t, the reward is 0, and the next state is s_{t+1} based on the price fluctuation model above; when exercising at time t, the reward is $\max(0, K - s_t)$, where K is the strike price, and the next state is the exit state, which is an absorbing state. In each experiment, we choose a value of the price up probability p_u^0 and generate 10000 trajectories of $T+1$ time steps for policy learning, where we fix $T = 20$, $K = 100$, $c_u = 1.02$, $c_d = 0.98$, and $s_0 = K + \epsilon$ with ϵ being a random number in the range of $[-5, 5]$ in all our simulations. In contrast to the Gambler's problem, p_u^0 is unknown to the policy learning algorithm and an empirical estimation $\hat{p}_u^0$ is used for all competing algorithms. To test a learned policy, we evaluate the total reward under different price fluctuation models with different price up probabilities p_u diverging from p_u^0 up to about 0.1 in KL divergence. For each testing price model, we report the total reward averaged over $1,000,000$ runs. Figure 3 shows the test results for the policy obtained by our distributionally robust value iteration v.s. that obtained by standard value iteration. It can be observed that the averaged total reward from our robust policy is more stable over different testing price models for all data generating p_u^0, in particular, it performs considerably better under worst-cases then the actual price-up probability is much higher than the data generating p_u^0. Such risk-averse behavior is consistent with our theoretical results and demonstrates the effectiveness of the proposed distributionally robust formulations.



(a) $p_u^0 = 0.5$



(b) $p_u^0 = 0.6$

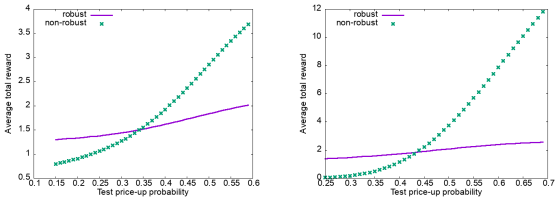


(c) $p_u^0 = 0.7$

Figure 3: Performance of robust v.s. regular policies in the American put option simulation. The x-axis is the probability of the test prices going up, and the y-axis is the average total reward.

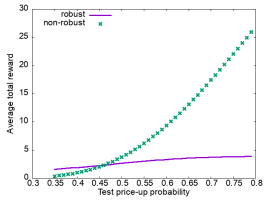
We further test our algorithm on a simulated American call option problem using the same parameters

as in the above put option problem except that the reward function when exercising the option at time step t is changed to $\max(0, s_t - K)$. Test results are shown in Figure 4, where it can be observed, again, that the averaged total reward resulting from our robust policy is more stable than that from the standard value iteration policy. With varying test probabilities, our algorithm performs considerably better in worst cases when the actual price-up probability is much lower than the training data generating probability p_u^0, demonstrating again the effectiveness of our proposed distributionally robust formulations.



(a) $p_u^0 = 0.4$

(b) $p_u^0 = 0.5$



(c) $p_u^0 = 0.6$

Figure 4: Performance of robust v.s. regular policies in the American call option simulation. The x-axis is the probability of the test prices going up, and the y-axis is the average total reward.

To demonstrate the scalability of our algorithms, we extend the American put option example to simulate the trading of a collection of different put options with the same stopping time, where the underlying assets are different stocks with different price transition probabilities. The reward will be the weighted sum of the payoffs of all the options. In this example, the size of the state space grows exponentially when the number of stocks increases. When simulating a collection of 4 stocks for a maximum execution time of 20, with price ranges of about 100 and price precision of 2 decimal places, the state space size grows to over 10^{17}, where we use a simple state aggregation function approximation to reduce the computational complexity of the problem. The results are shown in Table 1, where R_{rob} denotes the average reward obtained using our robust policy, and $R_{non-rob}$ denotes the average reward obtained using the policy from standard value iteration. We use different price transition probabilities for different stocks to generate the data samples, and we test the policies in an environment where the price up probabilities are higher than those in the training environment to demonstrate the risk-averse behavior of the robust policies, with KL divergence of 0.08, 0.16, 0.25, 0.34, respectively, for each N_{stocks} example.

N_{stocks}	1	2	3	4
N_{states}	$2 \cdot 10^5$	$2 \cdot 10^9$	$2 \cdot 10^{13}$	$2 \cdot 10^{17}$
R_{rob}	1.38	1.37	1.34	1.31
$R_{non-rob}$	0.13	0.64	0.47	0.64

Table 1: Trading a collection of American put options. The reward is computed as the average payoff of all the options.

5 Conclusion

First, we have provided a distributionally robust formulation for offline policy learning in RL. Second, we proposed a novel distributionally robust policy learning algorithm that is able to learn a robust policy under adversarial perturbation of transition probabilities and rewards. Third, we established the first finite sample guarantee on the distributionally robust regret of the policy that is learned by our algorithm. There are still many interesting directions that are worth further exploration. For instance, generalization of our algorithm and theory to the non-tabular setting, which is extremely challenging and require significant improvements to the current techniques. Another interesting direction would be to extend our results to the Wasserstein perturbation, which is entirely different from the current KL-divergence based framework and it is more flexible in term of the support of the transition model.

6 Acknowledgement

Material in this paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397. Additional support is gratefully acknowledged from NSF grants 1915967, 1820942 ,1838576 and from the China Merchant Bank.

References

S. Athey and S. Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

T. Başar and P. Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

D. P. Bertsekas. Dynamic programming and optimal control 3rd edition, volume ii. *Belmont, MA: Athena Scientific*, 2011.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming.*, volume 3 of *Optimization and neural computation series*. Athena Scientific, 1996.

D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

J. Blanchet and Y. Kang. Semi-supervised learning based on distributionally robust optimization. *Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods*, 5:1–33, 2020.

J. Blanchet, Y. Kang, and K. Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3): 830–857, 2019.

J. Chen and N. Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

J. J. Choi, D. Laibson, B. C. Madrian, and A. Metrick. Reinforcement learning and savings behavior. *The Journal of finance*, 64(6):2515–2534, 2009.

Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3): 653–664, 2017.

K. Drew. California robot teaching itself to walk like a human toddler. *NBC News*, Dec 2015.

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

J. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

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

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

P. Dupuis, M. R. James, and I. Petersen. Robust properties of risk-sensitive control. *Mathematics of Control, Signals and Systems*, 13(4):318–332, 2000.

P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2): 115–166, 2018.

N. Fournier and A. Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

S. Fujimoto, D. Meger, and D. Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062, 2019.

R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv: Optimization and Control*, 2016.

S. Gu, E. Holly, T. Lillicrap, and S. Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE, 2017.

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

C. Huang, S. Lucey, and D. Ramanan. Learning policies for adaptive tracking with deep feature cascades. *ICCV*, pages 105–114, 2017.

G. Iyengar. Robust dynamic programming. *Math. Oper. Res.*, 30:257–280, 05 2005.

N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

N. Kallus. Balanced policy evaluation and learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8895–8906. Curran Associates, Inc., 2018.

N. Kallus and A. Zhou. Confounding-robust policy improvement. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9269–9279. Curran Associates, Inc., 2018.

J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

A. Lazaric, M. Ghavamzadeh, and R. Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13(98): 3041–3074, 2012.

Y. Li, C. Szepesvari, and D. Schuurmans. Learning exercise policies for american options. In *Artificial Intelligence and Statistics*, pages 352–359, 2009.

D. G. Luenberger, Y. Ye, et al. *Linear and nonlinear programming*, volume 2. Springer, 1984.

A. R. Mahmood, H. P. van Hasselt, and R. S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 3014–3022, 2014.

J. Maitin-Shepard, M. Cusumano-Towner, J. Lei, and P. Abbeel. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In *2010 IEEE International Conference on Robotics and Automation*, pages 2308–2315, 2010.

J. Morimoto and K. Doya. Robust reinforcement learning. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 1061–1067. MIT Press, 2001.

R. Munos, T. Stepleton, A. Harutyunyan, and M. Bellemare. Safe and efficient off-policy reinforcement learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1054–1062. Curran Associates, Inc., 2016.

Y. Nevmyvaka, Y. Feng, and M. Kearns. Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd international conference on Machine learning*, pages 673–680, 2006.

I. R. Petersen, M. R. James, and P. Dupuis. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control*, 45(3):398–412, 2000.

W. B. Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2007.

F. Sadeghi and S. Levine. Cad2rl: Real single-image flight without a single real image. *arXiv preprint arXiv:1611.04201*, 2016.

J. Schulman, J. Ho, A. X. Lee, I. Awwal, H. Bradlow, and P. Abbeel. Finding locally optimal, collision-free trajectories with sequential convex optimization. In *Robotics: science and systems*, volume 9, pages 1–10. Citeseer, 2013.

N. Si, F. Zhang, Z. Zhou, and J. Blanchet. Distributionally robust policy evaluation and learning in offline contextual bandits. In *International Conference on Machine Learning (ICML)*, 2020.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

C. Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

V. A. Ugrinovskii. Robust h infinity control in the presence of stochastic uncertainty. *International Journal of Control*, 71(2):219–237, 1998.

W. Wiesemann, D. Kuhn, and B. Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

E. M. Wolff, U. Topcu, and R. M. Murray. Robust control of uncertain markov decision processes with temporal logic specifications. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 3372–3379, 2012.

L. Xie and C. E. de Souza. Robust h infinity control for linear systems with norm-bounded time-varying uncertainty. In *29th IEEE Conference on Decision and Control*, pages 1034–1035. IEEE, 1990.

H. Xu and S. Mannor. Distributionally robust markov decision processes. In *Advances in Neural Information Processing Systems*, pages 2505–2513, 2010.

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

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

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

Supplementary Materials

7 Technical Lemmas

First of all, we introduce two ancillary concentration inequalities.

Lemma 2 (Fournier and Guillin (2015), Concentration inequality for Wasserstein distance). *For $\mu \in \mathcal{P}(\mathbb{R})$, we consider an i.i.d. sequence $(X_k)_{k\geq 1}$ of μ-distributed random variables and, for all $n \geq 1$, the empirical measure*

$$\mu_n := \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k}.$$

Assume that there exists $\gamma > 0$ such that $\mathcal{E}_{2,\gamma}(\mu) := \int_{\mathbb{R}} \exp(\gamma|x|^2)\mu(dx) < \infty$. Then for all $n \geq 1$, all $x > 0$,

$$\mathbb{P}\left(\mathcal{W}(\mu_n, \mu) \geq x\right) \leq C \exp\left(-cnx^2\right),$$

where the Wasserstein distance $\mathcal{W}(\mu_n, \mu)$ is defined by

$$\mathcal{W}(\mu_n, \mu) := \inf_{\pi \in \Pi(\mu_n, \mu)} \left\{ \int |x - y| \pi(dx, dy) \right\},$$

and the positive constant C and c depends only on γ and $\mathcal{E}_{2,\gamma}(\mu)$.

Lemma 3 (Hoeffding's inequality). *Let $X_1, \cdots, X_n$ be independent random variables such that $X_i \in [a_i, b_i]$ almost surely for all $i = 1, 2, \cdots, n$. Then for every $t > 0$,*

$$\mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} (X_i - \mathbb{E}X_i) \right| \geq t \right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2} \right).$$

Next, we state and prove the Lemma 4 that characterize the boundedness of the dual parameter of data-driven Distributionally Robust Optimization. It is the key technical step in the proofs of Lemma 5 and Theorem 1.

Lemma 4. *Let $X \sim P$ be a random variable with $X \in [0, M]$, and P_n denotes its empirical distribution of sample size n. For $\delta > 0$, for any*

$$\alpha^* \in \arg\max_{\alpha \geq 0} \left\{ -\alpha \log\left(\mathbb{E}_P \left[e^{-X/\alpha} \right] \right) - \alpha\delta \right\}, \tag{7}$$

(1) *$\alpha^* = 0$. Furthermore, assume that the support of X is finite. Then there exists a constant $N' := N'(\varepsilon, \delta, P)$, such that $n \geq N'$, with probability at least $1 - \varepsilon$, we have*

$$0 \in \arg\max_{\alpha \geq 0} \left\{ -\alpha \log\left(\mathbb{E}_{P_n} \left[e^{-X/\alpha} \right] \right) - \alpha\delta \right\}.$$

(2) *$\alpha^* > 0$. Then there exists a constant $N'' := N^{''}(\varepsilon, \delta, P)$, such that for any $n \geq N''$, with probability at least $1 - \varepsilon$, there exists a*

$$\widehat{\alpha}^* \in \arg\max_{\alpha \geq 0} \left\{ -\alpha \log\left(\mathbb{E}_{P_n} \left[e^{-X/\alpha} \right] \right) - \alpha\delta \right\},$$

such that $\alpha^*, \widehat{\alpha}^* \in [\underline{\alpha}, \bar{\alpha}]$, where $\underline{\alpha} > 0$ is independent of n and $\bar{\alpha} = M/\delta$.

Proof. First of all, note that $-\alpha \log\left(\mathbb{E}_P\left[e^{-X/\alpha}\right]\right) - \alpha\delta$ is a concave function of α, thus α^* is the solution of a concave optimization problem. Moreover, follows from simple calculus,

$$\lim_{\alpha\to 0}\left[-\alpha \log\left(\mathbb{E}_P\left[e^{-X/\alpha}\right]\right) - \alpha\delta\right] = \operatorname{ess\,inf} X,$$

where $\operatorname{ess\,inf} X$ is the essential infimum of X.

Case 1. Suppose $\alpha^* = 0$. Since $X \sim P$ is bounded, by Proposition 2 in Hu and Hong (2013), we have $\alpha^* = 0$ if and only if $\kappa := \mathbb{P}(X = \operatorname{ess\,inf} X) > 0$ and $\log\kappa + \delta \geq 0$. Since δ is chosen by us, we can ignore the edge case $\log(\kappa) + \delta = 0$ by introducing randomness on δ. Without loss of generality, we may assume that $\log(\kappa) + \delta > 0$.

Let $\mathcal{S}_X$ be the support of X, and p_x denotes the probability $\mathbb{P}(X = x)$ for all $x \in \mathcal{S}_X$. Consider $X_1, \cdots, X_n, \cdots$ i.i.d. $\sim P$, we have

$$\mathbb{P}\left(\min_{1\leq i\leq n} X_i \neq \operatorname{ess\,inf} X\right) \leq \left(1 - \min_{x\in\mathcal{S}_X} p_x\right)^n \leq \varepsilon/2,$$

whenever $n \geq \log\left(\frac{1}{1-\min_{x\in\mathcal{S}_X} p_x}\right)^{-1} \log\left(\frac{2}{\varepsilon}\right)$. Moreover, follows from the Hoeffding's inequality (Lemma 3), for $n \geq 2M^2 \log\left(\frac{4}{\varepsilon}\right) / \left(\kappa - e^{-\delta}\right)^2$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i = \operatorname{ess\,inf} X) - \kappa\right| \geq \frac{\kappa - e^{-\delta}}{2}\right) \leq 2e^{-2n\left(\frac{\kappa - e^{-\delta}}{2}\right)^2/M^2} \leq \varepsilon/2.$$

Define $\kappa_n := \frac{1}{n}\sum_{i=1}^n \mathbb{1}\left(X_i = \min_{1\leq i\leq n} X_i\right) > 0$. From the above results, with probability at least $1-\varepsilon$, we have

$$|\kappa_n - \kappa| < \frac{\kappa - e^{-\delta}}{2}.$$

as long as $n \geq N'(\varepsilon, \delta, P) := \log\left(\frac{1}{1-\min_{x\in\mathcal{S}_X} p_x}\right)^{-1} \log\left(\frac{2}{\varepsilon}\right) \vee 2M^2 \log\left(\frac{4}{\varepsilon}\right) / \left(\kappa - e^{-\delta}\right)^2$. Consequently,

$$\log\kappa_n \geq \log\left(\kappa - \frac{\kappa - e^{-\delta}}{2}\right) > -\delta,$$

which implies $\widehat{\alpha}_n^* = 0$.

Case 2. Suppose $\alpha^* > 0$. We first show that α^* is upper bounded by $\bar{\alpha} = M/\delta$. On one hand,

$$\sup_{\alpha\geq 0}\left\{-\alpha \log\left(\mathbb{E}_P\left[e^{-X/\alpha}\right]\right) - \alpha\delta\right\} \geq \lim_{\alpha\to 0}\left[-\alpha \log\left(\mathbb{E}_P\left[e^{-X/\alpha}\right]\right) - \alpha\delta\right] = \operatorname{ess\,inf} X \geq 0.$$

On the other hand,

$$-\alpha \log\left(\mathbb{E}_P\left[e^{-X/\alpha}\right]\right) - \alpha\delta \leq -\alpha \log\left(e^{-M/\alpha}\right) - \alpha\delta \leq M - \alpha\delta.$$

Hence the optimal α^* is bounded by M/δ. Note that if X is non-degenerate, the optimization problem in equation 7 admits unique solution. However, the problem is trivially optimized at $\alpha^* = 0$ when X is degenerate. Hence the α^* in Case 2 is unique. Given this observation, let

$$\tau := \min\left\{\underline{\alpha} \log\left(\mathbb{E}_P\left[e^{-X/\underline{\alpha}}\right]\right) + \underline{\alpha}\delta, \bar{\alpha} \log\left(\mathbb{E}_P\left[e^{-X/\bar{\alpha}}\right]\right) + \bar{\alpha}\delta\right\} - \left(\alpha^* \log\left(\mathbb{E}_P\left[e^{-X/\alpha^*}\right]\right) + \alpha^*\delta\right),$$

where $\underline{\alpha} := \alpha^*/2$. Then we have $\tau > 0$. Given $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, we have

$$\begin{aligned}\left|\alpha \log\left(\mathbb{E}_{P_n}\left[e^{-X/\alpha}\right]\right) + \alpha\delta - \left[\alpha \log\left(\mathbb{E}_P\left[e^{-X/\alpha}\right]\right) + \alpha\delta\right]\right| &= \alpha\left|\log\left(1 + \frac{\mathbb{E}_{P_n}\left[e^{-X/\alpha}\right] - \mathbb{E}_P\left[e^{-X/\alpha}\right]}{\mathbb{E}_P\left[e^{-X/\alpha}\right]}\right)\right| \\ &\leq 2\alpha e^{M/\alpha}\left|\mathbb{E}_{P_n}\left[e^{-X/\alpha}\right] - \mathbb{E}_P\left[e^{-X/\alpha}\right]\right|.\end{aligned}$$

where the last inequality follows from the fact that $|\log(1+x)| \leq 2|x|$ when $|x| \leq 1/2$. Now by Hoeffding's inequality (Lemma 3),

$$\mathbb{P}\left(\left|\mathbb{E}_{P_n}\left[e^{-X/\alpha}\right] - \mathbb{E}_P\left[e^{-X/\alpha}\right]\right| \geq \frac{\tau}{2}\left(2\alpha e^{M/\alpha}\right)^{-1}\right) \leq 2e^{-n\tau^2/(8M^2e^{2M/\alpha})}.$$

Therefore, for $n \geq N''(\varepsilon, \delta, P) := \max_{\alpha \in \{\underline{\alpha}, \alpha^*, \bar{\alpha}\}} \left\{ \frac{8M^2 e^{2M/\alpha}}{\tau^2} \log\left(\frac{6}{\varepsilon}\right) \right\}$, with probability at least $1-\varepsilon$,

$$\max_{\alpha \in \{\underline{\alpha}, \alpha^*, \bar{\alpha}\}} \left| \alpha \log\left(\mathbb{E}_{P_n}\left[e^{-X/\alpha}\right]\right) + \alpha\delta - \left[\alpha \log\left(\mathbb{E}_{P}\left[e^{-X/\alpha}\right]\right) + \alpha\delta\right] \right| \leq \tau/2.$$

Hence,

$$\begin{aligned}
& \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left\{ -\alpha \log\left(\mathbb{E}_{P_n}\left[e^{-X/\alpha}\right]\right) - \alpha\delta \right\} \\
\geq & -\alpha^* \log\left(\mathbb{E}_{P_n}\left[e^{-X/\alpha^*}\right]\right) - \alpha^*\delta \\
\geq & -\alpha^* \log\left(\mathbb{E}_{P}\left[e^{-X/\alpha^*}\right]\right) - \alpha^*\delta - \tau/2 \\
\geq & \max\left\{ -\underline{\alpha} \log\left(\mathbb{E}_{P}\left[e^{-X/\underline{\alpha}}\right]\right) - \underline{\alpha}\delta, -\bar{\alpha} \log\left(\mathbb{E}_{P}\left[e^{-X/\bar{\alpha}}\right]\right) - \bar{\alpha}\delta \right\} + \tau/2 \\
\geq & \max\left\{ -\underline{\alpha} \log\left(\mathbb{E}_{P_n}\left[e^{-X/\underline{\alpha}}\right]\right) - \underline{\alpha}\delta, -\bar{\alpha} \log\left(\mathbb{E}_{P_n}\left[e^{-X/\bar{\alpha}}\right]\right) - \bar{\alpha}\delta \right\}.
\end{aligned}$$

Follows from the concavity of $-\alpha \log\left(\mathbb{E}_{P_n}\left[e^{-X/\alpha}\right]\right) - \alpha\delta$ (with respect to α) that $\widehat{\alpha}_n^* \in [\underline{\alpha}, \bar{\alpha}]$. □

8 Lemma 5

The following lemma provides a sample complexity bound on the gap between the robust reward $R_\delta^{\mathrm{rob}}(s, \pi(s))$ and empirical robust reward $\widehat{R}_\delta^{\mathrm{rob}}(s, \pi(s))$.

Lemma 5. *Given $\delta > 0$ and a policy $\pi \in \Pi$. For any $\varepsilon > 0$, there exists a constant $N := N(\varepsilon, \delta, \nu^0_{s,\pi(s)})$, such that for any $n > N$, with probability at least $1-\varepsilon$, we have*

$$\left| R_\delta^{\mathrm{rob}}(s, \pi(s)) - \widehat{R}_\delta^{\mathrm{rob}}(s, \pi(s)) \right| \leq \frac{2R_{\max} \exp\left(R_{\max}/\underline{\alpha}\right)}{\delta \underline{\alpha}} \sqrt{\frac{1}{c} \log\left(\frac{2C}{\varepsilon}\right)} \frac{1}{n(s, \pi(s))},$$

where the positive constant $\underline{\alpha}$ is presented in Lemma 4, and both C and c are constants determined by Lemma 2.

Proof. Recalled from the proof of Lemma 4 that $\alpha_s^* = 0$ implies $\mathbb{P}_{r(s,\pi(s)) \sim \nu^0_{s,\pi(s)}}(r(s, \pi(s)) = \operatorname{ess\,inf} r(s, \pi(s))) > 0$. Hence, as Assumption 1 is enforced, we have $\alpha_s^* > 0$. By Lemma 4, we know that for $n \geq N''(\varepsilon/2, \delta, \nu^0_{s,\pi(s)})$, with probability at least $1-\varepsilon/2$, there exists optimal α_s^* and $\widehat{\alpha}_s^*$ that are contained in an interval $[\underline{\alpha}, \bar{\alpha}]$ bounded away from zero. Here $\bar{\alpha} = R_{\max}/\delta$, and $\underline{\alpha}$ is a positive constant depends on $\nu^0_{s,\pi(s)}$ and δ. Hence

$$\begin{aligned}
& \left| R_\delta^{\mathrm{rob}}(s, \pi(s)) - \widehat{R}_\delta^{\mathrm{rob}}(s, \pi(s)) \right| \\
\leq & \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left\{ \left[-\alpha \log\left(\mathbb{E}_{\nu^0_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right]\right) - \alpha\delta \right] - \left[-\alpha \log\left(\mathbb{E}_{\widehat{\nu}_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right]\right) - \alpha\delta \right] \right\} \\
\leq & \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left\{ \left| \alpha \log\left(1 + \frac{\mathbb{E}_{\nu^0_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right] - \mathbb{E}_{\widehat{\nu}_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right]}{\mathbb{E}_{\widehat{\nu}_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right]} \right) \right| \right\} \\
\leq & \frac{2R_{\max}}{\delta} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left\{ \frac{\left| \mathbb{E}_{\nu^0_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right] - \mathbb{E}_{\widehat{\nu}_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right] \right|}{\mathbb{E}_{\widehat{\nu}_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right]} \right\} \\
\leq & \frac{2R_{\max}}{\delta e^{-R_{\max}/\underline{\alpha}}} \sup_{\alpha \in [\underline{\alpha}, \bar{\alpha}]} \left\{ \left| \mathbb{E}_{\nu^0_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right] - \mathbb{E}_{\widehat{\nu}_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right] \right| \right\}
\end{aligned}$$

where we have used the fact that $|\ln(1+x)| \leq 2|x|$ for $|x| \leq 1/2$. For any $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, the function $e^{-x/\alpha}$ is a Lipschitz function on $[0, \infty)$, and its Lipschitz constant is bounded by $1/\underline{\alpha}$. Hence, by Lemma 2 and the dual

representation of Wasserstein distance, with probability at least $1-\varepsilon/2$,

$$\sup_{\alpha\in[\underline{\alpha},\bar{\alpha}]}\left\{\left|\mathbb{E}_{\nu^0_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right]-\mathbb{E}_{\widehat{\nu}^{(n)}_{s,\pi(s)}}\left[e^{-r(s,\pi(s))/\alpha}\right]\right|\right\}\le\frac{1}{\underline{\alpha}}\mathcal{W}\left(\nu^0_{s,\pi(s)},\widehat{\nu}^{(n)}_{s,\pi(s)}\right)$$
$$\le\frac{1}{\underline{\alpha}}\sqrt{\frac{1}{c}\log\left(\frac{2C}{\varepsilon}\right)}\frac{1}{\sqrt{n(s,\pi(s))}}.$$

□

9 Proof of Theorem 1

Proof. By Assumption 3 and the sampling scheme, the expected value $\mathbb{E}[n(s,a)]$ is bounded below by $\eta n/|\mathcal{S}|$. By Hoeffding's inequality (Lemma 3), for $n\ge N_1:=\frac{2|\mathcal{S}|^2}{\eta^2}\log\left(\frac{4|\mathcal{S}|}{\varepsilon}\right)$, with probability at least $1-\varepsilon/2$,

$$n(s,\pi(s))\ge\frac{\eta}{2|\mathcal{S}|}\cdot n,\quad\forall s\in\mathcal{S}.$$

Observe that

$$\beta\log\left(\sum_{s'\in\mathcal{S}}\widehat{p}_{s,\pi(s)}(s')e^{-\widehat{V}^{\mathrm{rob},\pi}(s')/\beta}\right)\le\beta\log\left(\sum_{s'\in\mathcal{S}}\widehat{p}_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}e^{\|\widehat{V}^{\mathrm{rob},\pi}-V^{\mathrm{rob},\pi}\|_\infty/\beta}\right)$$
$$=\|\widehat{V}^{\mathrm{rob},\pi}-V^{\mathrm{rob},\pi}\|_\infty+\beta\log\left(\sum_{s'\in\mathcal{S}}\widehat{p}_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}\right),$$

and similarly, we have

$$\beta\log\left(\sum_{s'\in\mathcal{S}}\widehat{p}_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}\right)\le\|\widehat{V}^{\mathrm{rob},\pi}-V^{\mathrm{rob},\pi}\|_\infty+\beta\log\left(\sum_{s'\in\mathcal{S}}\widehat{p}_{s,\pi(s)}(s')e^{-\widehat{V}^{\mathrm{rob},\pi}(s')/\beta}\right).$$

Combing the above facts with equation 5, equation 6 and Lemma 5, for each state $s\in\mathcal{S}$, when $n\ge N_{2,s}:=N\left(\frac{\varepsilon}{6|\mathcal{S}|},\delta,\nu^0_{s,\pi(s)}\right)$ (defined in Lemma 5), with probability at least $1-\varepsilon/(6|\mathcal{S}|)$,

$$\left|\widehat{V}^{\mathrm{rob},\pi}(s)-V^{\mathrm{rob},\pi}(s)\right|\le\frac{C_{s,\pi(s)}\sqrt{\frac{1}{c}\log\left(\frac{12C|\mathcal{S}|}{\varepsilon}\right)}}{\delta\sqrt{\eta n/(2|S|)}}+\gamma\|\widehat{V}^{\mathrm{rob},\pi}-V^{\mathrm{rob},\pi}\|_\infty$$
$$+\gamma\left|\sup_{\beta\ge0}\left\{-\beta\log\left(\sum_{s'\in\mathcal{S}}\widehat{p}_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}\right)-\beta\delta\right\}-\right.$$
$$\left.\sup_{\beta\ge0}\left\{-\beta\log\left(\sum_{s'\in\mathcal{S}}p^0_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}\right)-\beta\delta\right\}\right|,$$

where the constant $C_{s,\pi(s)},c$ and C are determined in Lemma 5. Note that $C_{s,\pi(s)}$ is independent of δ. Let β^* denote an the optimal solution of

$$\sup_{\beta\ge0}\left\{-\beta\log\left(\sum_{s'\in\mathcal{S}}p^0_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}\right)-\beta\delta\right\},$$

and $\widehat{\beta}^*$ denote an optimal solution of

$$\sup_{\beta\ge0}\left\{-\beta\log\left(\sum_{s'\in\mathcal{S}}\widehat{p}_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}\right)-\beta\delta\right\}.$$

Next we discuss the upper bound of the above inequality via different values of β^*.
Case 1, $\beta^* = 0$. By Lemma 4, for $n \geq N_{3,s} := N'\left(\frac{\varepsilon}{6|\mathcal{S}|}, \delta, p^0_{s,\pi(s)}\right)$ (defined in Lemma 4), with high probability there is a optimal $\widehat{\beta}^*$ equals 0. Hence

$$\begin{aligned}\left|\widehat{V}^{\mathrm{rob},\pi}(s) - V^{\mathrm{rob},\pi}(s)\right| \leq & \frac{C_{s,\pi(s)}\sqrt{\frac{1}{c}\log\left(\frac{12C|\mathcal{S}|}{\varepsilon}\right)}}{\delta\sqrt{\eta n/(2|S|)}} + \gamma\|\widehat{V}^{\mathrm{rob},\pi} - V^{\mathrm{rob},\pi}\|_\infty \\ & + \gamma\left|\min_{\{i:(s_i,a_i)=(s,\pi(s))\}} V^{\mathrm{rob},\pi}(s'_i) - \operatorname*{ess\,inf}_{s'\sim p^0_{s,\pi(s)}(\cdot)} V^{\mathrm{rob},\pi}(s')\right|.\end{aligned}$$

Since $|\mathcal{S}|$ is finite, we have $\min_{\{s':p^0_{s,\pi(s)}(s')>0\}} p^0_{s,\pi(s)}(s') > 0$, and

$$\mathbb{P}\left(\min_{\{i:(s_i,a_i)=(s,\pi(s))\}} V^{\mathrm{rob},\pi}(s'_i) \neq \operatorname*{ess\,inf}_{s'\sim p^0_{s,\pi(s)}(\cdot)} V^{\mathrm{rob},\pi}(s')\right) \leq \left(1 - \min_{\{s':p^0_{s,\pi(s)}(s')>0\}} p^0_{s,\pi(s)}(s')\right)^{\frac{\eta n}{2|\mathcal{S}|}}.$$

As a result. for $n \geq N_{4,s} := \log\left(\frac{1}{1-\min_{\{s':p^0_{s,\pi(s)}(s')>0\}} p^0_{s,\pi(s)}(s')}\right)^{-1} \log\left(\frac{6|\mathcal{S}|}{\varepsilon}\right)$, with probability $1 - \varepsilon/(6|\mathcal{S}|)$, we have

$$\min_{\{i:(s_i,a_i)=(s,\pi(s))\}} V^{\mathrm{rob},\pi}(s'_i) = \operatorname*{ess\,inf}_{s'\sim p^0_{s,\pi(s)}(\cdot)} V^{\mathrm{rob},\pi}(s').$$

Hence,

$$\left|\widehat{V}^{\mathrm{rob},\pi}(s) - V^{\mathrm{rob},\pi}(s)\right| \leq \frac{C_{s,\pi(s)}\sqrt{\frac{1}{c}\log\left(\frac{12C|\mathcal{S}|}{\varepsilon}\right)}}{\delta\sqrt{\eta n/(2|S|)}} + \gamma\|\widehat{V}^{\mathrm{rob},\pi} - V^{\mathrm{rob},\pi}\|_\infty.$$

By taking the supremum over s on the left-hand side, with an union bound of probability, we have for $n \geq \max\{N_1, \max_s N_{2,s}, \max_s N_{3,s}, \max_s N_{4,s}\}$, with probability at least $1 - \varepsilon/2 - |\mathcal{S}| \cdot (\varepsilon/(6|\mathcal{S}|) \cdot 3) = 1 - \varepsilon$ that

$$\left|\widehat{V}^{\mathrm{rob},\pi}(s) - V^{\mathrm{rob},\pi}(s)\right| \leq \frac{\max_s C_{s,\pi(s)}\sqrt{\frac{1}{c}\log\left(\frac{2C}{\varepsilon}\right)}}{\delta(1-\gamma)}\sqrt{\frac{2|\mathcal{S}|}{\eta n}}.$$

Case 2, $\beta^* > 0$. By Lemma 4 once again, when $n \geq N'_{3,s} := N''\left(\frac{\varepsilon}{6|\mathcal{S}|}, \delta, p^0_{s,\pi(s)}\right)$ (defined in Lemma 4), with probability at least $1 - \varepsilon/(6|\mathcal{S}|)$, we have $\beta^*, \widehat{\beta}^* \in [\underline{\beta}, \bar{\beta}]$, where $\underline{\beta} > 0$ and $\bar{\beta} = R_{\max}/((1-\gamma)\delta)$.
Similar to what we have done in Lemma 5,

$$\begin{aligned}&\left|\sup_{\beta\geq 0}\left\{-\beta\log\left(\sum_{s'\in\mathcal{S}}\widehat{p}_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}\right) - \beta\delta\right\} - \sup_{\beta\geq 0}\left\{-\beta\log\left(\sum_{s'\in\mathcal{S}}p^0_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}\right) - \beta\delta\right\}\right| \\ &\leq \sup_{\beta\in[\underline{\beta},\bar{\beta}]}\left\{\left|\left[-\beta\log\left(\sum_{s'\in\mathcal{S}}\widehat{p}_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}\right) - \beta\delta\right] - \left[-\beta\log\left(\sum_{s'\in\mathcal{S}}p^0_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}\right) - \beta\delta\right]\right|\right\} \\ &\leq \sup_{\beta\in[\underline{\beta},\bar{\beta}]}\left\{\left|\beta\log\left(1 + \frac{\sum_{s'\in\mathcal{S}}p^0_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta} - \sum_{s'\in\mathcal{S}}\widehat{p}_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}}{\sum_{s'\in\mathcal{S}}\widehat{p}_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}}\right)\right|\right\} \\ &\leq \frac{2R_{\max}}{(1-\gamma)\delta\exp\left(-R_{\max}/(\underline{\beta}(1-\gamma))\right)}\sup_{\beta\in[\underline{\beta},\bar{\beta}]}\left\{\left|\sum_{s'\in\mathcal{S}}p^0_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta} - \sum_{s'\in\mathcal{S}}\widehat{p}_{s,\pi(s)}(s')e^{-V^{\mathrm{rob},\pi}(s')/\beta}\right|\right\} \\ &\leq \frac{2R_{\max}}{(1-\gamma)\delta\exp\left(-R_{\max}/(\underline{\beta}(1-\gamma))\right)}\sum_{s'\in\mathcal{S}}\left|p^0_{s,\pi(s)}(s') - \widehat{p}_{s,\pi(s)}(s')\right|.\end{aligned}$$

Lastly, by Hoeffding's inequality (Lemma 3), with probability at least $1-\varepsilon/(6|\mathcal{S}|)$, we have

$$\sum_{s'\in\mathcal{S}}\left|p^0_{s,\pi(s)}(s')-\widehat{p}_{s,\pi(s)}(s')\right|\le\sqrt{\frac{1}{2}\log\left(\frac{12|\mathcal{S}|^2}{\varepsilon}\right)}\frac{1}{\sqrt{\eta n/(2|\mathcal{S}|)}}.$$

Hence,

$$\left|\widehat{V}^{\mathrm{rob},\pi}(s)-V^{\mathrm{rob},\pi}(s)\right|\le\frac{C'_{s,\pi(s)}\sqrt{\frac{1}{c}\log\left(\frac{12C|\mathcal{S}|}{\varepsilon}\right)}}{\delta\sqrt{\eta n/(2|\mathcal{S}|)}}+\gamma\|\widehat{V}^{\mathrm{rob},\pi}-V^{\mathrm{rob},\pi}\|_\infty,$$

where $C'_{s,\pi(s)}$ is a constant depends on $\nu^0_{s,\pi(s)}$ and $p^0_{s,\pi(s)}$. Taking the supremum over s on the left-hand side yields the desired result. □

10 Proof of Theorem 2

Proof. For each state $s\in\mathcal{S}$, we have $V^{\mathrm{rob},*}(s)-V^{\mathrm{rob},\widehat{\pi}^{\mathrm{rob},*}_\delta}(s)\ge 0$ by definition. Follows from Theorem 1,

$$\begin{aligned}
V^{\mathrm{rob},*}(s)-V^{\mathrm{rob},\widehat{\pi}^{\mathrm{rob},*}_\delta}(s)&\le\left|V^{\mathrm{rob},*}(s)-\widehat{V}^{\mathrm{rob},*}(s)\right|+\|\widehat{V}^{\mathrm{rob},\widehat{\pi}^{\mathrm{rob},*}_\delta}-V^{\mathrm{rob},\widehat{\pi}^{\mathrm{rob},*}_\delta}\|_\infty\\
&=\left|\sup_\pi V^{\mathrm{rob},\pi}(s)-\sup_\pi\widehat{V}^{\mathrm{rob},\pi}(s)\right|+\|\widehat{V}^{\mathrm{rob},\widehat{\pi}^{\mathrm{rob},*}_\delta}-V^{\mathrm{rob},\widehat{\pi}^{\mathrm{rob},*}_\delta}\|_\infty\\
&\le\sup_\pi\left|V^{\mathrm{rob},\pi}(s)-\widehat{V}^{\mathrm{rob},\pi}(s)\right|+\|\widehat{V}^{\mathrm{rob},\widehat{\pi}^{\mathrm{rob},*}_\delta}-V^{\mathrm{rob},\widehat{\pi}^{\mathrm{rob},*}_\delta}\|_\infty\\
&\le 2\sup_\pi\|V^{\mathrm{rob},\pi}-\widehat{V}^{\mathrm{rob},\pi}\|_\infty.
\end{aligned}$$

Thus $\|V^{\mathrm{rob},*}-V^{\mathrm{rob},\widehat{\pi}^{\mathrm{rob},*}_\delta}\|_\infty\le 2\sup_\pi\|V^{\mathrm{rob},\pi}-\widehat{V}^{\mathrm{rob},\pi}\|_\infty$. Note that there are the total number of fixed policy from $\mathcal{S}$ to $\mathcal{A}$ is bounded by $|\mathcal{A}|^{|\mathcal{S}|}$. Moreover, from Lemma 4 we know that $\max_\pi N(\varepsilon,\delta,\nu^0,\mathcal{P}^0,\pi)$ is essentially maximize among constants N that induced by all possible $p^0_{s,a}$ and $\nu^0_{s,a}$, hence it is in fact a constant $N(\varepsilon,\delta,\nu^0,\mathcal{P}^0)$ only depends on ε, δ and the original environment (ν^0 and $\mathcal{P}^0$). Similarly, we also have $\max_\pi C_\pi$ is in fact maximizing over $C_{s,a}$'s that induced by all possible state and action pairs. In other words, $\max_\pi C_\pi$ is constant that only depends on ν^0 and $\mathcal{P}^0$. Now, follows from Theorem 1, when $n\ge N(\varepsilon/|\mathcal{A}|^{|\mathcal{S}|},\delta,\nu^0,\mathcal{P}^0)$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_\pi\|V^{\mathrm{rob},\pi}-\widehat{V}^{\mathrm{rob},\pi}\|_\infty\ge\frac{\max_\pi C_\pi\sqrt{\frac{1}{c}\log\left(\frac{C|\mathcal{S}|\cdot|\mathcal{A}|^{|\mathcal{S}|}}{\varepsilon}\right)}}{1-\gamma}\sqrt{\frac{|\mathcal{S}|}{\eta n}}\right)\\
\le\ &\sum_\pi\mathbb{P}\left(\|V^{\mathrm{rob},\pi}-\widehat{V}^{\mathrm{rob},\pi}\|_\infty\ge\frac{C_\pi\sqrt{\frac{1}{c}\log\left(\frac{C|\mathcal{S}|\cdot|\mathcal{A}|^{|\mathcal{S}|}}{\varepsilon}\right)}}{1-\gamma}\sqrt{\frac{|\mathcal{S}|}{\eta n}}\right)\\
\le\ &\frac{\varepsilon}{|\mathcal{A}|^{|\mathcal{S}|}}\cdot|\mathcal{A}|^{|\mathcal{S}|}=\varepsilon.
\end{aligned}$$

Finally, observe that $|\mathcal{S}|\log\left(|\mathcal{S}|\cdot|\mathcal{A}|^{|\mathcal{S}|}/\varepsilon\right)=|\mathcal{S}|\log\left(|\mathcal{S}|/\varepsilon\right)+|\mathcal{S}|^2\log(|\mathcal{A}|)\le|\mathcal{S}|^2\log\left(|\mathcal{S}|\cdot|\mathcal{A}|/\varepsilon\right)$, we arrive at the desired result. □