ORVIT: Near-Optimal Online Distributionally Robust Reinforcement Learning

Debamita Ghosh[1], **George K. Atia**[1,2], **Yue Wang**[1,2]

[1]Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816
[2]Department of Computer Science, University of Central Florida, Orlando, FL 32816
{de881780, george.atia, yue.wang}@ucf.edu

Abstract

Reinforcement learning (RL) faces significant challenges in real-world deployments due to the sim-to-real gap, where policies trained in simulators often underperform in practice due to mismatches between training and deployment conditions. Distributionally robust RL addresses this issue by optimizing worst-case performance over an uncertainty set of environments and providing an optimized lower bound on deployment performance. However, existing studies typically assume access to either a generative model or offline datasets with broad coverage of the deployment environment—assumptions that limit their practicality in unknown environments without prior knowledge. In this work, we study the more realistic and challenging setting of online distributionally robust RL, where the agent interacts only with a single unknown training environment while aiming to optimize its worst-case performance. We focus on general f-divergence-based uncertainty sets, including χ^2 and KL divergence balls, and propose a computationally efficient algorithm with sublinear regret guarantees under minimal assumptions. Furthermore, we establish a minimax lower bound of online learning, demonstrating the near-optimality of our approach. Extensive experiments across diverse environments further confirm the robustness and efficiency of our algorithm, validating our theoretical findings.

1 Introduction

Reinforcement learning (RL) is a powerful framework for solving complex decision-making problems and has achieved major success in simulation-based domains like video games (Silver et al. 2016; Zha et al. 2021) and generative AI (Ouyang et al. 2022; Du et al. 2023; Cao et al. 2024). However, applying RL to real-world domains such as autonomous driving (Kiran et al. 2021) and healthcare (Wang et al. 2018) remains challenging. The major challenge is due to the infeasibility of direct real-world training under these environments, where RL agents are instead generally trained in simulation and then deployed in the real world. However, real-world deployments are often susceptible to environment uncertainties from, e.g., unpredictable noise, unmodeled perturbations, and even adversarial attacks, which cannot be fully captured by the training environments (Padakandla, KJ, and Bhatnagar 2020; Rajeswaran et al. 2017), resulting in the *sim-to-real gap* and severe performance degradation in practical deployments (Kober, Bagnell, and Peters 2013; Peng et al. 2018; Zhao, Queralta, and Westerlund 2020).

Distributionally robust RL (Iyengar 2005; Pinto et al. 2017) offers a promising way to bridge the sim-to-real gap by training policies that remain reliable under environment shifts. It handles uncertainty by defining a set of plausible environments centered around the training model and then optimizing for the worst-case scenario conservatively. This strategy ensures a guaranteed lower bound on performance when the true environment lies within the uncertainty set, while also enhancing generalization and robustness to unforeseen conditions (Vinitsky et al. 2020; Hou et al. 2020; Rajeswaran et al. 2017; Pattanaik et al. 2018).

Despite its extensive studies, most existing methods on distributionally robust RL cannot be directly adapted to real-world applications due to their restrictive assumptions on data collection. Most studies either rely on a *generative model* of the training environment (Panaganti and Kalathil 2022; Xu, Panaganti, and Kalathil 2023; Shi et al. 2023) that can freely generate data, or assume *offline datasets* that comprehensively cover the unknown optimal policy (Blanchet et al. 2023; Shi and Chi 2024; Tang, Liu, and Xu 2024; Wang, Sun, and Zou 2024; Liu and Xu 2024b; Panaganti et al. 2022; Wang, Shi, and Chi 2024), which generally cannot be guaranteed in practice, as the practical environments are unknown and data is often sparse, and the agent needs to explore the environments by itself.

This motivates the studies of distributionally robust RL with *online interactions* (Liu, Wang, and Xu 2024; Lu et al. 2024; Liu and Xu 2024a; He et al. 2025), where the agent strategically explores the unknown training environment and optimizes for the worst-case performance. Due to the mismatch between the environment that generates training data (nominal), and the one used to evaluate robustness (worst-case), online distributionally robust RL is an off-target problem (Liu and Xu 2024b; Holla 2021), leading to a major challenge known as the *information deficit*: the states covered by the worst-case or deployment environment may not be visited during training, yet the agent must still act reliably in these unfamiliar states (He et al. 2025). This lack of exposure can significantly increase the difficulties of online learning.

Existing studies of online distributionally robust RL generally addresses the challenges through additional

structural assumptions, e.g., existence of fail-states (Liu, Wang, and Xu 2024; Liu and Xu 2024a; Lu et al. 2024). These assumptions greatly simplify the problem by making worst-case shifts predictable, effectively eliminating the information deficit and allowing efficient learning. A more recent work (He et al. 2025) bypasses these structural assumptions, but relies on some coverage-type assumption. However, in realistic settings where deployment dynamics are unknown, these assumptions are impractical or hard to justify, greatly limiting the practical use of these studies. A critical question hence naturally arises:

Can we develop an online distributionally robust RL algorithm with near-optimal sample-complexity?

1.1 Contributions

In this paper, we answer this question by designing algorithms for online distributionally robust RL with near-optimal sample complexity, without relying on structural assumptions such as fail-states. Our major contributions are summarized as follows [1].

- We propose an optimistic model-based meta-algorithm for online distributionally robust RL, named f-ORVIT, with implementations under two important uncertainty set structures: χ^2-ORVIT and KL-ORVIT. These algorithms use plug-in estimates of the nominal kernel and introduce data-driven penalty terms tailored to the respective uncertainty sets. Notably, our algorithm is simpler and admits a more efficient implementation than the one in (He et al. 2025), which requires an additional optimization oracle.
- We prove that χ^2-ORVIT and KL-ORVIT are data-efficient: they achieve an ε-optimal robust policy with $\tilde{\mathcal{O}}\big(H^5(1+\sigma)SA/\varepsilon^2\big)$ and $\tilde{\mathcal{O}}\big(H^5\exp(2H^2)SA/(P^\star_{\min}\varepsilon^2)\big)$ samples, respectively. Here, $P^\star_{\min}$ denotes the minimum positive entry of the nominal transition probability induced by the optimal robust policy. Our results do not require any additional assumptions and achieve improved complexity compared to previous work (He et al. 2025) (Refer to Table 1 for details and Section 3 for relevant notation).
- We further develop hard instances to derive the minimax lower bound for online distributionally robust RL, highlighting the fundamental difficulty of the problem. Our results show that for any online algorithm, there exists a hard instance requiring a sample complexity of at least $\Omega\big(H^5(1+\sigma)SA/\varepsilon^2\big)$ and $\Omega\big(H^5SA/(P^\star_{\min}\varepsilon^2)\big)$ to find an ε-optimal policy under χ^2-RMDP and KL-RMDP. This demonstrates the near-optimality of our algorithm, which is minimax-optimal up to logarithmic factors under the χ^2 set, and matches the lower bound up to the dependency on H under the KL set.
- We validate our methods through extensive experiments on the Gambler's problem and Frozen Lake, demonstrating strong performance under significant distribution shifts and supporting our theoretical findings.

[1]The extended and complete version of our paper can be found at https://arxiv.org/abs/2508.03768.

2 Related Work

In this section, we briefly review the most relevant literature on distributionally robust and risk-sensitive RL.

Distributionally Robust RL. The framework of distributionally robust RL is first proposed and studied in (Iyengar 2005; Nilim and El Ghaoui 2005; Wiesemann, Kuhn, and Rustem 2013; Mannor, Mebel, and Xu 2016). Subsequent learning results for distributionally robust RL under *generative* models (Yang, Zhang, and Zhang 2022; Panaganti and Kalathil 2022; Xu, Panaganti, and Kalathil 2023; Shi et al. 2023) or *offline* data (Shi and Chi 2024; Blanchet et al. 2023; Tang, Liu, and Xu 2024; Wang, Sun, and Zou 2024; Liu and Xu 2024b; Panaganti et al. 2022; Wang, Shi, and Chi 2024) achieve sample complexity guarantees but rely on strong data-access assumptions and do not address the purely online, single-environment setting considered here.

Online Distributionally Robust RL and Our Contributions. Recent advances move toward interactive, robust RL with unknown dynamics. Several works study online robust RL or robust control under structural or coverage conditions: (Liu and Xu 2024a; Liu, Wang, and Xu 2024; Lu et al. 2024) analyze TV-based uncertainty sets and obtain upper/lower bounds by assuming, for example, fail-states or vanishing minimal robust values, which restrict how the worst-case model can differ from the nominal one and mitigate the information-deficit issue. (He et al. 2025) further removes such structural constraints but requires a supremal visitation ratio (a strong coverage condition) to ensure sufficient exploration. In contrast, we study finite-horizon online robust RL with general (s,a)-rectangular f-divergence ambiguity and develop algorithms specialized to χ^2 and KL sets that achieve near-optimal regret and sample complexity *without* fail-state assumptions, vanishing-value conditions, or visitation-ratio requirements.

Risk-Sensitive RL. Our f-divergence-based formulation is mathematically connected to dual representations used in coherent and entropic risk measures (Ahmadi-Javid 2012; Cheridito and Li 2009) and to risk-sensitive RL methods based on entropic or alternative risk criteria (e.g., entropic-VaR and Gini deviation) (Ni and Lai 2022; Luo et al. 2023). However, these approaches typically assume known or generative models, or focus on asymptotic performance, and do not provide online, instance-wise regret guarantees under f-divergence uncertainty sets. By contrast, our work operationalizes these dual ideas in the fully online setting with unknown dynamics and provides sharp guarantees tailored to χ^2 and KL ambiguity.

3 Preliminaries and Problem Formulation

Distributionally Robust Markov Decision Process (RMDPs). Distributionally robust RL can be formulated as an episodic finite-horizon distributionally RMDP (Iyengar 2005), represented by $(\mathcal{S},\mathcal{A},H,\mathcal{P},r)$, where the set $\mathcal{S}=\{1,\dots,S\}$ is the finite state space, $\mathcal{A}=\{1,\dots,A\}$ is the finite action space, H is the horizon length, $r=\{r_h\}_{h=1}^H$ is

the collection of reward functions, where each $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$, and $\mathcal{P} = \{\mathcal{P}_h\}_{h=1}^H$ is an uncertainty set of transition kernels. At step h, the agent is at state s_h and takes an action a_h, receives the reward $r_h(s_h, a_h)$, and is transited to the next state s_{h+1} following an arbitrarily transition kernel $P_h \in \mathcal{P}_h$.

We consider the standard (s,a)-rectangular uncertainty set with divergence ball-structure (Wiesemann, Kuhn, and Rustem 2013). Specifically, there is a collection of *nominal* transition kernels $P^\star = \{P_h^\star\}_{h=1}^H$, where each $P_h^\star : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$. The uncertainty set, centered around the nominal transition kernel, is defined as $\mathcal{P} = \mathcal{U}^\sigma(P^\star) = \bigotimes_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \mathcal{U}_h^\sigma(s,a)$, and $\mathcal{U}_h^\sigma(s,a) = \big\{P \in \Delta(\mathcal{S}) : D(P, P_h^\star(\cdot|s,a)) \le \sigma\big\}$, containing all the transition kernels that differ from $P^\star$ up to some uncertainty level $\sigma \ge 0$, under some probability divergence functions (Iyengar 2005; Panaganti and Kalathil 2022; Yang, Zhang, and Zhang 2022). We mainly focus on the f-divergence uncertainty set, as defined below (a more detail discussion is provided in (Ghosh, Atia, and Wang 2025, Sec. 3)):

Definition 1 (f-Divergence Uncertainty Set)**.** *For each (s,a) pair, the uncertainty set is defined as:*

$$\mathcal{U}_h^\sigma(s,a) = \Big\{P \in \Delta(\mathcal{S}) : D_f\Big(P, P_h^\star(\cdot|s,a)\Big) \le \sigma\Big\}, \quad (1)$$

where $D_f\big(P, P_h^\star(\cdot|s,a)\big) = \sum_{s'\in\mathcal{S}} f\left(\frac{P(s')}{P_h^\star(s'|s,a)}\right) P_h^\star(s'|s,a)$ is the f-divergence (Sason and Verdú 2016).

We refer to the distributionally RMDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$ with $\mathcal{P}$ being a f-divergence uncertainty set as an f-RMDP.

Policy and Robust Value Function. The agent's strategy of taking actions is captured by a Markov policy $\pi := \{\pi_h\}_{h=1}^H$, with $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ for each step $h \in [H]$, where $\pi_h(\cdot|s)$ is the probability of taking actions at the state s in step h. In RMDPs, the performance of a policy is captured by the worst-case performance, defined as the robust value functions. Specifically, for any policy π and step $h \in [H]$, we define the *robust value function* and the *robust state-action value function* as the expected cumulative reward under the worst-case transition kernel within the uncertainty set:

$$V_h^{\pi,\sigma}(s) = \inf_{P\in\mathcal{U}^\sigma(P^\star)} \mathbb{E}_{\pi,P}\left[\sum_{t=h}^H r_t(s_t,a_t)\Big| s_h = s\right], \quad (2)$$

$$Q_h^{\pi,\sigma}(s,a) = \inf_{P\in\mathcal{U}^\sigma(P^\star)} \mathbb{E}_{\pi,P}\left[\sum_{t=h}^H r_t(s_t,a_t)\Big| s_h = s, a_h = a\right],$$

where the expectation is taken with respect to the state-action trajectories induced by policy π under the transition P.

The goal of RMDP is to find the optimal robust policy $\pi^\star := \{\pi_h^\star\}$ that maximizes the robust value function:

$$\pi^\star = \arg\max_{\pi\in\Pi} V_1^{\pi,\sigma}(s_1), \forall s_1 \in \mathcal{S}, \quad (3)$$

where Π is the set of policies. In other words, the optimal robust policy $\pi^\star$ maximizes the worst case expected total rewards in all possible testing environments. For simplicity and without loss of generality, we assume in the sequel that the initial state $s_1 \in \mathcal{S}$ is fixed.

Algorithm 1: Optimistic Robust Value Iteration with f-Divergence Uncertainty Set (f-ORVIT)

Input: uncertainty level $\sigma > 0$.
Initialize: Dataset $\mathbb{D} = \emptyset$
for episode $k = 1, \dots, K$ **do**
  * **NOMINAL TRANSITION ESTIMATION** *
    Compute the transition kernel estimator $\widehat{P}_h^k(s,a,s')$ as given in (6).
  * **OPTIMISTIC ROBUST PLANNING** *
    Set $\overline{V}_{H+1}^k(\cdot) = \underline{V}_{H+1}^k(\cdot) = 0$
    for step $h = H, \dots, 1$ **do**
      for $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
        Update $\overline{Q}_h^k(s,a)$ by (7) and $\underline{Q}_h^k(s,a)$ by (8).
      end for
      for $\forall s \in \mathcal{S}$ **do**
        Update $\pi_h^k(\cdot)$, $\overline{V}_h^k(\cdot)$ and $\underline{V}_h^k(\cdot)$ by (12).
      end for
    end for
  * **EXECUTION OF POLICY AND DATA COLLECTION** *
    Receive initial state $s_1^k \in \mathcal{S}$
    for step $h = 1, \dots, H$ **do**
      Take action $a_h^k \sim \pi_h^k(\cdot \mid s_h^k)$, observe reward $r_h(s_h^k, a_h^k)$ and next state s_{h+1}^k.
    end for
    Set $\mathbb{D} = \mathbb{D} \cup \{(s_h^k, a_h^k, s_{h+1}^k)\}_{h=1}^H$.
end for
Output: Randomly (uniformly) return a policy from $\{\pi^k\}_{k=1}^K$.

Online Distributionally Robust RL. In this work, we consider distributionally robust RL with online interaction. In particular, the agent aims to learn the optimal robust policy $\pi^\star$ in (3), through active interactions with the nominal environment $P^\star$ over $K \in \mathbb{N}$ episodes. At the beginning of episode k, the agent observes initial state s_1^k and selects a policy π^k based on its history. It then collects a trajectory by executing π^k in $P^\star$, and updates the policy for the next episode. Therefore, the goal of the agent is to minimize the *cumulative robust regret* over K episodes, defined as

$$\text{Regret}(K) := \sum_{k=1}^K \left[V_1^{\star,\sigma}(s_1^k) - V_1^{\pi^k,\sigma}(s_1^k)\right]. \quad (4)$$

Note that this robust regret extends the regret in standard MDPs (Auer, Jaksch, and Ortner 2008) by measuring the cumulative robust value gap between the optimal policy $\pi^\star$ and the learner's policies $\{\pi^k\}_{k=1}^K$.

We also evaluate performance through *sample complexity*, defined as the minimum number of samples $T = KH$ needed to learn an ε-optimal robust policy $\widehat{\pi}$ that satisfies

$$V_1^{\star,\sigma}(s_1) - V_1^{\widehat{\pi},\sigma}(s_1) \le \varepsilon. \quad (5)$$

4 Optimistic Robust Value Iteration (ORVIT)

In this section, we introduce *Optimistic Robust Value Iteration with f-Divergence Uncertainty Set (f-ORVIT)*, a meta-algorithm for episodic finite-horizon RMDPs with interactive data collection under f-divergence uncertainty sets, as defined in Definition 1. f-ORVIT is flexible and can be applied to various f-divergences, with a focus on the χ^2-divergence and KL-divergence. The algorithm, detailed in Algorithm 1, balances exploration and exploitation by building confidence intervals directly around the robust value function, avoiding the complexity of modeling full transition dynamics. By leveraging the structure of the f-divergence, it uses adaptive bonuses that reflect both uncertainty and robustness. Inspired by UCB-VI (Azar, Osband, and Munos 2017), this leads to tighter confidence bounds, less dependence on state space size, and more efficient learning in uncertain environments.

4.1 Algorithm Design: f-ORVIT

Our algorithm follows a value iteration framework and integrates optimistic estimation to derive an optimistic estimation of the robust value function. In each episode k, f-ORVIT proceeds in three stages as follows.

Stage 1: Nominal Transition Estimation (Line 4). At the beginning of each episode $k \in [K]$, we maintain an estimate of the transition kernel $P^\star$ of the training environment by using the historical data $\mathbb{D} = \{(s_h^\tau, a_h^\tau, s_{h+1}^\tau)\}_{\tau=1,h=1}^{k-1,H}$ collected from the interaction with the training environment. Specifically, f-ORVIT updates the empirical transition kernel for $(h, s, a, s') \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, as follows

$$\widehat{P}_h^k(s'|s,a) = \begin{cases} \frac{N_h^k(s,a,s')}{N_h^k(s,a)}, & \text{if } N_h^k(s,a) > 0 \\ 1/|\mathcal{S}|, & \text{if } N_h^k(s,a) = 0, \end{cases} \tag{6}$$

where the counts $N_h^k(s,a,s')$ and $N_h^k(s,a)$ are calculated on the current dataset $\mathbb{D}$ by $N_h^k(s,a,s') = \sum_{\tau=1}^{k-1} \mathbf{1}\{(s_h^\tau, a_h^\tau, s_{h+1}^\tau) = (s,a,s')\}$ and $N_h^k(s,a) = \sum_{s' \in \mathcal{S}} N_h^k(s,a,s')$. Our algorithm follows a model-based approach, as it requires explicit model estimation. While this incurs a large memory cost, we emphasize that distributionally robust RL is inherently challenging in the model-free setting: the worst-case expectation is non-linear in the nominal transition kernel, making model-free estimation either biased or extremely sample-inefficient (Wang et al. 2023; Liu et al. 2022; Wang, Zou, and Wang 2024; Zhang et al. 2025).

Stage 2: Optimistic Robust Planning (Lines 5–13). Given the empirical transition model $\widehat{P}^k$ obtained, f-ORVIT performs optimistic robust planning to construct the policy π^k for episode k. Specifically, we construct an optimistic estimation of the robust value function for policy execution. Such an approach, known as the Upper-Confidence-Bound (UCB) method, is shown to be effective in online interactive learning in vanilla RL (Azar, Osband, and Munos 2017; Zanette and Brunskill 2019; Zhang, Ji, and Du 2021; Ménard et al. 2021; Domingues et al. 2021). Specifically, interacting based on an optimistic estimation encourages the agent to explore the less visited state-action pairs.

Toward this goal, we update our estimation as follows, to ensure the estimation is optimistic. At each episode k, f-ORVIT maintains a bonus term to account for the difference between the robust value function of the estimated model $\hat{P}$ and the true robust value function. We then add this term to the estimation, which is constructed following the robust Bellman equation, to ensure its optimism. Namely, for each $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, we update the estimation as

$$\overline{Q}_h^k(s,a) = \min\left\{\overline{R}_h^k(s,a) + B_{k,h}^f(s,a),\ H\right\}, \tag{7}$$

$$\underline{Q}_h^k(s,a) = \max\left\{\underline{R}_h^k(s,a) - B_{k,h}^f(s,a),\ 0\right\}. \tag{8}$$

Each of these estimates (7) and (8) consists of two components: an estimated robust Bellman operator $\overline{R}_h^k(s,a)$ or $\underline{R}_h^k(s,a)$, computed as

$$\overline{R}_h^k(s,a) = r_h(s,a) + \mathbb{E}_{\widehat{\mathcal{U}}_h^\sigma(s,a)}[\overline{V}_{h+1}^k], \tag{9}$$

$$\underline{R}_h^k(s,a) = r_h(s,a) + \mathbb{E}_{\widehat{\mathcal{U}}_h^\sigma(s,a)}[\underline{V}_{h+1}^k], \tag{10}$$

and a bonus term $B_{k,h}^f(s,a) \geq 0$. We denote $\mathbb{E}_{\mathcal{U}^\sigma(s,a)}[V] := \inf_{P \in \mathcal{U}^\sigma(s,a)} \mathbb{E}_P[V]$. The bonus term is constructed (we will discuss the construction later) to ensure the estimation becomes a confidence interval of the true robust value function, i.e., $Q_h^{\star,\sigma} \in [\underline{Q}_h^k(s,a), \overline{Q}_h^k(s,a)]$, with high probability.

With the optimistic estimation $\overline{Q}_h^k$, we then set the execution policy for the k-th episode as the greedy policy with respect to the optimistic Q-estimate:

$$\pi_h^k(\cdot \mid s) = \arg\max_{a \in \mathcal{A}} \overline{Q}_h^k(s,a), \tag{11}$$

and update the robust value function estimation of $V_h^{\star,\sigma}$ as

$$\overline{V}_h^k(s) = \max_{a \in \mathcal{A}} \overline{Q}_h^k(s,a), \quad \underline{V}_h^k(s) = \max_{a \in \mathcal{A}} \underline{Q}_h^k(s,a). \tag{12}$$

We remark that although the lower estimation in (8) does not affect the choice of the execution policy, it is critical for constructing valid exploration bonus terms and for establishing strong theoretical guarantees, and the algorithm leverages both upper and lower bounds to guide exploration. This strategy–optimistic robust planning–enables structured, uncertainty-aware exploration, effectively balancing the competing objectives of exploration, exploitation, and distributional robustness.

Stage 3: Execution of Policy and Data Collection (Lines 14–20). After evaluating the policy $\{\pi_h^k\}_{h=1}^H$ for episode k, the learner takes action based on π_h^k and observes reward $r_h(s_h^k, a_h^k)$ and next state s_{h+1}^k, which gets appended to the historical dataset collected till episode $k-1$.

4.2 Bonus of f-ORVIT Under RMDP

We then instantiate our meta-algorithm for RMDPs under both χ^2-divergence and KL-divergence by explicitly constructing the corresponding bonus terms and estimation procedures.

- χ^2**-RMDP:** We denote $B^f_{k,h}(s,a) := B^{\chi^2}_{k,h}(s,a)$ as

$$\sqrt{\frac{\sigma c_1 L \mathrm{Var}_{\widehat{P}^k_h(\cdot|s,a)}\left[\left(\frac{\overline{V}^k_{h+1}+\underline{V}^k_{h+1}}{2}\right)\right]}{N^k_h(s,a)\vee 1}} + \sqrt{\frac{\sigma}{K}} \quad (13)$$
$$+ \frac{2\sqrt{\sigma}\mathbb{E}_{\widehat{P}^k_h(\cdot|s,a)}\left[\overline{V}^k_{h+1}-\underline{V}^k_{h+1}\right]}{H} + \frac{3c_2\sqrt{\sigma}H^2SL}{\sqrt{N^k_h(s,a)\vee 1}},$$

  where $L = \log\left(\frac{S^3AH^2K^{3/2}}{\delta}\right)$, and $c_1, c_2 > 0$ are absolute constants. The term δ is a pre-selected failure probability.

- **KL-RMDP:** We denote $B^f_{k,h}(s,a) := B^{\mathrm{KL}}_{k,h}(s,a)$ as

$$\frac{2c_f H}{\sigma}\sqrt{\frac{L}{\left(N^k_h(s,a)\vee 1\right)\widehat{P}^k_{\min,h}(s,a)}} + \sqrt{\frac{1}{K}}, \quad (14)$$

  where $\widehat{P}^k_{\min,h}(s,a) = \min_{s'\in\mathcal{S}}\{\widehat{P}^k_h(s'|s,a) : \widehat{P}^k_h(s'|s,a) > 0\}$, $L = \log\left(\frac{S^3AH^2K^{3/2}}{\delta}\right)$, and $c_f > 0$ is an absolute constant.

Under these constructions, $\overline{Q}^k_h$ and $\underline{Q}^k_h$ remain valid confidence bounds (as shown in Lemma K.1 and K.5 in (Ghosh, Atia, and Wang 2025)). Importantly, we will also show that the carefully designed bonus (13) and (14) ensure the confidence region is tight, resulting in a near-optimal regret bound for our algorithm.

4.3 Theoretical Guarantees

We then develop theoretical analysis of our algorithm, under both χ^2 and KL divergence uncertainty sets.

Regret Bound. We first study the regret bound of our method. For the KL-divergence uncertainty set, we adopt the following standard assumption (Yang, Zhang, and Zhang 2022; Shi et al. 2023), which ensures the regularity of the dual formulation of the distributionally robust optimization over the KL-divergence uncertainty set.

Assumption 1. *We assume there exists a constant $P^\star_{\min} > 0$, such that for any $(h,s,a,s') \in [H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}$, if $P^\star_h(s'|s,a) > 0$, then $P^\star_h(s'|s,a) > P^\star_{\min}$.*

We then present the regret bound of our algorithm.

Theorem 1 (Regret Bound of f-ORVIT)**.** *Consider the χ^2 and KL divergence uncertainty sets. For any $\delta\in(0,1)$ and uncertainty radius $\sigma > 0$, with probability at least $1-\delta$, the regret of our f-ORVIT algorithm with corresponding bonus term as* (13) *and* (14) *can be bounded as:*

- *For χ^2 divergence uncertainty set,*

$$\textit{Regret}(K) = \tilde{\mathcal{O}}\left(\sqrt{H^4(1+\sigma)SAK}\right); \quad (15)$$

- *For KL divergence uncertainty set, under Assumption 1,*

$$\textit{Regret}(K) = \tilde{\mathcal{O}}\left(\sqrt{\frac{H^4\exp\left(2H^2\right)SAK}{P^\star_{\min}\sigma^2}}\right). \quad (16)$$

*where $f(K) = \tilde{\mathcal{O}}(g(K))$ means $f(K) \le c\cdot g(K)\cdot$ **Poly**$(\log(K))$ for some constant c and some polynomial of $\log(K)$.*

Our results imply that our algorithm achieves a sublinear regret of $\tilde{\mathcal{O}}(\sqrt{K})$ in both uncertainty sets, ensuring efficient robust policy learning from interactive data. Notably, the exponential term in the KL case is standard due to the complicated structure of the uncertainty set, e.g., (Blanchet et al. 2023; Panaganti and Kalathil 2022).

Sample Complexity. As a direct consequence, we derive the sample complexity to learn an ε-optimal policy through χ^2-ORVIT and KL-ORVIT. Using a standard online-to-batch conversion (Cesa-Bianchi, Conconi, and Gentile 2001), we have the following results.

Corollary 1 (Sample Complexity of f-ORVIT)**.** *Under the same setup in Theorem 1, with probability at least $1-\delta$, f-ORVIT obtains an ε-optimal policy with*

$$T = KH = \begin{cases} \tilde{\mathcal{O}}\left(\frac{H^5(1+\sigma)SA}{\varepsilon^2}\right), & \chi^2\textit{-RMDP} \\ \tilde{\mathcal{O}}\left(\frac{H^5\exp\left(2H^2\right)SA}{P^\star_{\min}\sigma^2\varepsilon^2}\right), & \textit{KL-RMDP} \end{cases} \quad (17)$$

number of samples.

The sample complexity bound for χ^2-RMDP shows linear dependence on the uncertainty radius σ, consistent with prior generative model results (Shi et al. 2023; Yang, Zhang, and Zhang 2022). In particular, Shi et al. (2023, Theorem 3) proved a near-optimal sample complexity of order $\mathcal{O}\left(\frac{SA(1+\sigma)H^4_\gamma}{\varepsilon^2}\right)$ for infinite-horizon stationary χ^2-RMDP with the effective horizon $H_\gamma = \frac{1}{1-\gamma}$, thus our bound aligns with this result while requiring no access to a generative model (note that the non-stationary nature of finite horizon MDPs generally requires an additional H in complexity (Shi and Chi 2024)), showing the tightness of our algorithm.

Our sample complexity bounds for KL-RMDP align with known KL-robust results under generative and offline settings (Yang, Zhang, and Zhang 2022; Panaganti and Kalathil 2022; Wang, Sun, and Zou 2024; Shi and Chi 2024; Blanchet et al. 2023); for example, (Blanchet et al. 2023) obtains $\tilde{\mathcal{O}}\left(\frac{S^2C^\star H^4 e^H}{\sigma^2\varepsilon^2}\right)$ for offline setting, and our bounds exhibit comparable KL-specific dependence while requiring only online interaction. The exponential and $P^\star_{\min}$-dependent terms in our analysis follow directly from the dual form of the KL ball, where robust expectations involve log-MGFs; this equivalently induces $(P^\star_{\min})^{-2}$-type factors (Blanchet et al. 2023), reflecting the genuine difficulty of protecting against adversarial mass on extremely low-probability states. Because our guarantees hold uniformly for all $\sigma > 0$, they are conservative when σ is very small (the ambiguity set is near-nominal), while necessarily incurring a $1/(P^\star_{\min}\sigma^2)$-type cost when rare transitions dominate, as also observed in (Blanchet et al. 2023; Shi and Chi 2024). Overall, these dependencies capture the fundamental hardness of KL-robust control rather than proof artifacts, indicating that our KL-RMDP bounds are effectively tight relative to the best-known KL-robust benchmarks.

5 Lower Bound for Online RMDP

To further understand the hardness of online distributionally robust RL and assess the tightness of the upper bounds presented in theorem 1, we now study corresponding minimax lower bounds for χ^2-RMDP and KL-RMDP.

Theorem 2 (Minimax Lower Bound of Online Distributionally Robust RL)**.** *For any learning algorithm ξ, there exist an f-RMDP $\mathcal{M}$ with the following regret bound with ξ, as long as $K \geq A$:*

$$\mathbb{E}[\textit{Regret}_{\mathcal{M}}(\xi, K)] = \begin{cases} \Omega\Big(\sqrt{H^4(1+\sigma)SAK}\Big), & \chi^2\textit{-RMDP} \\ \Omega\Big(\sqrt{\frac{H^4SAK}{(1-P^\star_{\min})\sigma^2}}\Big), & \textit{KL-RMDP} \end{cases} \tag{18}$$

where $f(K) = \Omega(g(K))$ means $\limsup_{K\to\infty} \frac{f(K)}{g(K)} > 0$.

Noting that the previous results are all for generative model settings (Shi et al. 2023) or offline setting (Shi and Chi 2024; Liu and Xu 2024b), our results stand for the first minimax lower bound for the more involved online setting. We also note that our upper bounds in Theorem 1 match with the lower bound in Theorem 2 in parameters K, S, A, up to some logarithmic factors. This implies that our f-ORVIT is nearly minimax optimal. Hence, our algorithm is the first online distributionally robust RL algorithm to achieve near-optimal sample complexity without structural assumptions.

6 Comparisons with Prior Works

We then compare our results with two most related prior studies on online distributionally robust RL (Lu et al. 2024; He et al. 2025). Other related works are discussed in (Ghosh, Atia, and Wang 2025, Sec. 2). Compared to the existing works (Lu et al. 2024; He et al. 2025), our results enjoy two major advantages: *better applicability*, and *better sample efficiency*.

The work (Lu et al. 2024) studies online learning under the TV-divergence uncertainty set. Their results and methods heavily rely on the additional assumption of the fail-state condition and vanishing minimal states, which effectively addresses the information deficit in their case. However, such simplifications do not extend to general f-divergences, such as χ^2 or KL, whose worst-case solutions become more difficult (Iyengar 2005). More recent work (He et al. 2025) studies all three uncertainty sets, however, their studies also rely on an assumption of the coverage of the worst-case kernel by the nominal environment. Specifically, they assume a supremal visitation ratio between the visitation distributions under the nominal and the worst-case kernels which is polynomial in S, A, H: $C_{vr} = \textbf{Poly}(S, A, H)$. Such an assumption similarly bypasses the information deficit as in (Lu et al. 2024), inspired by the offline RL literature (Li et al. 2024; Shi et al. 2023). However, both assumptions can be infeasible in practice, as there is no such prior knowledge on the distributionally RMDPs. Moreover, their implementations require an additional optimization oracle. In contrast, our method makes no additional assumptions. We design confidence-aware updates that are fully data-driven and based on the dual representation of uncertainty sets, without any oracle. As a result, our analysis applies broadly to RMDPs with general f-divergences, providing robust learning guarantees while addressing the information deficit.

Moreover, our method has a better data efficiency (also see Table 1). Specifically, ORBIT incurs a regret of $\tilde{\mathcal{O}}\big(\sqrt{C_{vr}S^3AH^4K}\big)$ for χ^2 and $\tilde{\mathcal{O}}\left(\left(1+\frac{H\sqrt{S}}{\sigma P^\star_{\min}}\right)\left(\sqrt{C_{vr}SAH^2K}\right)\right)$ for KL, which are largely sub-optimal. In contrast, our regret bound does not depend on C_{vr}, providing more general results, and improve scaling in parameters, leading to tighter guarantees.

On the other hand, our studies on the minimax lower bound provide an assumption-free result, specifying the dependence on S, A, H, σ, which presents a more detailed and accurate study on distributionally robust online learning compared to (He et al. 2025). Moreover, our lower bounds indicate the near-optimality of our method.

Our method hence leads to both sharper theoretical bounds and more applicable and efficient learning in practice.

7 Numerical Experiments

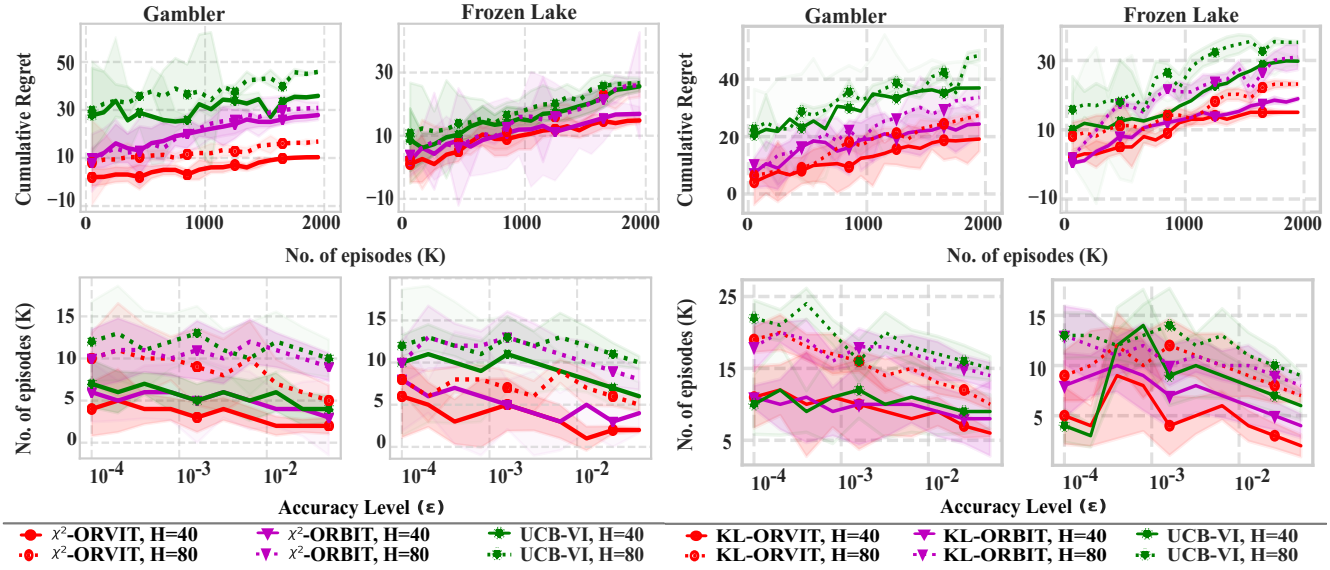
In this section, we evaluate the effectiveness of f-ORVIT using two challenging environments: the Gambler's problem (Sutton, Barto et al. 1998; Shi and Chi 2024) and the Frozen Lake environment (Brockman et al. 2016). We defer the detailed environment setup to (Ghosh, Atia, and Wang 2025, Sec 7.1). We compare with two baselines: (i) the standard non-robust UCB-VI algorithm (Azar, Osband, and Munos 2017), and (ii) the robust ORBIT algorithm (χ^2-ORBIT for χ^2-RMDP; KL-ORBIT for KL-RMDP) for online distributionally robust RL (He et al. 2025). For all methods, we set $\sigma = 0.05/0.1$ in χ^2-RMDP/KL-RMDP, obtain their output policies, and evaluate them under the corresponding RMDPs. We mainly focus on two performance criteria: robust regret (w.r.t. episode number K) and sample complexity to learn an ε-optimal robust policy. Our simulations are averaged over 10 independent runs and reported with confidence intervals.

Robust Regret. We first evaluate and compare the robust regrets of all algorithms. For both the χ^2-divergence and KL-divergence uncertainty models, our proposed algorithms–χ^2-ORVIT and KL-ORVIT–consistently demonstrate superior performance in terms of robust regret. As shown in Figures 1a (for χ^2-RMDP), and Figures 1b (for KL-RMDP), the cumulative regret of all considered algorithms grows sub-linearly with the number of episodes K for a fixed horizon H. Compared to the non-robust UCB-VI, χ^2-ORVIT and KL-ORVIT consistently achieve lower regret across all configurations. The performance gap between our algorithms and UCB-VI becomes even more pronounced as the horizon increases, highlighting the effectiveness of our distributionally robust approach in managing model uncertainty. This strong and consistent performance makes χ^2-ORVIT and KL-ORVIT more efficient and effective choices for complex and uncertain environments, verifying our theoretical results.

Sample Complexity. The sample complexity behavior further reinforces the advantages of our proposed algorithms. Figures 1a (for χ^2-RMDP), along with Figures 1b (for

Model assumption	Algorithm	Regret	Lower bound
Vanishing minimal value (Lu et al. 2024)	TV-OPROVI	$\tilde{\mathcal{O}}\big(\sqrt{\min\{H,\sigma^{-1}\}H^2SAK}\big)$	N/A
Supremal visitation ratio C_{vr} (He et al. 2025)	TV-ORBIT	$\tilde{\mathcal{O}}\big(C_{vr}S^2AH^2+\sqrt{C_{vr}H^4S^3AK}\big)$	$\Omega\big(\sqrt{C_{vr}K}\big)$
	χ^2-ORBIT	$\tilde{\mathcal{O}}\big(C_{vr}S^2AH^2+\sqrt{C_{vr}H^4S^3AK}\big)$	
	KL-ORBIT	$\tilde{\mathcal{O}}\Big(\Big(1+\frac{H\sqrt{S}}{\sigma P^\star_{\min}}\Big)\big(C_{vr}SAH+\sqrt{C_{vr}H^2SAK}\big)\Big)$	
No additional assumption **(our work)**	χ^2-ORVIT	$\tilde{\mathcal{O}}\big(\sqrt{H^4(1+\sigma)SAK}\big)$	$\Omega\big(\sqrt{H^4(1+\sigma)SAK}\big)$
	KL-ORVIT	$\tilde{\mathcal{O}}\Big(\sqrt{H^4\exp(2H^2)SAK/P^\star_{\min}\sigma^2}\Big)$	$\Omega\Big(\sqrt{H^4SAK/P^\star_{\min}\sigma^2}\Big)$

Table 1: Comparison between f-ORVIT and prior results on online RMDPs.



(a) Regret and Sample Complexity comparisons of χ^2-RMDP ($\sigma = 0.05$)

(b) Regret and Sample Complexity comparisons of KL-RMDP ($\sigma = 0.1$)

Figure 1: Performance on Gambler's problem ($S = 20$) and Frozen Lake (4×4) under χ^2-RMDP and KL-RMDP.

KL-RMDP), illustrate the number of episodes K required to achieve a given accuracy level ε. In both the Gambler and Frozen-Lake environments, χ^2-ORVIT and KL-ORVIT consistently require fewer episodes to reach the same level of accuracy compared to the benchmark algorithms. This advantage is especially noticeable at higher accuracy levels, i.e., for larger values of ε. This sample-efficiency advantage becomes more significant as the horizon H increases, indicating that the robust formulation of both χ^2-ORVIT and KL-ORVIT is more sample-efficient and scales better in deeper decision-making scenarios. Overall, the results highlight the superior sample complexity of χ^2-ORVIT and KL-ORVIT, reinforcing their effectiveness in efficiently learning robust policies under distributional uncertainty, specifically χ^2/KL-divergence uncertainty sets.

8 Conclusion

In this paper, we studied online distributionally robust RL under χ^2 and KL divergence uncertainty sets. Our algorithm, f-ORVIT, does not require strong structural assumptions yet still achieves a sub-linear robust regret bound. Moreover, we derived the minimax regret lower for online distributionally robust learning, verifying the near-optimality of our algorithm. Hence, we provided the first tight performance guarantees for online distributionally robust RL under these uncertainty sets. Extensive experiments on diverse environments validate our theoretical results and demonstrate the practical robustness and efficiency of our method.

Acknowledgments

This work was supported by DARPA under Agreement No. HR0011-24-9-0427 and NSF under Award CCF-2106339. The authors thank the anonymous reviewers for their constructive feedback.

References

Ahmadi-Javid, A. 2012. Entropic Value-at-Risk: A New Coherent Risk Measure. *Journal of Optimization Theory and*

Applications, 155(3): 1105–1123.

Auer, P.; Jaksch, T.; and Ortner, R. 2008. Near-Optimal Regret Bounds for Reinforcement Learning. *Advances in Neural Information Processing Systems*, 21.

Azar, M. G.; Osband, I.; and Munos, R. 2017. Minimax Regret Bounds for Reinforcement Learning. In *International Conference on Machine Learning*, 263–272. PMLR.

Blanchet, J.; Lu, M.; Zhang, T.; and Zhong, H. 2023. Double Pessimism is Provably Efficient for Distributionally Robust Offline Reinforcement Learning: Generic Algorithm and Robust Partial Coverage. *Advances in Neural Information Processing Systems*, 36: 66845–66859.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. OpenAI Gym. *arXiv preprint arXiv:1606.01540*.

Cao, Y.; Zhao, H.; Cheng, Y.; Shu, T.; Chen, Y.; Liu, G.; Liang, G.; Zhao, J.; Yan, J.; and Li, Y. 2024. Survey on Large Language Model-Enhanced Reinforcement Learning: Concept, Taxonomy, and Methods. *IEEE Transactions on Neural Networks and Learning Systems*.

Cesa-Bianchi, N.; Conconi, A.; and Gentile, C. 2001. On the Generalization Ability of On-Line Learning Algorithms. *Advances in Neural Information Processing Systems*, 14.

Cheridito, P.; and Li, T. 2009. Risk Measures on Orlicz Hearts. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 19(2): 189–214.

Domingues, O. D.; Ménard, P.; Kaufmann, E.; and Valko, M. 2021. Episodic Reinforcement Learning in Finite MDPs: Minimax Lower Bounds Revisited. In *Algorithmic Learning Theory*, 578–598. PMLR.

Du, Y.; Watkins, O.; Wang, Z.; Colas, C.; Darrell, T.; Abbeel, P.; Gupta, A.; and Andreas, J. 2023. Guiding Pretraining in Reinforcement Learning with Large Language Models. In *Proc. International Conference on Machine Learning (ICML)*, 8657–8677. PMLR.

Ghosh, D.; Atia, G. K.; and Wang, Y. 2025. ORVIT: Near-Optimal Online Distributionally Robust Reinforcement Learning. arXiv:2508.03768.

He, Y.; Liu, Z.; Wang, W.; and Xu, P. 2025. Sample Complexity of Distributionally Robust Off-Dynamics Reinforcement Learning with Online Interaction. In *Forty-second International Conference on Machine Learning*.

Holla, J. A. 2021. *On the Off-Dynamics Approach to Reinforcement Learning*. McGill University (Canada).

Hou, L.; Pang, L.; Hong, X.; Lan, Y.; Ma, Z.; and Yin, D. 2020. Robust Reinforcement Learning with Wasserstein Constraint. *arXiv preprint arXiv:2006.00945*.

Iyengar, G. N. 2005. Robust Dynamic Programming. *Mathematics of Operations Research*, 30(2): 257–280.

Kiran, B. R.; Sobh, I.; Talpaert, V.; Mannion, P.; Al Sallab, A. A.; Yogamani, S.; and Pérez, P. 2021. Deep Reinforcement Learning for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6): 4909–4926.

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

Li, G.; Shi, L.; Chen, Y.; Chi, Y.; and Wei, Y. 2024. Settling the Sample Complexity of Model-Based Offline Reinforcement Learning. *The Annals of Statistics*, 52(1): 233–260.

Liu, Z.; Bai, Q.; Blanchet, J.; Dong, P.; Xu, W.; Zhou, Z.; and Zhou, Z. 2022. Distributionally Robust Q-Learning. In *Proc. International Conference on Machine Learning (ICML)*, 13623–13643. PMLR.

Liu, Z.; Wang, W.; and Xu, P. 2024. Upper and Lower Bounds for Distributionally Robust Off-Dynamics Reinforcement Learning. *arXiv preprint arXiv:2409.20521*.

Liu, Z.; and Xu, P. 2024a. Distributionally Robust Off-Dynamics Reinforcement Learning: Provable Efficiency with Linear Function Approximation. In *International Conference on Artificial Intelligence and Statistics*, 2719–2727. PMLR.

Liu, Z.; and Xu, P. 2024b. Minimax Optimal and Computationally Efficient Algorithms for Distributionally Robust Offline Reinforcement Learning. *Advances in Neural Information Processing Systems*, 37: 86602–86654.

Lu, M.; Zhong, H.; Zhang, T.; and Blanchet, J. 2024. Distributionally Robust Reinforcement Learning with Interactive Data Collection: Fundamental Hardness and Near-Optimal Algorithm. *The Thirty-eighth Annual Conference on Neural Information Processing Systemss*.

Luo, Y.; Liu, G.; Poupart, P.; and Pan, Y. 2023. An Alternative to Variance: Gini Deviation for Risk-averse Policy Gradient. *Advances in Neural Information Processing Systems*, 36: 60922–60946.

Mannor, S.; Mebel, O.; and Xu, H. 2016. Robust MDPs with k-Rectangular Uncertainty. *Mathematics of Operations Research*, 41(4): 1484–1509.

Ménard, P.; Domingues, O. D.; Shang, X.; and Valko, M. 2021. UCB Momentum Q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, 7609–7618. PMLR.

Ni, X.; and Lai, L. 2022. Risk-Sensitive Reinforcement Learning Via Entropic-VaR Optimization. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, 953–959. IEEE.

Nilim, A.; and El Ghaoui, L. 2005. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research*, 53(5): 780–798.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Padakandla, S.; KJ, P.; and Bhatnagar, S. 2020. Reinforcement Learning Algorithm for Non-stationary Environments. *Applied Intelligence*, 50(11): 3590–3606.

Panaganti, K.; and Kalathil, D. 2022. Sample Complexity of Robust Reinforcement Learning with a Generative Model.

In *International Conference on Artificial Intelligence and Statistics*, 9582–9602. PMLR.

Panaganti, K.; Xu, Z.; Kalathil, D.; and Ghavamzadeh, M. 2022. Robust Reinforcement Learning using Offline Data. *Advances in Neural Information Processing Systems*, 35: 32211–32224.

Pattanaik, A.; Tang, Z.; Liu, S.; Bommannan, G.; and Chowdhary, G. 2018. Robust Deep Reinforcement Learning with Adversarial Attacks. In *Proc. International Conference on Autonomous Agents and MultiAgent Systems*, 2040–2042.

Peng, X. B.; Andrychowicz, M.; Zaremba, W.; and Abbeel, P. 2018. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 3803–3810. IEEE.

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A. 2017. Robust Adversarial Reinforcement Learning. In *International Conference on Machine Learning*, 2817–2826. PMLR.

Rajeswaran, A.; Ghotra, S.; Ravindran, B.; and Levine, S. 2017. EPOpt: Learning Robust Neural Network Policies Using Model Ensembles. In *Proc. International Conference on Learning Representations (ICLR)*.

Sason, I.; and Verdú, S. 2016. f-Divergence Inequalities. *IEEE Transactions on Information Theory*, 62(11): 5973–6006.

Shi, L.; and Chi, Y. 2024. Distributionally Robust Model-Based Offline Reinforcement Learning with Near-Optimal Sample Complexity. *Journal of Machine Learning Research*, 25(200): 1–91.

Shi, L.; Li, G.; Wei, Y.; Chen, Y.; Geist, M.; and Chi, Y. 2023. The Curious Price of Distributional Robustness in Reinforcement Learning with a Generative Model. *Advances in Neural Information Processing Systems*, 36: 79903–79917.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; and Hassabis, D. 2016. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, 529(7587): 484–489.

Sutton, R. S.; Barto, A. G.; et al. 1998. *Reinforcement Learning: An Introduction*, volume 1. MIT press Cambridge.

Tang, C.; Liu, Z.; and Xu, P. 2024. Robust Offline Reinforcement Learning with Linearly Structured f-Divergence Regularization. *arXiv preprint arXiv:2411.18612*.

Vinitsky, E.; Du, Y.; Parvate, K.; Jang, K.; Abbeel, P.; and Bayen, A. 2020. Robust Reinforcement Learning using Adversarial Populations. *arXiv preprint arXiv:2008.01825*.

Wang, H.; Shi, L.; and Chi, Y. 2024. Sample Complexity of Offline Distributionally Robust Linear Markov Decision Processes. *arXiv preprint arXiv:2403.12946*.

Wang, L.; Zhang, W.; He, X.; and Zha, H. 2018. Supervised Reinforcement Learning with Recurrent Neural Network for Dynamic Treatment Recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2447–2456.

Wang, Y.; Sun, Z.; and Zou, S. 2024. A Unified Principle of Pessimism for Offline Reinforcement Learning under Model Mismatch. *Advances in Neural Information Processing Systems*, 37: 9281–9328.

Wang, Y.; Velasquez, A.; Atia, G. K.; Prater-Bennette, A.; and Zou, S. 2023. Model-Free Robust Average-Reward Reinforcement Learning. In *Proc. International Conference on Machine Learning (ICML)*, 36431–36469. PMLR.

Wang, Y.; Zou, S.; and Wang, Y. 2024. Model-Free Robust Reinforcement Learning with Sample Complexity Analysis. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*.

Wiesemann, W.; Kuhn, D.; and Rustem, B. 2013. Robust Markov Decision Processes. *Mathematics of Operations Research*, 38(1): 153–183.

Xu, Z.; Panaganti, K.; and Kalathil, D. 2023. Improved Sample Complexity Bounds for Distributionally Robust Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, 9728–9754. PMLR.

Yang, W.; Zhang, L.; and Zhang, Z. 2022. Toward Theoretical Understandings of Robust Markov Decision Processes: Sample Complexity and Asymptotics. *The Annals of Statistics*, 50(6): 3223–3248.

Zanette, A.; and Brunskill, E. 2019. Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds. In *International Conference on Machine Learning*, 7304–7312. PMLR.

Zha, D.; Xie, J.; Ma, W.; Zhang, S.; Lian, X.; Hu, X.; and Liu, J. 2021. DouZero: Mastering DouDizhu with Self-Play Deep Reinforcement Learning. In *Proc. International Conference on Machine Learning (ICML)*, 12333–12344. PMLR.

Zhang, C.; Farhat, Z. U.; Atia, G. K.; and Wang, Y. 2025. Model-Free Offline Reinforcement Learning with Enhanced Robustness. In *Proc. International Conference on Learning Representations (ICLR)*.

Zhang, Z.; Ji, X.; and Du, S. 2021. Is Reinforcement Learning More Difficult Than Bandits? A Near-optimal Algorithm Escaping the Curse of Horizon. In *Conference on Learning Theory*, 4528–4531. PMLR.

Zhao, W.; Queralta, J. P.; and Westerlund, T. 2020. Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 737–744. IEEE.