A Sample Article Using IEEEtran.cls for IEEE Journals and Transactions

IEEE Publication Technology, *Staff, IEEE,*

***Abstract*—In reinforcement learning (RL), addressing generalization across different environments is essential, especially given uncertain model perturbations. Our work introduces the Robust Conservative Policy Iteration (RCPI) algorithm, employing the distributionally-robust optimization (DRO) framework with the bilevel optimization algorithm to solve robust Markov Decision Processes (MDPs). This novel approach ensures monotonic policy improvement in worst-case scenarios, with theoretical guarantees of convergence to an optimal policy under mild assumptions, providing the iteration complexity of $\mathcal{O}(\frac{1}{1-\gamma}\frac{1}{\epsilon^2})$ and the sample complexity of $\mathcal{O}(\epsilon^{-5})$ to achieve the ϵ-accuracy in the worst-case value function. Empirical tests on synthetic environments demonstrate RCPI's superiority in deriving resilient and reliable policies, outperforming traditional strategies.**

***Index Terms*—Article submission, IEEE, IEEEtran, journal, LaTeX, paper, template, typesetting.**

I. Introduction

REINFORCEMENT learning (RL) plays a pivotal role in modern machine learning research, particularly in the domain of training RL agents to perform well in dynamic environments while maintaining high performance even for the worst-case scenario [1]–[4]. Many real-world applications are rooted in the development of this field, such as control [5], power system [6], robotics [7] and autonomous driving [8]. This challenge of generalization often encounters a significant drop in performance due to model discrepancy, a phenomenon where the training environment does not accurately represent the deployed environment. The standard approach addressing this issue is to consider the model uncertainty [9]–[12] - the variability and unpredictability in the environment's transition - and to focus on maximizing the agent's worst-case performance during training.

Modeling such complex training environments necessitates an extension of the classical Markov Decision Process (MDP) framework to robust MDPs, the foundation of which has been extensively studied by [9], [13]–[15]. Solving robust MDPs presents a significantly greater challenge than the standard scenario, as the optimal policy may not be deterministic, and solutions heavily depend on environment uncertainty [9]. There has been a surge in research focused on developing algorithms that optimize for the worst-case performance, ensuring theoretical soundness. [13] demonstrated the feasibility of solving robust MDPs; however, their proposed value iteration method requires prior knowledge of the worst-case transitions. Subsequent studies have developed two common approaches to resolve this issue. One is the *sample-based method* that leverages data from a nominal transition model [16]–[18]. These approaches aim to learn robust policies without explicitly computing worst-case scenarios. Another is the *model-based method* that estimate the worst-case transition model during the learning process [19], [20]. These techniques attempt to solve the inner minimization problem of finding the worst-case dynamics. Notably, recent work by Kumar et al. [21] has shown that for certain uncertainty sets, the worst-case transition model is a rank-one perturbation of the nominal model, indicating that the worst-case transition model can be directly learned from the nominal transition model. This insight suggests that the two approaches above may be more closely related than previously thought.

In this context, the policy gradient method emerges as a promising approach for addressing the challenges posed by robust MDPs due to their efficiency and ability to directly optimize decision-making policies. Many algorithms are proposed with theoretical convergence guarantees, such as Robust Policy Mirror Descent [16], which adapts mirror descent to robust MDPs; Double-Loop Robust Policy Gradient [19], which alternates between policy updates and worst-case transition estimation; Robust Policy Iteration [22], which extends classical policy iteration to the robust setting; and Monotonic Robust Policy Optimization [23], which ensures monotonic improvement under model discrepancies. Despite these advances, existing work either lacks of a finite-sample complexity analysis [23] or lacks of the monotonicity of robust policy improvements [16], [19], [22], which presents significant challenges remain in robust RL. The challenge of establishing finite-sample complexity guarantees based on the theoretically-guaranteed monotonicity of robust policy improvements underscore the need for further research to bridge the gap between theoretical guarantees and practical applicability. Addressing these open problems requires innovative algorithmic approaches that can efficiently navigate the complex landscape of robust MDPs while maintaining computational tractability and theoretical soundness.

A. Contributions

Building upon these advancements, we introduce a novel policy-based algorithm to address the challenges inherent in solving robust MDPs. By leveraging the strengths of the distributionally robust optimization (DRO) framework, we reformulate the problem of solving robust MDPs as a bilevel optimization problem. This innovative approach enables the adaptation of existing bilevel optimization algorithms, such

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received October 8, 2023; revised December 8, 2023.

as [24], to effectively solve robust MDPs. Our contributions include:

- We propose the Robust Conservative Policy Iteration (RCPI) algorithm, which extends traditional conservative policy iteration [25] by incorporating the DRO framework to handle environmental uncertainties. Our algorithm ensures monotonic policy improvement in the worst-case scenario, akin to the guarantees of the original conservative policy iteration approach.
- Under some mild assumptions, we demonstrate that our proposed RCPI algorithm offers theoretical convergence guarantees to an optimal policy for the worst-case scenario. The iteration complexity of achieving the ϵ-accuracy in the robust value function is $\mathcal{O}(\frac{1}{1-\gamma}\frac{1}{\epsilon^2})$, and the sample complexity is $\mathcal{O}(\epsilon^{-5})$.
- We conduct experiments on synthetic environments to validate our algorithm's performance empirically. The results indicate the RCPI algorithm's superior ability to find robust policies that perform well across a range of environments, thereby significantly outperforming traditional methods in terms of resilience and reliability.

B. Related Work

a) Value-based approaches with model uncertainty: The value-based method is also an attractive direction of a robust RL algorithm. [10] designs the robust Q-learning algorithm for a certain uncertainty model. To generalize it to a more general setting, [26], [27] utilize the technique from the distributionally-robust optimization (DRO) theory, which leads to a simple but elegant design of robust Q-learning algorithm for solving robust MDPs with a much more general uncertainty set. Their developed technique of combining robust Q-learning and DRO theory can potentially be generalized to the policy-based method, which motivates us to design the RCPI method.

b) Policy-based approaches with model uncertainty: [28] designs the policy gradient method for a specific uncertainty model. [16] extends this method by considering the robust mirror ascent on a more general uncertainty model; by selecting appropriate Bregman divergence, this result demonstrates the convergence guarantee of projected policy gradient [29] and natural policy gradient [30]. The recent work [19] presents the double-loop robust policy gradient method, which adapts the project policy gradient method to the scenario of robust MDPs; the update rule of the policy gradient relies on estimating the worst-case transition probability, which can be effectively solved using value-iteration and gradient-based approaches. Extending this idea, we adopt a bilevel optimization framework to solve the RCPI algorithm, which provides a new perspective to analyze and understand robust RL algorithms.

c) Distributionally Robust Optimization (DRO): The recent development of DRO theory is one of the most crucial components of our proposed RCPI algorithm. The DRO aims to solve the optimization problem over a set of data distribution (usually called the *uncertainty set*) instead of a single data distribution [31]–[33]. The common approach to solve the DRO problem is to treat the uncertainty set as a constraint of the data distribution and then re-write it as the dual form [34]. Recently, [35] has extended this result to a general non-convex objective function with a theoretical convergence guarantee. The RL problem inherently contains dynamic and diverse data distributions, which makes the DRO framework an ideal choice to characterize such uncertainty.

II. Preliminaries

In this section, we recap mathematical notations and concepts used in this work.

A. Robust reinforcement learning

We mainly focus on the discounted infinite-horizon Markov Decision Processes (MDP), which is defined as a five-tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote finite state and action spaces, respectively. The transition probability $P(s'|s,a)$ represents the probability of transitioning from state s to state s' by taking action a. The reward function $r : \mathcal{S} \to [0,1]$ assigns rewards for each state. The discounted factor $\gamma \in (0,1)$ quantifies the diminishing value of future rewards.

Instead of considering a single transition probability P, we are interested in resolving the impact of *model uncertainty* (or *environment shifting*) that is ubiquitous in real-world applications. Let P_0 be the nominal transition and ϕ be a divergent function. Given a threshold $c > 0$, we define $\mathcal{U}_{s,a} := \{u_{s,a} = P_u(\cdot|s,a) - P_0(\cdot|s,a) : \max_{s,a \in \mathcal{S}\times\mathcal{A}} \phi(P_0(\cdot|s,a), P_u(\cdot|s,a) \leqslant c\}$. Then, the uncertainty set is defined as $\mathcal{U} = \times_{s,a} \mathcal{U}_{s,a}$.

We note that by our definition of uncertainty set, the uncertainty u and the transition probability P_u are interchangeable. In this sense, we define the value function of the uncertainty u as the value function of the transition probability P_u:

$$\begin{aligned} V_u^\pi(s) &:= E\big[\sum_{t=0}^{\infty} \gamma^t r(s_t) \mid s_0 = s, P_u, \pi\big] \\ &:= E_{\tau|\pi,u}\big[\sum_{t=0}^{\infty} \gamma^t r(s_t) \mid s_0 = s\big], \end{aligned}$$

where we use "$\tau|\pi, u$" to represent the trajectory generated by P_u and π. The robust value function is the worst-case value function over all uncertainties; that is

$$V^\pi(s) := \min_u V_u^\pi(s).$$

Moreover, we denote the expectation of robust value function with respect to its initial state distribution μ_0 as $\eta(\pi) := E_{s\sim\mu_0} V^\pi(s)$. For simplicity, we assume μ_0 is supported by a singleton. Similarly, we also define the robust Q-function and the robust advantage function as

$$\begin{aligned} Q^\pi(s,a) &:= E_{\tau|\pi,u}\big[\sum_{t=0}^{\infty} \gamma^t r(s_t) \mid s_0 = s, a_0 = a\big]; \\ A^\pi(s,a) &:= Q^\pi(s,a) - V^\pi(s). \end{aligned}$$

B. Distributionally robust optimization

The main technique used in our work to develop the robust RL algorithm is the distributionally robust optimization (DRO) framework, which seeks to optimize an objective function under the worst-case scenario from a set of possible probability distributions. It considers the following optimization problem

$$\max_{\pi\in\Pi}\min_{u\in\mathcal{U}}\mathbb{E}_{x\sim u}[f(\pi,x)],$$

where π is the parameter we aim to optimize, $\mathcal{U}$ is the uncertainty set, and $f(\pi,u)$ is the objective function, typically representing the value function to be maximized.

The uncertainty set is usually represented as a soft constraint. For the uncertainty set $\mathcal{U}:=\{u: d_\psi(u,u_0)\leqslant\epsilon\}$, our aim turns to solve

$$\max_{\pi\in\Pi}\min_{u\in\mathcal{U}}\left[\mathbb{E}_{x\sim u}[f(\pi,x)]+\lambda d(u,u_0)\right],$$

where u_0 is called the nominal distribution and $\lambda>0$ is the regularization coefficient. By [35], this unconstrained max-min optimization problem can be equivalently written as

$$\max_{\pi,\eta}\mathbb{E}_{x\sim u_0}\left[\lambda\psi^*\left(\frac{f(\pi;x)-\eta}{\lambda}\right)+\eta\right].$$

Solving this optimization problem has one advantage which makes the DRO framework an ideal characterization for robust RL: we only need to sample from the nominal distribution u_0 to obtain a stochastic gradient.

III. The Robust Conservative Policy Iteration Algorithm

In this section, we introduce the formulation of a robust conservative policy iteration algorithm. This algorithm represents a bridge between the distributionally robust optimization (DRO) and robust reinforcement learning.

A. Derive the robust conservative policy iteration

Updating the policy π to π', it is crucial to quantify the performance improvement under the worst-case scenario. This requirement leads us to establish a robust policy improvement lemma, described as follows:

Lemma 1. *Let $\eta(\pi)$ and $\eta(\pi')$ be the expected robust value function of π and π', respectively. Then*

$$\eta(\pi')\geqslant\eta(\pi)+\min_u\sum_s\rho_{\pi',u}(s)\sum_a\pi'(a|s)A^\pi(s,a). \quad (1)$$

Proof. See Lemma 4, Appendix A. □

This lemma indicates that updating π to π' has a non-negative improvement, which suffices to require $\sum_a\pi'(a|s)A^\pi(s,a)\geqslant 0$. The classical approach considers the deterministic policy $\pi'(s)=\arg\max_a A^\pi(s,a)$, which improves the policy at the state-action pair with a positive robust advantage value and nonzero visitation probability $\rho_{\pi',u}(s)$. However, as pointed out by [36], it is usually hard to directly optimize (1) since sampling from $\rho_{\pi',u}$ would be hard. To resolve this issue, we adopt the following local approximation of η as used in the TRPO algorithm [36] and the original conservative policy iteration algorithm [25]:

$$\begin{aligned}L_\pi(\pi'):&=\eta(\pi)+\min_u\sum_s\rho_{\pi,u}(s)\sum_a\pi'(a|s)A^\pi(s,a)\\&=\eta(\pi)+\min_u E_{(s,a)\sim\rho_{\pi,u}\otimes\pi'}A^\pi(s,a)\\&=\eta(\pi)+\min_u E_{(s,a)\sim\rho_{\pi,u}\otimes\pi}\frac{\pi'(a|s)}{\pi(a|s)}A^\pi(s,a).\end{aligned}$$

In this approximation, we replace $\rho_{\pi',u}$ with $\rho_{\pi,u}$, which ignores the change in visitation measure caused by the policy update. The following theorem shows that this approximation is generally accurate when two conditions are satisfied: (i) the two policies π and π' are sufficiently closed; (ii) the corresponding worst-case transition probability of π and π' (that is, u and u') are sufficiently closed.

Theorem 1. *Let $\epsilon=\max_{s,a}|A^\pi(s,a)|$, $\alpha_{\pi,\pi'}=\max_s D_{TV}(\pi,\pi')$, and $\beta_{u,u'}=\max_{s',s,a}|P_u-P_{u'}|$. Denote the worst-case transition of π is $u(\pi)$; that is, $u(\pi)=\arg\min_u V_u^\pi(s_0)$. Then*

$$\begin{aligned}\eta(\pi')-L_\pi(\pi')\geqslant&-\frac{\gamma^2}{(1-\gamma)^3}\left(\beta_{u,u'}^2+2\alpha_{\pi,\pi'}\beta_{u,u'}\right)\\&-\left[\frac{2\gamma}{(1-\gamma)^2}(2\alpha_{\pi,\pi'}^2+\alpha_{\pi,\pi'}\beta_{u,u'})\right]\epsilon.\end{aligned}$$

Moreover, the equality holds if $\pi'=\pi$; that is,

$$\eta(\pi)=L_\pi(\pi).$$

Proof. See Appendix B. □

Here, $\alpha_{\pi,\pi'}=\max_s D_{TV}(\pi,\pi')$ quantifies the total variation distance between policies π and π' and and $\beta_{u,u'}=\max_{s',s,a}|P_u-P_{u'}|$ measures the distance between their worst-case transitions. When $\pi=\pi'$, both distances are vanishing. The theorem's inequality establishes a lower bound for the gap between the robust expected return $\eta(\pi')$ and its local approximation $L_\pi(\pi')$, decomposing into penalties for significant discrepancies between two policies $\max_s D_{TV}(\pi,\pi')$ and their corresponding worst-case transitions $\beta_{u,u'}=\max_{s',s,a}|P_u-P_{u'}|$. Therefore, it is guaranteed to have a monotone policy improvement if we define the policy iteration step as

$$\begin{aligned}\pi'\leftarrow\arg\max_{\pi'}\min_u E_{(s,a)\sim\rho_{\pi,u}\otimes\pi}\Big[&L_{\pi_i}(\pi')\\&-\frac{\gamma^2}{(1-\gamma)^3}\left(\beta_{u,u'}^2+2\alpha_{\pi,\pi'}\beta_{u,u'}\right)\\&-\left[\frac{2\gamma}{(1-\gamma)^2}(2\alpha_{\pi,\pi'}^2+\alpha_{\pi,\pi'}\beta_{u,u'})\right]\epsilon\Big].\end{aligned}$$

where u and u' are the worst-case uncertainty of π and π', respectively. This iteration formula turns the robust policy optimization problem into a distributionally robust optimization (DRO) problem. We can re-write it into its dual-form [35]:

$$\begin{aligned}\max_{\pi',\eta}E_{(s,a)\sim\rho_{\pi,0}\otimes\pi}\Bigg[&\lambda\psi^*\Bigg(\frac{1}{\lambda}\Big(\frac{\pi'(a|s)}{\pi(a|s)}A^\pi(s,a)\\&-\frac{\gamma^2}{(1-\gamma)^3}\left(\beta^2+2\alpha\beta\right)\end{aligned}$$

$$- \Big[\frac{2\gamma}{(1-\gamma)^2}(2\alpha^2+\alpha\beta)\Big]\epsilon - \eta\Big)\Big) + \eta\Bigg]$$

where $\lambda > 0$ is the regularization coefficient. To address the challenge of unknown worst-case uncertainty denoted as u', we adopt the bilevel optimization approach as below:

$$\begin{aligned}
\text{(Upper)} \quad & \max_{\pi',\eta} E_{(s,a)\sim\rho_{\pi,0}\otimes\pi}\Bigg[\lambda\psi^*\Bigg(\frac{1}{\lambda}\Big(\frac{\pi'(a|s)}{\pi(a|s)}A^{\pi}(s,a) \\
& - \frac{\gamma^2}{(1-\gamma)^3}\Big(\beta^2+2\alpha\beta\Big) \\
& - \Big[\frac{2\gamma}{(1-\gamma)^2}(2\alpha^2+\alpha\beta)\Big]\epsilon - \eta\Big)\Big) + \eta\Bigg], \qquad (2) \\
\text{(Lower)} \quad & u' = \arg\min_{u} V_u^{\pi'}(s_0).
\end{aligned}$$

The lower-level optimization problem has been widely employed in the analysis of robust MDPs; moreover, [19] shows that when the uncertainty set is convex, the lower-level problem is also convex and gives a gradient-based method to solve this problem. For completeness, we also provide an alternative proof for this statement in Theorem 4. Employing this convexity, we can apply the primal-dual method given by [24] to solve this bilevel optimization problem. For more details, we re-formulate our problem into [24]'s framework and derive the sample complexity in Section D. As a summary, we obtain the following Algorithm 1.

Algorithm 1 Robust conservative policy iteration (RCPI) algorithm

```
Initialize π_0
loop
  Compute all robust advantage values A^{π_i}(s,a)
  π ← π_i
  Solve π' from the constrained bilevel optimization problem (2)
  π_{i+1} ← π'
  if η(π_{i+1}) − η(π_i) < ϵ then
    return π_i
  end if
  i ← i + 1
end loop
```

B. Properties of the proposed algorithm

In this subsection, we discuss some main properties of our derived algorithm RCPI.

Theorem 2. *Let $\{\pi_i\}_{i=1,2,\dots}$ be the sequence policies generated by Algorithm 1.*

(i) $\{\eta(\pi_i)\}_{i=1,2,\dots}$ is a non-decreasing sequence.

(ii) Let the algorithm terminate if $\eta(\pi_{i+1})-\eta(\pi_i) < \epsilon$. Then the output policy $\pi := \pi_i$ and $\pi' := \pi_{i+1}$ satisfy

$$\eta(\pi^*) - \eta(\pi) < C_{\pi'}\epsilon + \mathcal{E},$$

where $C_{\pi'}$ and $\mathcal{E}$ are non-negative constants depending on π, π', and π^*.

(iii) Suppose all policies are restricted to $\Pi_\delta = \{\delta U + (1-\delta)\pi : \pi \in \Delta\}$. Under the same termination condition as (ii), the output policy of Algorithm 1 requires at most $\mathcal{O}(\frac{1}{1-\gamma}\frac{1}{\epsilon})$ iterations to achieve ϵ/δ-accuracy.

Proof. See Appendix C. □

The first term (i) implies the convergence of Algorithm 1. Since the sequence of expected robust value functions $\{\eta(\pi_i)\}_{i=1,2,\dots}$ associated with the policy dynamics $\{\pi_i\}_{i=1,2,\dots}$ is non-decreasing, the sequence $\{\eta(\pi_i)\}_{i=1,2,\dots}$ must be convergent, given that all value functions are bounded above by $\frac{1}{1-\gamma}$. Nonetheless, this monotonicity alone does not ensure that $\lim_{i\to\infty}\eta(\pi_i)$ equals $\eta(\pi^*)$. To address this, item (ii) examines the algorithm's behavior upon termination, describing the discrepancy between the output and optimal policies. Meanwhile, item (iii) provides convergence guarantees and iteration complexity under some mild assumptions.

Discussions on the assumptions: In item (iii), we assume all policies are restricted to $\Pi_\delta = \{\delta U + (1-\delta)\pi : \pi \in \Delta\}$; it can be replaced with Π_ϵ so the optimal policy within this constraint will be an ϵ-approximation of the optimal policy. Then the termination condition should be adjusted to $\eta(\pi_{i+1})-\eta(\pi_i) < \epsilon^2$, ensuring that the output policy π attains ϵ-accuracy relative to the optimal policy, with a final iteration complexity of $\mathcal{O}(\frac{1}{1-\gamma}\frac{1}{\epsilon^2})$.

Total sample complexity: While the proposition primarily addresses iteration complexity, it is also crucial to consider sample complexity, i.e., the amount of data required to achieve ϵ-accuracy. Utilizing the bilevel optimization algorithm Proximal-PDBO [24] to solve (2), we find that approximately $\mathcal{O}(\epsilon^{-3})$ steps are needed to reach ϵ^2 accuracy. It implies a total sample complexity of $\mathcal{O}(\epsilon^{-5})$.

IV. Experiments

To validate the effectiveness of our proposed RCPI algorithm (Algorithm 1), we conducted experiments using the Frozen Lake environment [37]. This environment provides a testbed for our robust RL algorithm due to its inherent uncertainties and potential for catastrophic failures (falling into holes). The Frozen Lake environment consists of a 4×4 grid where the agent must navigate from the start state (left-top corner) to a goal state (right-bottom corner) while avoiding holes. The standard environment features slippery ice, introducing stochasticity. We further modified this setup to incorporate additional uncertainty, simulating varying ice conditions that affect the probability of slipping. Figure 1 illustrates the map configuration.

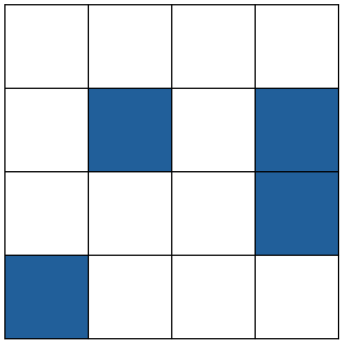

Fig. 1. The map configuration of the Frozen Lake environment [37]. The agent starts in the top-left corner and aims to reach the bottom-right corner. In the nominal model, the environment is deterministic: when the agent chooses a direction, it moves one step in that direction with probability 1. In the uncertain environment, the agent has a probability p of slipping an additional step.

We adapted the episodic environment to an infinite-horizon continuous learning setting by resetting the agent to the start position upon reaching the goal or falling into a hole. When the agent arrives the goal, it receives the reward 1; when the agent falls into a hole, it recieves the reward -1. Our experiments aimed to evaluate the performance of our proposed RCPI algorithm (Algorithm 1) in the worst-case environment. More explicitly, we set the probability of being slippery to be $p = 0.2$ and use KL-divergence to define the uncertainty set. Figure 2 presents the episodic rewards for both the non-robust policy iteration and our proposed RCPI algorithm.

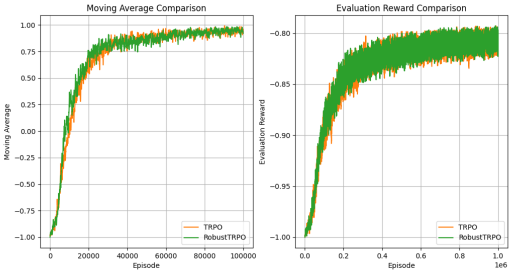

Fig. 2. Comparison of episodic rewards. Left: Performance in the nominal environment. Right: Performance in the worst-case environment.

The left panel of Figure 2 demonstrates that our RCPI algorithm performs comparably to the non-robust version in the nominal environment. More importantly, the right panel illustrates that the policy learned by RCPI exhibits significantly improved performance in the worst-case environment. These results underscore the robustness and effectiveness of our proposed approach in handling environmental uncertainties.

V. Conclusion

In this paper, we introduced the RCPI algorithm, a novel approach to solving robust MDPs that combines conservative policy iteration with distributionally robust optimization. Our key contributions include developing RCPI with guaranteed monotonic policy improvement in worst-case scenarios, providing theoretical convergence guarantees with an iteration complexity of $O(\frac{1}{1-\gamma}\frac{1}{\epsilon^2})$ and sample complexity of $O(\epsilon^{-5})$, and empirically validating its performance in the Frozen Lake environment. These results demonstrate RCPI's ability to learn policies robust to environmental uncertainties, outperforming non-robust methods in worst-case scenarios. Our work bridges the gap between theoretical guarantees and practical applicability in uncertain environments, opening new avenues for robust RL algorithm development. Future research directions include extending RCPI to continuous state and action spaces, integrating function approximation techniques, exploring multi-agent settings, and conducting more extensive empirical studies. The RCPI algorithm represents a significant advancement in robust reinforcement learning, offering a promising approach for developing reliable autonomous systems capable of operating in complex, real-world scenarios with inherent uncertainties.

Acknowledgments

This should be a simple paragraph before the References to thank those individuals and institutions who have supported your work on this article.

Appendix A
Useful Lemmas

In this section, we provide some necessary lemmas used to prove our main results.

Lemma 2 (Robust Bellman equation)**.** *The robust state-action value function Q^π and the robust state value function V^π satisfy*

$$Q^\pi(s,a) = r(s) + \gamma \min_{u\in\mathcal{U}} \sum_{s'\in S} P_u(s'|s,a)V^\pi(s') \tag{3}$$

for all $(s,a) \in S \times A$.

Proof. See Eq.(2.5) in [16]. □

Lemma 3 (Robust policy improvement lemma)**.** *Let $\eta(\pi)$ and $\eta(\pi')$ be the expected robust value function of π and π', respectively. Then*

$$\begin{aligned}&\eta(\pi') - \eta(\pi)\\ &= \min_{u\in\mathcal{U}} \sum_s \rho_{\pi',u}(s) \sum_a \pi'(a|s)\Big[r(s) + \gamma \sum_{s'\in S} P_u(s'|s,a)V^\pi(s') - V^\pi(s)\Big].\end{aligned} \tag{4}$$

Proof. The robust expected value function $\eta(\pi)$ is defined as $\eta(\pi) = E_{s_0\sim\mu_0} V^\pi(s_0)$; therefore, the improvement from π to π' is given by

$$\begin{aligned}&\eta(\pi') - \eta(\pi)\\ &= E_{s_0\sim\mu_0} V^{\pi'}(s_0) - E_{s_0\sim\mu_0} V^\pi(s_0)\\ &= -E_{s_0\sim\mu_0} V^\pi(s_0) + E_{s_0\sim\mu_0} \min_u E_{\tau|\pi',u}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t)\Big]\\ &= E_{s_0\sim\mu_0}\Bigg[- V^\pi(s_0) + \min_u E_{\tau|\pi',u}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t)\Big]\Bigg]\end{aligned}$$

Without loss of generality, we fix a starting state s_0. Then $\eta(\pi) = V^\pi(s_0)$ can be considered as a real number. Shifting a constant doesn't affect the minimization, so we get

$$\begin{aligned}
&\eta(\pi') - \eta(\pi) \\
&= \min_u E_{\tau|\pi',u}\left[-V^\pi(s_0) + \Big[\sum_{t=0}^{\infty}\gamma^t r(s_t)\Big]\right] \\
&= \min_u E_{\tau|\pi',u}\left[\sum_{t=0}^{\infty}\gamma^t\Big(r(s_t) + \gamma V^\pi(s_{t+1}) - V^\pi(s_t)\Big)\right] \\
&= \min_u E_{\tau|\pi',u}\left[\sum_{t=0}^{\infty}\gamma^t\Big(r(s_t) + \gamma E_{s_{t+1}\sim P_{u,\pi'}(\cdot|s_t)} V^\pi(s_{t+1}) - V^\pi(s_t)\Big)\right]
\end{aligned}$$

Define the shortcut notation

$$\begin{aligned}
\mathsf{A}_u(s) &:= r(s) + \gamma E_{s'\sim P_{u,\pi'}(\cdot|s)} V^\pi(s') - V^\pi(s) \\
&= r(s) + \gamma \sum_{s'\in S} P_{u,\pi'}(s'|s)V^\pi(s') - V^\pi(s) \\
&= r(s) + \gamma \sum_{s'\in S}\sum_{a\in A}\pi'(a|s)P_u(s'|s,a)V^\pi(s') - V^\pi(s) \\
&= \sum_{a\in A}\pi'(a|s)\Big[r(s) + \gamma\sum_{s'\in S}P_u(s'|s,a)V^\pi(s') - V^\pi(s)\Big]
\end{aligned}$$

Then we obtain

$$\begin{aligned}
\eta(\pi') &= \eta(\pi) + \min_u E_{\tau|\pi',u}\left[\sum_{t=0}^{\infty}\gamma^t \mathsf{A}_u(s_t)\right] \\
&= \eta(\pi) + \min_u \sum_{t=0}^{\infty}\sum_s P_u(s_t = s|\pi')\gamma^t \mathsf{A}_u(s) \\
&= \eta(\pi) + \min_u \sum_s \rho_{\pi',u}(s)\mathsf{A}_u(s).
\end{aligned}$$

We expand the shortcut notation $\mathsf{A}_u(s)$ and get the following policy improvement equality:

$$\begin{aligned}
\eta(\pi') = \eta(\pi) + \min_{u\in\mathcal{U}} \sum_s \rho_{\pi',u}(s)\sum_a \pi'(a|s)\Big[r(s) \\
+ \gamma\sum_{s'\in S}P_u(s'|s,a)V^\pi(s') - V^\pi(s)\Big].
\end{aligned}$$

Then the proof is completed. □

Lemma 4. *Let $\eta(\pi)$ and $\eta(\pi')$ be the expected robust value function of π and π', respectively.*

(i) Let $A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s)$ be the robust advantage function. Then

$$\eta(\pi') - \eta(\pi) \geqslant \min_u \sum_s \rho_{\pi',u}(s)\sum_a \pi'(a|s)A^\pi(s,a). \tag{5}$$

(ii) Moreover, if the worst-case uncertainty of the policy π is achieved at u^*, then

$$\eta(\pi') - \eta(\pi) \leqslant \big(\max_s \rho_{\pi',u^*}(s)\big)\sum_s\sum_a \pi'(a|s)A^\pi(s,a). \tag{6}$$

Proof. (i) First, we derive the lower bound of the right-hand side of (4):

$$\begin{aligned}
\eta(\pi') =& \eta(\pi) + \min_{u\in\mathcal{U}} \sum_s \rho_{\pi',u}(s)\sum_a \pi'(a|s)\Big[r(s) \\
&+ \gamma\sum_{s'\in S}P_u(s'|s,a)V^\pi(s') - V^\pi(s)\Big] \\
\geqslant& \eta(\pi) + \min_{u\in\mathcal{U}} \sum_s \rho_{\pi',u}(s)\sum_a \pi'(a|s)\min_{u\in\mathcal{U}_{s,a}}\Big[r(s) \\
&+ \gamma\sum_{s'\in S}P_u(s'|s,a)V^\pi(s') - V^\pi(s)\Big].
\end{aligned}$$

By Lemma 2,

$$\min_{u\in\mathcal{U}_{s,a}}\Big[r(s) + \gamma\sum_{s'\in S}P_u(s'|s,a)V^\pi(s') - V^\pi(s)\Big] = Q^\pi(s,a) - V^\pi(s).$$

This recovers the desired result:

$$\eta(\pi') \geqslant \eta(\pi) + \min_u \sum_s \rho_{\pi',u}(s)\sum_a \pi'(a|s)A^\pi(s,a).$$

(ii) Now we derive the upper bound of the right-hand side of (4). Let the worst-case uncertainty of the policy π is achieved at u^*. Then

$$\begin{aligned}
\eta(\pi') &= \eta(\pi) + \min_{u\in\mathcal{U}} \sum_s \rho_{\pi',u}(s)\sum_a \pi'(a|s)\Big[r(s) \\
&\quad + \gamma\sum_{s'\in S}P_u(s'|s,a)V^\pi(s') - V^\pi(s)\Big] \\
&\leqslant \eta(\pi) + \sum_s \rho_{\pi',u^*}(s)\sum_a \pi'(a|s)\Big[r(s) \\
&\quad + \gamma\sum_{s'\in S}P_{u^*}(s'|s,a)V^\pi(s') - V^\pi(s)\Big] \\
&\overset{(i)}{=} \eta(\pi) + \sum_s \rho_{\pi',u^*}(s)\sum_a \pi'(a|s)A^\pi(s,a) \\
&= \eta(\pi) + \max_s \rho_{\pi',u^*}(s)\sum_s\sum_a \pi'(a|s)A^\pi(s,a).
\end{aligned}$$

where (i) applies the robust Bellman equation (Lemma 2).

□

Lemma 5. *Let $\epsilon = \max_{s,a}|A^\pi(s,a)|$, $\alpha_{\pi,\pi'} = \max_s D_{TV}(\pi,\pi')$, and $\beta_{u,u'} = \max_{s',s,a}|P_u(s'|s,a) - P_{u'}(s'|s,a)|$. Then for any policy π' and any uncertainty $u' \in \mathcal{U}$,*

$$\Big\|\gamma V^\pi_{u(\pi)}\Delta^{u',\pi'}_{u(\pi),\pi}\Big\|_\infty \leqslant \frac{\gamma\beta_{u,u'}}{1-\gamma} + 2\alpha_{\pi,\pi'}\epsilon.$$

Proof. The s-th item of the vector $(\gamma V^\pi_{u(\pi)}\Delta^{u',\pi'}_{u(\pi),\pi})$ is decomposed to the distance between two uncertainty u' and $u(\pi)$ and the distance between two policy π and π' as the following:

$$\begin{aligned}
&|(\gamma V^\pi_{u(\pi)}\Delta^{u',\pi'}_{u(\pi),\pi})_s| \\
&= \left|\sum_{s',a}\Big[\gamma P_{u'}(s'|s,a)\pi'(a|s)V^\pi(s') - \gamma P_{u(\pi)}(s'|s,a)\pi(a|s)V^\pi(s')\Big]\right| \\
&= \Bigg|\sum_{s',a}\Big[\gamma\big[P_{u'}(s'|s,a) - P_{u(\pi)}(s'|s,a) + P_{u(\pi)}(s'|s,a)\big]\pi'(a|s)V^\pi(s') \\
&\quad - \gamma P_{u(\pi)}(s'|s,a)\pi(a|s)V^\pi(s')\Big]\Bigg| \\
&= \Big|\sum_{s',a}\Big[\gamma\big[P_{u'}(s'|s,a) - P_{u(\pi)}(s'|s,a)\big]\pi'(a|s)V^\pi(s')\Big]
\end{aligned}$$

$$+\sum_{s',a}\Big[\gamma P_{u(\pi)}(s'|s,a)\pi'(a|s)V^{\pi}(s')-\gamma P_{u(\pi)}(s'|s,a)\pi(a|s)V^{\pi}(s')\Big]\Big|$$

$$\leqslant \frac{\gamma\beta_{u,u'}}{1-\gamma}+\Bigg|\sum_{s',a}\Big[\gamma P_{u(\pi)}(s'|s,a)\pi'(a|s)V^{\pi}(s')$$

$$-\gamma P_{u(\pi)}(s'|s,a)\pi(a|s)V^{\pi}(s')\Big]\Bigg|$$

We bound the second term

$$\Bigg|\sum_{s',a}\Big[\gamma P_{u(\pi)}(s'|s,a)\pi'(a|s)V^{\pi}(s')-\gamma P_{u(\pi)}(s'|s,a)\pi(a|s)V^{\pi}(s')\Big]\Bigg|$$

by applying the robust Bellman equation $Q^{\pi}(s,a)=r(s)+\gamma\sum_{s'}P_{u(\pi)}(s'|s,a)V^{\pi}(s')$.

$$\begin{aligned}
&\Bigg|\sum_{s',a}\Big[\gamma P_{u(\pi)}(s'|s,a)\pi'(a|s)V^{\pi}(s')\\
&\qquad-\gamma P_{u(\pi)}(s'|s,a)\pi(a|s)V^{\pi}(s')\Big]\Bigg|\\
=&\Bigg|\sum_{a}\big(\pi'(a|s)-\pi(a|s)\big)\left[Q^{\pi}(s,a)-r(s)\right]\Bigg|\\
=&\Bigg|\sum_{a}\big(\pi'(a|s)-\pi(a|s)\big)\,A^{\pi}(s,a)\Bigg|\\
\leqslant&\sum_{a}\big|\pi'(a|s)-\pi(a|s)\big|\max_{a}A^{\pi}(s,a)\\
\leqslant&2\alpha_{\pi,\pi'}\epsilon.
\end{aligned}$$

Plugging it back, we get

$$|(\gamma V^{\pi}_{u(\pi)}\Delta^{u',\pi'}_{u(\pi),\pi})_s|\leqslant\frac{\gamma\beta_{u,u'}}{1-\gamma}+2\alpha_{\pi,\pi'}\epsilon.$$

It completes the proof. □

APPENDIX B
PROOF OF THEOREM 1

Theorem 3. *Let* $\epsilon=\max_{s,a}|A^{\pi}(s,a)|$, $\alpha_{\pi,\pi'}=\max_s D_{TV}(\pi,\pi')$, *and* $\beta_{u,u'}=\max_{s',s,a}|P_u-P_{u'}|$. *Denote the worst-case transition of* π *is* $u(\pi)$*; that is,* $u(\pi)=\arg\min_u V^{\pi}_u(s_0)$. *Then*

$$\begin{aligned}
\eta(\pi')-L_{\pi}(\pi')\geqslant&-\frac{\gamma^2}{(1-\gamma)^3}\Big(\beta^2_{u,u'}+2\alpha_{\pi,\pi'}\beta_{u,u'}\Big)\\
&-\Big[\frac{2\gamma}{(1-\gamma)^2}(2\alpha^2_{\pi,\pi'}+\alpha_{\pi,\pi'}\beta_{u,u'})\Big]\epsilon.
\end{aligned}$$

Moreover, the equality holds if $\pi'=\pi$*; that is,*

$$\eta(\pi)=L_{\pi}(\pi).$$

Proof. Recall that $P_{u,\pi}(s'|s)$ represents the probability of moving from the state s to the new state s' over the policy π. It can be written as a $S\times S$ matrix. In this sense, we define the matrix $G_{u,\pi}=(1-\gamma P_{u,\pi})^{-1}$ for any given u and π. Also, we define $\Delta^{u,\pi}_{u',\pi'}=P_{u,\pi}-P_{u',\pi'}$ for any given u',u,π',π. Our goal is to lower bound $\eta(\pi')-L_{\pi}(\pi')$. We start from the following decomposition:

$$\eta(\pi')-L_{\pi}(\pi')=\eta(\pi')-\eta(\pi)+\eta(\pi)-L_{\pi}(\pi')$$

$$=\Big(\eta(\pi')-\eta(\pi)\Big)+\Big(L_{\pi}(\pi)-L_{\pi}(\pi')\Big),$$

where the second equality is implied by the definition of L_{π}.

a) Bound $\eta(\pi')-\eta(\pi)$*:* The proof of this part is adapted from the perturbation theory proof of [36]. Recall that the worst-case expected value function $\eta(\pi)=rG_{u(\pi),\pi}\rho_0=r\rho_{u(\pi),\pi}$ and $\eta(\pi')=rG_{u(\pi'),\pi'}\rho_0=r\rho_{u(\pi'),\pi'}$.

$$G^{-1}_{u(\pi),\pi}-G^{-1}_{u(\pi'),\pi'}=\gamma(P_{u(\pi'),\pi'}-P_{u(\pi),\pi})$$

$$\begin{aligned}
&G_{u(\pi),\pi}\Big(G^{-1}_{u(\pi),\pi}-G^{-1}_{u(\pi'),\pi'}\Big)G_{u(\pi'),\pi'}\\
&\qquad=\gamma G_{u(\pi),\pi}\Big(P_{u(\pi'),\pi'}-P_{u(\pi),\pi}\Big)G_{u(\pi'),\pi'}
\end{aligned}$$

$$G_{u(\pi'),\pi'}-G_{u(\pi),\pi}=\gamma G_{u(\pi),\pi}\Big(P_{u(\pi'),\pi'}-P_{u(\pi),\pi}\Big)G_{u(\pi'),\pi'},$$

where we define

$$\Delta:=\Delta^{u(\pi'),\pi'}_{u(\pi),\pi}=P_{u(\pi'),\pi'}-P_{u(\pi),\pi}.$$

Then we obtain Eq.(47) of the TRPO paper [36]:

$$G_{u(\pi'),\pi'}=G_{u(\pi),\pi}+\gamma G_{u(\pi),\pi}\Delta G_{u(\pi'),\pi'}.$$

Also Eq.(48):

$$\begin{aligned}
G_{u(\pi'),\pi'}&=G_{u(\pi),\pi}+\gamma G_{u(\pi),\pi}\Delta\Big[G_{u(\pi),\pi}+\gamma G_{u(\pi),\pi}\Delta G_{u(\pi'),\pi'}\Big]\\
&=G_{u(\pi),\pi}+\gamma G_{u(\pi),\pi}\Delta G_{u(\pi),\pi}\\
&\quad+\gamma^2 G_{u(\pi),\pi}\Delta G_{u(\pi),\pi}\Delta G_{u(\pi'),\pi'}.
\end{aligned}$$

This leads to

$$\begin{aligned}
\eta(\pi')-\eta(\pi)&=r\Big(G_{u(\pi'),\pi'}-G_{u(\pi),\pi}\Big)\rho_0\\
&=r\Big(\gamma G_{u(\pi),\pi}\Delta G_{u(\pi),\pi}\\
&\quad+\gamma^2 G_{u,\pi}\Delta G_{u(\pi),\pi}\Delta G_{u(\pi'),\pi'}\Big)\rho_0\\
&=\gamma rG_{u(\pi),\pi}\Delta G_{u(\pi),\pi}\rho_0\\
&\quad+\gamma^2 rG_{u(\pi),\pi}\Delta G_{u(\pi),\pi}\Delta G_{u(\pi'),\pi'}\rho_0,
\end{aligned}$$

Also, the first term $rG_{u(\pi),\pi}\Delta G_{u(\pi),\pi}\rho_0=V^{\pi}\Delta\rho_{\pi,u(\pi)}$.

b) Upper bound $L_{\pi}(\pi')-L_{\pi}(\pi)$*:* We need an upper bound for $L_{\pi}(\pi')-L_{\pi}(\pi)$ that contains the term $V^{\pi}\Delta\rho_{\pi,u(\pi)}$.

$$\begin{aligned}
L_{\pi}(\pi')-L_{\pi}(\pi)=&\Big[\eta(\pi)+\min_u\sum_s\rho_{\pi,u}(s)\sum_a\pi'(a|s)A^{\pi}(s,a)\Big]\\
&-\Big[\eta(\pi)+\min_u\sum_s\rho_{\pi,u}(s)\sum_a\pi(a|s)A^{\pi}(s,a)\Big]\\
\leqslant&\sum_s\rho_{\pi,u(\pi)}(s)\sum_a\pi'(a|s)A^{\pi}(s,a)\\
&-\sum_s\rho_{\pi,u(\pi)}(s)\sum_a\pi(a|s)A^{\pi}(s,a).
\end{aligned}$$

We will investigate each term. By Proposition 2.2 [16],

$$\begin{aligned}
&\sum_s\rho_{\pi,u(\pi)}(s)\sum_a\pi'(a|s)A^{\pi}(s,a)\\
=&\sum_s\rho_{\pi,u(\pi)}(s)\sum_a\pi'(a|s)\Big[r(s)+\min_u\sum_{s'}P_u(s'|s,a)V^{\pi}(s')-V^{\pi}(s)\Big]\\
\leqslant&\sum_s\rho_{\pi,u(\pi)}(s)\sum_a\pi'(a|s)\Big[r(s)+\sum_{s'}P_{u(\pi')}(s'|s,a)V^{\pi}(s')-V^{\pi}(s)\Big]
\end{aligned}$$

$$\sum_s\rho_{\pi,u(\pi)}(s)\sum_a\pi(a|s)A^{\pi}(s,a)$$

$$= \sum_s \rho_{\pi,u(\pi)}(s) \sum_a \pi(a|s) \Big[r(s) + \sum_{s'} P_{u(\pi)}(s'|s,a) V^\pi(s') - V^\pi(s) \Big].$$

Then the gap between $L_\pi(\pi')$ and $L_\pi(\pi)$ becomes

$$\begin{aligned}
&L_\pi(\pi') - L_\pi(\pi) \\
&\leqslant \sum_s \rho_{\pi,u(\pi)}(s) \sum_a \pi'(a|s) \sum_{s'} P_{u(\pi')}(s'|s,a) V^\pi(s') \\
&\quad - \sum_s \rho_{\pi,u(\pi)}(s) \sum_a \pi(a|s) \sum_{s'} P_{u(\pi)}(s'|s,a) V^\pi(s') \\
&= \sum_s \rho_{\pi,u(\pi)}(s) \sum_a \sum_{s'} \Big[\pi'(a|s) P_{u(\pi')}(s'|s,a) \\
&\quad - \pi(a|s) P_{u(\pi)}(s'|s,a) \Big] V^\pi(s') \\
&= \sum_s \rho_{\pi,u(\pi)}(s) \sum_{s'} \Big[P_{u(\pi'),\pi'}(s'|s) - P_{u(\pi),\pi}(s'|s) \Big] V^\pi(s') \\
&= V^\pi \Delta \rho_{\pi,u(\pi)}.
\end{aligned}$$

Now we are back to the original decomposition

$$\begin{aligned}
\eta(\pi') - L_\pi(\pi') &= \eta(\pi') - \eta(\pi) + \eta(\pi) - L_\pi(\pi') \\
&= \Big(\eta(\pi') - \eta(\pi) \Big) + \Big(L_\pi(\pi) - L_\pi(\pi') \Big).
\end{aligned}$$

We combine two bounds together.

$$\begin{aligned}
\eta(\pi') - \eta(\pi) =& V^\pi \Delta \rho_{\pi,u(\pi)} \\
&+ \gamma^2 r G_{u(\pi),\pi} \Delta G_{u(\pi),\pi} \Delta G_{u(\pi'),\pi'} \rho_0, \\
L_\pi(\pi') - L_\pi(\pi) \leqslant& V^\pi \Delta \rho_{\pi,u(\pi)}.
\end{aligned}$$

Then

$$\eta(\pi') - L_\pi(\pi') \geqslant \gamma^2 r G_{u(\pi),\pi} \Delta G_{u(\pi),\pi} \Delta G_{u(\pi'),\pi'} \rho_0.$$

Lastly, we decomposite $\Delta := \Delta_{u(\pi),\pi}^{u(\pi'),\pi'} = P_{u(\pi'),\pi'} - P_{u(\pi),\pi}$ as

$$\begin{aligned}
\Delta &= P_{u(\pi'),\pi'} - P_{u(\pi),\pi'} + P_{u(\pi),\pi'} - P_{u(\pi),\pi} \\
&= \Delta_{u(\pi),\pi'}^{u(\pi'),\pi'} + \Delta_{u(\pi),\pi}^{u(\pi),\pi'}.
\end{aligned}$$

We have

$$\begin{aligned}
&\gamma^2 r G_{u(\pi),\pi} \Delta G_{u(\pi),\pi} \Delta G_{u(\pi'),\pi'} \rho_0 \\
&= \gamma^2 r G_{u(\pi),\pi} \Big[\Delta_{u(\pi),\pi'}^{u(\pi'),\pi'} + \Delta_{u(\pi),\pi}^{u(\pi),\pi'} \Big] G_{u(\pi),\pi} \\
&\quad \times \Big[\Delta_{u(\pi),\pi'}^{u(\pi'),\pi'} + \Delta_{u(\pi),\pi}^{u(\pi),\pi'} \Big] G_{u(\pi'),\pi'} \rho_0
\end{aligned}$$

Now it suffices to bound $\big| \gamma^2 r G_{u(\pi),\pi} \Delta G_{u(\pi),\pi} \Delta G_{u(\pi'),\pi'} \rho_0 \big|$. If we have the bound

$$\big| \gamma^2 r G_{u(\pi),\pi} \Delta G_{u(\pi),\pi} \Delta G_{u(\pi'),\pi'} \rho_0 \big| \leqslant \mathcal{E}$$

then we automatically obtain the lower bound

$$\gamma^2 r G_{u(\pi),\pi} \Delta G_{u(\pi),\pi} \Delta G_{u(\pi'),\pi'} \rho_0 \geqslant -\mathcal{E}.$$

By the Hölder's inequality,

$$\begin{aligned}
&\big| \gamma^2 r G_{u(\pi),\pi} \Delta G_{u(\pi),\pi} \Delta G_{u(\pi'),\pi'} \rho_0 \big| \\
&\leqslant \big\| \gamma^2 r G_{u(\pi),\pi} \Delta \big\|_\infty \big\| G_{u(\pi),\pi} \Delta G_{u(\pi'),\pi'} \rho_0 \big\|_1
\end{aligned}$$

To bound the first term $\big\| \gamma^2 r G_{u(\pi),\pi} \Delta \big\|_\infty = \gamma \big\| \gamma V_{u(\pi)}^\pi \Delta_{u(\pi),\pi}^{u(\pi'),\pi'} \big\|_\infty$, we let $u' = u(\pi')$. Then we directly apply Lemma 5 and get

$$\Big\| \gamma V_{u(\pi)}^\pi \Delta_{u(\pi),\pi}^{u',\pi'} \Big\|_\infty \leqslant \frac{\gamma \beta_{u,u'}}{1-\gamma} + 2\alpha_{\pi,\pi'} \epsilon.$$

The second term can be bounded as

$$\begin{aligned}
\big\| G_{u(\pi),\pi} \Delta G_{u(\pi'),\pi'} \rho_0 \big\|_1 &\leqslant \big\| G_{u(\pi),\pi} \big\|_1 \big\| \Delta \big\|_1 \big\| G_{u(\pi'),\pi'} \big\|_1 \big\| \rho_0 \big\|_1 \\
&= \frac{1}{1-\gamma} \times \big\| \Delta \big\|_1 \times \frac{1}{1-\gamma} \times 1 \\
&= \frac{\big\| \Delta \big\|_1}{(1-\gamma)^2} \leqslant \frac{\beta_{u,u'} + 2\alpha_{\pi,\pi'}}{(1-\gamma)^2}.
\end{aligned}$$

In summary, we obtain

$$\begin{aligned}
&\big| \gamma^2 r G_{u(\pi),\pi} \Delta G_{u(\pi),\pi} \Delta G_{u(\pi'),\pi'} \rho_0 \big| \\
&\leqslant \gamma \times \left(\frac{\gamma \beta_{u,u'}}{1-\gamma} + 2\alpha_{\pi,\pi'} \epsilon \right) \times \frac{\beta_{u,u'} + 2\alpha_{\pi,\pi'}}{(1-\gamma)^2}.
\end{aligned}$$

That is,

$$\begin{aligned}
&\gamma^2 r G_{u(\pi),\pi} \Delta G_{u(\pi),\pi} \Delta G_{u(\pi'),\pi'} \rho_0 \\
&\geqslant -\frac{\gamma^2}{(1-\gamma)^3} \Big(\beta_{u,u'}^2 + 2\alpha_{\pi,\pi'} \beta_{u,u'} \Big) \\
&\quad - \Big[\frac{2\gamma}{(1-\gamma)^2} (2\alpha_{\pi,\pi'}^2 + \alpha_{\pi,\pi'} \beta_{u,u'}) \Big] \epsilon.
\end{aligned}$$

Lastly, we give the equality condition. When π' is set to π, $L_\pi(\pi)$ exactly recovers $\eta(\pi)$ since

$$\begin{aligned}
L_\pi(\pi) &= \eta(\pi) + \min_u \sum_s \rho_{\pi,u}(s) \sum_a \pi(a|s) A^\pi(s,a) \\
&= \eta(\pi) + \min_u \sum_s \rho_{\pi,u}(s) \sum_a \pi(a|s) \big[Q^\pi(s,a) - V^\pi(s) \big] \\
&= \eta(\pi),
\end{aligned}$$

where the last equality holds because $\sum_a \pi(a|s) Q^\pi(s,a) = V^\pi(s)$. □

APPENDIX C
PROOF OF THEOREM 2

In this section, we provide the proof for Theorem 2. Here we give its full statement:

Proposition 1. *Let $\{\pi_i\}_{i=1,2,\dots}$ be the sequence policies generated by Algorithm 1.*

(i) $\{\eta(\pi_i)\}_{i=1,2,\dots}$ is a non-decreasing sequence.

(ii) Let the algorithm terminate if $\eta(\pi_{i+1}) - \eta(\pi_i) < \epsilon$. Then the output policy $\pi := \pi_i$ and $\pi' := \pi_{i+1}$ satisfy

$$\eta(\pi^*) - \eta(\pi) < C_{\pi'} \epsilon + \mathcal{E},$$

where $C_{\pi'} := \min_u \min_s \rho_{\pi',u}(s) \min_{\substack{s,a: \pi^*(a|s) \neq 0 \\ \pi'(a|s) \neq 0}} \{ \frac{\pi'(a|s)}{\pi^*(a|s)}, 1 \}$ and $\mathcal{E} := \sum_{\substack{\pi^*(a|s) \neq 0 \\ \pi'(a|s) = 0}} \big(\pi'(a|s) - \pi^*(a|s) \big) A^\pi(s,a)$.

(iii) Suppose all policies are restricted to $\Pi_\delta = \{\delta U + (1-\delta)\pi : \pi \in \Delta\}$. Under the same termination condition as (ii), the output policy of Algorithm 1 requires at most $\mathcal{O}(\frac{1}{1-\gamma} \frac{1}{\epsilon})$ iterations to achieve ϵ/δ-accuracy.

Proof. (i) First, we recall that Theorem 3 gives

$$\eta(\pi') - L_\pi(\pi') \geqslant -\frac{\gamma^2}{(1-\gamma)^3}\Big(\beta_{u,u'}^2 + 2\alpha_{\pi,\pi'}\beta_{u,u'}\Big) - \Big[\frac{2\gamma}{(1-\gamma)^2}(2\alpha_{\pi,\pi'}^2 + \alpha_{\pi,\pi'}\beta_{u,u'})\Big]\epsilon. \quad (7)$$

This result leads to our objective function in Algorithm 1:

$$M_i(\pi') := L_{\pi_i}(\pi') - \frac{\gamma^2}{(1-\gamma)^3}\Big(\beta_{u,u'}^2 + 2\alpha_{\pi,\pi'}\beta_{u,u'}\Big) - \Big[\frac{2\gamma}{(1-\gamma)^2}(2\alpha_{\pi,\pi'}^2 + \alpha_{\pi,\pi'}\beta_{u,u'})\Big]\epsilon.$$

We can further show the convergence of this proposed algorithm by adapting the standard argument of the minorization-maximization (MM) algorithm. By (7) and the above definition of M_i,

$$\eta(\pi_{i+1}) \geqslant M_i(\pi_{i+1}).$$

Also, when $\pi = \pi_i$, the gap between $\eta(\pi)$ and $L_\pi(\pi)$ is 0 due to the equality condition given in Theorem 3.

$$\eta(\pi_i) = M_i(\pi_i).$$

Then,

$$\eta(\pi_{i+1}) - \eta(\pi_i) \geqslant M_i(\pi_{i+1}) - M_i(\pi_i).$$

It leads to the policy iteration step in Algorithm 1. We need to solve

$$\pi_{i+1} = \arg\max\big[M_i(\pi_{i+1}) - M_i(\pi_i)\big]. \quad (8)$$

The maximization step guarantees $\{\eta(\pi_i)\}_{i=1,2,\dots}$ is a non-decreasing sequence; moreover, from the theory of MM algorithms, $\eta(\pi_i)$ converges to $\eta(\pi)$ for some π.

(ii) We let the algorithm terminates if $\eta(\pi_{i+1}) - \eta(\pi_i) < \epsilon$. Then the algorithm outputs the policy $\pi = \pi_i$ and we set $\pi' = \pi_{i+1}$. By the termination condition $\eta(\pi_{i+1}) - \eta(\pi_i) < \epsilon$, we get from Lemma 3,

$$\min_u \sum_s \rho_{\pi',u}(s) \sum_a \pi'(a|s)\Big[r(s) + \gamma \sum_{s'\in S} P_u(s'|s,a)V^\pi(s') - V^\pi(s)\Big] < \epsilon.$$

That is,

$$\begin{aligned}\epsilon &> \eta(\pi') - \eta(\pi)\\ &= \min_u \sum_s \rho_{\pi',u}(s) \sum_a \pi'(a|s)\Big[r(s) + \gamma \sum_{s'\in S} P_u(s'|s,a)V^\pi(s') - V^\pi(s)\Big]\\ &\overset{(i)}{\geqslant} \min_u \sum_s \rho_{\pi',u}(s) \sum_a \pi'(a|s)A^\pi(s,a).\end{aligned}$$

where (i) is by Lemma 4. We further decompose this term:

$$\begin{aligned}&\min_u \sum_s \rho_{\pi',u}(s) \sum_a \pi'(a|s)A^\pi(s,a)\\ =&\min_u \sum_s \rho_{\pi',u}(s)\Bigg[\sum_{\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)\neq 0}} \frac{\pi'(a|s)}{\pi^*(a|s)}\pi^*(a|s)A^\pi(s,a) + \sum_{\pi^*(a|s)=0} \pi^*(a|s)A^\pi(s,a) + \sum_{\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)=0}} \pi^*(a|s)A^\pi(s,a) - \sum_{\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)=0}} \pi^*(a|s)A^\pi(s,a) + \sum_{\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)=0}} \pi'(a|s)A^\pi(s,a)\Bigg].\end{aligned}$$

Let $C_{\pi',1} = \min_{s,\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)\neq 0}} \frac{\pi'(a|s)}{\pi^*(a|s)}$ and $C_{\pi',2} = 1$. Then we obtain

$$\begin{aligned}&\min_u \sum_s \rho_{\pi',u}(s) \sum_a \pi'(a|s)A^\pi(s,a)\\ \leqslant&\min_u \sum_s \rho_{\pi',u}(s)\Bigg[C_{\pi',1}\sum_{\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)\neq 0}} \pi^*(a|s)A^\pi(s,a) + C_{\pi',2}\Big(\sum_{\pi^*(a|s)=0} \pi^*(a|s)A^\pi(s,a) + \sum_{\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)=0}} \pi^*(a|s)A^\pi(s,a)\Big)\\ &- \sum_{\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)=0}} \pi^*(a|s)A^\pi(s,a) + \sum_{\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)=0}} \pi'(a|s)A^\pi(s,a)\Bigg].\end{aligned}$$

We denote the last two terms as

$$\mathcal{E} := -\sum_{\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)=0}} \pi^*(a|s)A^\pi(s,a) + \sum_{\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)=0}} \pi'(a|s)A^\pi(s,a)$$

as an error term which indicates the discrepancy between π' and π^* on actions that are not covered by π'. We note that when the policy during the policy iteration algorithm keeps sufficient exploration ability, there is no uncovered actions; in this case, the error term is exactly zero. We get

$$\begin{aligned}&\min_u \sum_s \rho_{\pi',u}(s) \sum_a \pi'(a|s)A^\pi(s,a)\\ \leqslant&\min_u \sum_s \rho_{\pi',u}(s)\Bigg[C_{\pi',1}\sum_{\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)\neq 0}} \pi^*(a|s)A^\pi(s,a) + C_{\pi',2}\Big(\sum_{\pi^*(a|s)=0} \pi^*(a|s)A^\pi(s,a) + \sum_{\substack{\pi^*(a|s)\neq 0\\ \pi'(a|s)=0}} \pi^*(a|s)A^\pi(s,a)\Big) + \mathcal{E}.\end{aligned}$$

Then we have:

$$\begin{aligned}&\min_u \sum_s \rho_{\pi',u}(s) \sum_a \pi'(a|s)A^\pi(s,a)\\ =&\min_u \sum_s \rho_{\pi',u}(s) \sum_a \frac{\pi'(a|s)}{\pi^*(a|s)}\pi^*(a|s)A^\pi(s,a)\\ >&\, C_{\pi'}[\eta(\pi^*) - \eta(\pi)],\end{aligned}$$

where $C_{\pi'} := \min_u \min_s \rho_{\pi',u}(s) \min\{C_{\pi',1}, C_{\pi',2}\}$. In summary, we obtain,

$$\begin{aligned}\epsilon &> \min_u \sum_s \rho_{\pi',u}(s) \sum_a \pi'(a|s)A^\pi(s,a)\\ &\geqslant \min_u \sum_s \rho_{\pi',u}(s) \sum_a \frac{\pi'(a|s)}{\pi^*(a|s)}\pi^*(a|s)A^\pi(s,a) + \mathcal{E}\end{aligned}$$

$$> C_{\pi'}[\eta(\pi^*) - \eta(\pi)] + \mathcal{E}.$$

Then the proof is completed. □

APPENDIX D
SOLVING THE BILEVEL OPTIMIZATION PROBLEM

In this section, we describe how to solve the bilevel optimization problem (2) given in our robust policy iteration by using existing techniques from Distributionally Robust Optimization (DRO) and Bilevel Optimization Algorithms. First, we recap the definition of the uncertainty set; in our work, we mainly consider the (s,a)-rectangular uncertainty set defined as

$$\begin{aligned}\mathcal{U}_{s,a} :=& \{u(\cdot|s,a) := P(\cdot|s,a) - P_0(\cdot|s,a)) \\ & \mid \ \phi(P(\cdot|s,a), P_0(\cdot|s,a)) \leqslant c\}\end{aligned}$$

and

$$\mathcal{U} := \bigtimes_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{U}_{s,a}.$$

The robust conservative policy iteration is given by solving the following DRO problem:

$$\max_{\pi'} \min_{u\in\mathcal{U}} E_{(s,a)\sim\rho_{\pi,u}\otimes\pi}\Bigg[L_{\pi_i}(\pi') \tag{9}$$

$$- \frac{\gamma^2}{(1-\gamma)^3}\Big(\beta_{u,u'}^2 + 2\alpha_{\pi,\pi'}\beta_{u,u'}\Big) \tag{10}$$

$$- \Big[\frac{2\gamma}{(1-\gamma)^2}(2\alpha_{\pi,\pi'}^2 + \alpha_{\pi,\pi'}\beta_{u,u'})\Big]\epsilon\Bigg]. \tag{11}$$

We need to re-written it as the form of bilevel optimization problem (2) by applying existing DRO analysis from [35]:

$$\begin{aligned}\text{(Upper)}\quad & \max_{\pi',\eta} E_{(s,a)\sim\rho_{\pi,0}\otimes\pi}\Bigg[\lambda\psi^*\Bigg(\frac{1}{\lambda}\Big(\frac{\pi'(a|s)}{\pi(a|s)}A^\pi(s,a) \\ & - \frac{\gamma^2}{(1-\gamma)^3}\Big(\beta^2 + 2\alpha\beta\Big) \\ & - \Big[\frac{2\gamma}{(1-\gamma)^2}(2\alpha^2 + \alpha\beta)\Big]\epsilon - \eta\Big)\Bigg) + \eta\Bigg], \\ \text{(Lower)}\quad & u' = \arg\min_u V_u^{\pi'}(s_0).\end{aligned}$$

Assuming we are using the stochastic Proximal-PDBO algorithm given in [24] to solve this bilevel optimization problem, the sample complexity of the robust conservative policy iteration (Algorithm 1) is given as $\epsilon^{-1.5}\times\epsilon^{-1} = \epsilon^{-2.5}$, where $\epsilon^{-1.5}$ is the complexity of solving the bilevel optimization problem (2) given in Corollary 2, [24] and ϵ^{-1} is the complexity of the policy iteration given in Theorem 3.

There are still two components that we aim to clarify in this section: (1) To directly apply the algorithm given in [24], we need to justify that the lower-level optimization is a convex optimization problem, and (2) to re-write the max-min optimization problem (9) into its duel form by applying [35], we need to construct the corresponding uncertainty set for the visitation measure.

A. Convexity of the lower-level optimization problem

First, we describe the convexity of the lower-level problem in (2). In general, the lower-level optimization problem cannot by strongly convex since there possibly exist multiple transitions with the same worst-case value function. Though this property has been pointed out by [19], we provide an alternative proof for completeness.

Theorem 4 (Convexity of Lower-Level Problem)**.** *Let V_u^π be the value function for the given uncertainty $u \in \mathcal{U}$ and the policy π. Suppose that the initial state is fixed at s_0. Then for every policy π, $u \mapsto V_u^\pi(s_0)$ is convex over $\mathcal{U}$.*

Proof. To establish this statement, we first consider that the constraint of the lower-level problem is based on a convex and compact uncertainty set $\mathcal{U}$. There is a bijection mapping from uncertainty u to transition probability P_u, given as $P_u(\cdot|s,a) = P_0(\cdot|s,a) + u(\cdot|s,a)$, which is linear so preserves convexity. Now, it suffices to consider the convexity over the set of transition probability. The objective function of the problem is expressed as the composition of two mappings, $f : X \mapsto X^{-1}$ and $g : X \mapsto 1 - \gamma X$. Since $f : X \mapsto X^{-1}$ is convex by [38] and $g : X \mapsto 1 - \gamma X$ is linear; the composition $f \circ g$ would also be convex, as the composition of convex functions preserves convexity. Therefore, given that the convexity of f is established, it follows that the objective function $P_u \mapsto r(1 - \gamma P_u)^{-1}\rho_0$ is convex in P_u, thereby concluding that $u \mapsto V_u^\pi(s_0)$ is convex over $\mathcal{U}$. □

B. Connecting the uncertainty set of transitions to visitation measures

We recall the definition of the uncertainty set $\mathcal{U}$, which is the (s,a)-rectangular uncertainty set induced by ϕ-divergence; that is, for a fixed *nominal transition* P_0 and a constant $c > 0$, the uncertainty set $\mathcal{U}$ is defined as

$$\mathcal{U} := \bigtimes_{s,a\in S\times A} \mathcal{U}_{s,a},$$

where $\mathcal{U}_{s,a} := \{u(\cdot|s,a) := P(\cdot|s,a) - P_0(\cdot|s,a)) : \phi(P(\cdot|s,a), P_0(\cdot|s,a)) \leqslant c\}$. Then we show that there exists a corresponding uncertainty set of the nominal visitation measure ρ_0 induced by a ψ-divergence. Then, to solve the DRO problem (9), it suffices to consider the uncertainty set of visitation measure

$$\{\rho : \psi(\rho, \rho_0) \leqslant c'\}$$

for the upper-level problem in (2). This leads to the dual-form of DRO problem:

$$\begin{aligned}\max_{\pi',\eta} E_{(s,a)\sim\rho_{\pi,0}\otimes\pi}\Bigg[\lambda\psi^*\Bigg(\frac{1}{\lambda}\Big(\frac{\pi'(a|s)}{\pi(a|s)}A^\pi(s,a) \\ - \frac{\gamma^2}{(1-\gamma)^3}\Big(\beta^2 + 2\alpha\beta\Big) \\ - \Big[\frac{2\gamma}{(1-\gamma)^2}(2\alpha^2 + \alpha\beta)\Big]\epsilon - \eta\Big)\Bigg) + \eta\Bigg].\end{aligned}$$

To build the connection between $\{\rho : \psi(\rho, \rho_0) \leqslant c'\}$ and the original uncertainty set $\mathcal{U}$, we consider the mapping described

as follows. When the policy π is given, each uncertainty $u \in \mathcal{U}$ corresponds to an occupancy measure ρ_u. We define this mapping:

$$\mathbf{\Lambda}_\pi : u \mapsto \rho_u.$$

More explicitly, let $\{s_0, s_1, \dots\}$ be the trajectory generated by the policy π and the uncertainty u. Then

$$\mathbf{\Lambda}_\pi(u)(s,a) := \sum_{t=0}^{\infty} \gamma^t \pi(a|s) \mathbf{P}_u(s_t = s|\pi).$$

Note that the mapping $\mathbf{\Lambda}_\pi$ is not an injection since for two different u and u', it is possible that their visitation measures are the same; that is, $\mathbf{\Lambda}_\pi(u) = \mathbf{\Lambda}_\pi(u')$. We consider the following construction ($\mathbf{P}$ is the nominal transition used to define the original uncertainty set):

- $\mathbf{\Lambda}_\pi(\mathcal{U}) := \{\mathbf{\Lambda}_\pi(u) : u \in \mathcal{U}\}$.
- $\mathcal{U}_\pi(\delta) = \{\rho : d_\phi(\rho \| \mathbf{\Lambda}_\pi(\mathbf{P})) \leqslant \delta\}$.

$\mathcal{U}_\pi(\delta)$ is the desired uncertainty set since we can directly apply the optimization paper. Then we can prove that the new uncertainty set over the visitation measure has the desired property:

Proposition 2. *The distributionally robust optimization (DRO) problem over $\mathcal{U}$,*

$$\min_{\theta \in \Theta} \max_{u \in \mathcal{U}} E_{\xi \sim \rho_{\pi,u}}[f(\theta;\xi)],$$

is equivalent to the $\mathbf{\Lambda}_\pi$-induced DRO problem

$$\min_{\theta \in \Theta} \max_{v \in \mathbf{\Lambda}_\pi(\mathcal{U})} E_{\xi \sim v}[f(\theta;\xi)].$$

That is, if $u^* := \arg\max_{u \in \mathcal{U}} E_{\xi \sim \rho_{\pi,u}}[f(\theta;\xi)]$, then

$$v^* := \arg\max_{v \in \mathbf{\Lambda}_\pi(\mathcal{U})} E_{\xi \sim v}[f(\theta;\xi)] = \mathbf{\Lambda}_\pi(u^*).$$

Conversely, given $v^* := \arg\max_{v \in \mathbf{\Lambda}_\pi(\mathcal{U})} E_{\xi \sim v}[f(\theta;\xi)]$, for any $u \in \mathbf{\Lambda}_\pi^{-1}(v^*)$,

$$E_{\xi \sim \rho_{\pi,u}}[f(\theta;\xi)] = \max_{u \in \mathcal{U}} E_{\xi \sim \rho_{\pi,u}}[f(\theta;\xi)].$$

Proof. It suffices to consider the change of variable.

$$\begin{aligned}
\max_{u \in \mathcal{U}} E_{\xi \sim \rho_{\pi,u}}[f(\theta;\xi)] &= \max_{u \in \mathcal{U}} \int f(\theta;\xi) d\rho_{\pi,u}(\xi) \\
&\overset{(i)}{=} \max_{\rho_{\pi,u} \in \mathbf{\Lambda}_\pi(\mathcal{U})} \int f(\theta;\xi) d\rho_{\pi,u}(\xi) \\
&= \max_{v \in \mathbf{\Lambda}_\pi(\mathcal{U})} \int f(\theta;\xi) dv(\xi) \\
&= \max_{v \in \mathbf{\Lambda}_\pi(\mathcal{U})} \mathbb{E}_{\xi \sim v} f(\theta;\xi).
\end{aligned}$$

Here, we replace u with $\mathbf{\Lambda}_\pi(u)$ and the constraint $\mathcal{U}$ is changed to its image $\mathbf{\Lambda}_\pi(\mathcal{U})$ in the step (i). □

However, the uncertainty set $\mathbf{\Lambda}_\pi(\mathcal{U})$ may not have the desired structure (e.g. convexity). It is much preferred to consider the uncertainty set of visitation measure induced by some divergence ψ; that is

$$\{\rho : \psi(\rho, \rho_0) \leqslant c'\},$$

where ρ_0 is the visitation measure corresponding to the nominal transition P_0. In the next theorem, we show that there exists such corresponding uncertainty set of the nominal visitation measure ρ_0 induced by the ψ-divergence shares the same worst-case uncertainty with the $\mathbf{\Lambda}_\pi(\mathcal{U})$.

Theorem 5. *Assume the mapping $\delta \mapsto \min_{\rho \in \mathcal{U}_\pi(\delta)} E_{x \sim \rho}[f(\pi;x)]$ is uniformly continuous in δ. Then there exists $\delta^\star$ such that*

$$\min_{\rho \in \mathcal{U}_\pi(\delta^\star)} E_{x \sim \rho}[f(\pi;x)] = \min_{\rho \in \mathbf{\Lambda}_\pi(\mathcal{U})} E_{x \sim \rho}[f(\pi;x)];$$

that is, solving the robust problem over the extended uncertainty set $\mathcal{U}_\pi(\delta^\star)$ is exactly same as the original robust problem.

Proof. Let $F_\pi(\delta) = \min_{\rho \in \mathcal{U}_\pi(\delta)} E_{x \sim \rho}[f(\pi;x)]$; it is a continuous function by our assumption. Define

$$\tau := \min_{\rho \in \mathbf{\Lambda}_\pi(\mathcal{U})} E_{x \sim \rho}[f(\pi;x)].$$

Then $f(0) \geqslant \tau$ and there exists δ such that $f(\delta) \leqslant \tau$. There must exist $\delta^\star$ such that $f(\delta^\star) = \tau$ by the continuity of f. □

REFERENCES

[1] C. Paduraru, D. Mankowitz, G. Dulac-Arnold, J. Li, N. Levine, S. Gowal, and T. Hester, "Challenges of real-world reinforcement learning: definitions, benchmarks and analysis," *Machine Learning*, vol. 110, pp. 2419 – 2468, 2021.

[2] L. Yu, J. Wang, and X. Zhang, "Robust reinforcement learning under model misspecification," *ArXiv*, vol. abs/2103.15370, 2021.

[3] J. Morimoto and K. Doya, "Robust reinforcement learning," *Neural Computation*, vol. 17, pp. 335–359, 2005.

[4] J. Moos, K. Hansel, H. Abdulsamad, S. Stark, D. Clever, and J. Peters, "Robust reinforcement learning: A review of foundations and recent advances," *Mach. Learn. Knowl. Extr.*, vol. 4, pp. 276–315, 2022.

[5] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.

[6] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, "Reinforcement learning for selective key applications in power systems: Recent advances and future challenges," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 2935–2958, 2022.

[7] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: a survey," in *2020 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2020, pp. 737–744.

[8] X. Hu, S. Li, T. Huang, B. Tang, R. Huai, and L. Chen, "How simulation helps autonomous driving: A survey of sim2real, digital twins, and parallel intelligence," *IEEE Transactions on Intelligent Vehicles*, 2023.

[9] W. Wiesemann, D. Kuhn, and B. Rustem, "Robust markov decision processes," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 153–183, 2013.

[10] Y. Wang and S. Zou, "Online robust reinforcement learning with model uncertainty," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7193–7206, 2021.

[11] S. Ma, Z. Chen, S. Zou, and Y. Zhou, "Decentralized robust v-learning for solving markov games with model uncertainty," *Journal of Machine Learning Research*, vol. 24, no. 371, pp. 1–40, 2023.

[12] K. Zhang, T. Sun, Y. Tao, S. Genc, S. Mallya, and T. Basar, "Robust multi-agent reinforcement learning with model uncertainty," *Advances in neural information processing systems*, vol. 33, pp. 10 571–10 583, 2020.

[13] A. Nilim and L. Ghaoui, "Robustness in markov decision problems with uncertain transition matrices," *Advances in neural information processing systems*, vol. 16, 2003.

[14] G. N. Iyengar, "Robust dynamic programming," *Mathematics of Operations Research*, vol. 30, no. 2, pp. 257–280, 2005.

[15] H. Xu and S. Mannor, "Distributionally robust markov decision processes," *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[16] Y. Li, G. Lan, and T. Zhao, "First-order policy optimization for robust markov decision process," *arXiv preprint arXiv:2209.10579*, 2022.

[17] W. Yang, H. Wang, T. Kozuno, S. M. Jordan, and Z. Zhang, "Avoiding model estimation in robust markov decision processes with a generative model," *arXiv preprint arXiv:2302.01248*, 2023.
[18] Z. Xu, K. Panaganti, and D. Kalathil, "Improved sample complexity bounds for distributionally robust reinforcement learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 9728–9754.
[19] Q. Wang, C. P. Ho, and M. Petrik, "Policy gradient in robust mdps with global convergence guarantee," in *International Conference on Machine Learning*. PMLR, 2023, pp. 35 763–35 797.
[20] K. Wang, U. Gadot, N. Kumar, K. Y. Levy, and S. Mannor, "EWok: Tackling robust markov decision processes via estimating worst kernel," 2024. [Online]. Available: https://openreview.net/forum?id=8WH6ZlDad6
[21] N. Kumar, E. Derman, M. Geist, K. Y. Levy, and S. Mannor, "Policy gradient for rectangular robust markov decision processes," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[22] K. P. Badrinath and D. Kalathil, "Robust reinforcement learning using least squares policy iteration with provable performance guarantees," in *International Conference on Machine Learning*. PMLR, 2021, pp. 511–520.
[23] Y. Jiang, C. Li, W. Dai, J. Zou, and H. Xiong, "Monotonic robust policy optimization with model discrepancy," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4951–4960.
[24] D. Sow, K. Ji, Z. Guan, and Y. Liang, "A primal-dual approach to bilevel optimization with multiple inner minima," *arXiv preprint arXiv:2203.01123*, 2022.
[25] S. M. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," in *International Conference on Machine Learning*, 2002. [Online]. Available: https://api.semanticscholar.org/CorpusID:31442909
[26] Z. Liu, Q. Bai, J. Blanchet, P. Dong, W. Xu, Z. Zhou, and Z. Zhou, "Distributionally robust q-learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 13 623–13 643.
[27] S. Wang, N. Si, J. Blanchet, and Z. Zhou, "A finite sample complexity bound for distributionally robust q-learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 3370–3398.
[28] Y. Wang and S. Zou, "Policy gradient method for robust reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23 484–23 526.
[29] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 4431–4506, 2021.
[30] S. M. Kakade, "A natural policy gradient," *Advances in neural information processing systems*, vol. 14, 2001.
[31] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," *arXiv preprint arXiv:1908.05659*, 2019.
[32] R. Gao and A. Kleywegt, "Distributionally robust stochastic optimization with wasserstein distance," *Mathematics of Operations Research*, vol. 48, no. 2, pp. 603–655, 2023.
[33] A. Shapiro, "Distributionally robust stochastic programming," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2258–2275, 2017.
[34] D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford, "Large-scale methods for distributionally robust optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8847–8860, 2020.
[35] J. Jin, B. Zhang, H. Wang, and L. Wang, "Non-convex distributionally robust optimization: Non-asymptotic analysis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2771–2782, 2021.
[36] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*. PMLR, 2015, pp. 1889–1897.
[37] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.
[38] K. Nordström, "Convexity of the inverse and moore–penrose inverse," *Linear algebra and its applications*, vol. 434, no. 6, pp. 1489–1512, 2011.