SAMPLE-EFFICIENT ONLINE DISTRIBUTIONALLY ROBUST REINFORCEMENT LEARNING VIA GENERAL FUNCTION APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The deployment of reinforcement learning (RL) agents in real-world tasks is frequently hampered by performance degradation caused by mismatches between the training and target environments. Distributionally Robust RL (DR-RL) offers a principled framework to mitigate this issue by learning a policy that maximizes worst-case performance over a specified uncertainty set of transition dynamics. Despite its potential, existing DR-RL research faces two key limitations: reliance on prior knowledge of the environment – typically access to a generative model or a large offline dataset – and a primary focus on tabular methods that do not scale to complex problems. In this paper, we bridge these gaps by introducing an online DR-RL algorithm compatible with general function approximation. Our method learns an optimal robust policy directly from environmental interactions, eliminating the need for prior models and enabling application to complex, high-dimensional tasks. Furthermore, our theoretical analysis establishes a near-optimal sublinear regret for the algorithm under the total variation uncertainty set, demonstrating that our approach is both sample-efficient and effective.

1 INTRODUCTION

Reinforcement Learning (RL) has emerged as a powerful paradigm for solving sequential decision-making problems. A central paradigm of RL is online learning, where an agent learns an optimal policy through direct trial-and-error interactions with an unknown environment, without relying on pre-collected datasets or high-fidelity simulators. This learning scheme has fueled significant achievements in complex simulator-based tasks, including video games (Silver et al., 2016; Zha et al., 2021; Berner et al., 2019; Vinyals et al., 2017) and generative AI (Ouyang et al., 2022; Cao et al., 2023; Black et al., 2023; Uehara et al., 2024; Zhang et al., 2024; Du et al., 2023; Cao et al., 2024). However, a critical vulnerability lies at the heart of conventional online RL algorithms. Vanilla RL typically optimizes an agent's policy under the implicit assumption that the environment's dynamics, while stochastic, are fixed and unchanging. In other words, the environment encountered during training is presumed identical to the one at deployment – an assumption often violated in practice and risky for real-world applications. An agent trained in this manner can become highly specialized to the exact conditions experienced during training, leading to a brittle policy that is dangerously unprepared for even minor variations. When deployed in dynamic settings such as autonomous driving (Kiran et al., 2021) or healthcare (Wang et al., 2018), an agent may confront unforeseen shifts, like a sudden change in road friction due to weather. A standard RL agent, never having been trained to consider such possibilities, may suffer a catastrophic drop in performance, leading to unsafe or costly outcomes.

The core of this issue is that vanilla online RL merely optimizes for expected performance within the training environment, but fails to account for potential perturbations or model mismatch upon deployment. Distributionally robust RL (DR-RL) (Iyengar, 2005; Pinto et al., 2017; Hu et al., 2022) offers a promising solution by instead optimizing for the worst-case performance over a pre-defined uncertainty set that captures potential model mismatches. By doing so, DR-RL can learn policies that are inherently resilient to environmental shifts, achieving reliable and safe performance even when encountering new conditions post-deployment (Goodfellow et al., 2014; Vinitsky et al., 2020; Abdullah et al., 2019; Hou et al., 2020; Rajeswaran et al., 2017; Atkeson & Morimoto, 2003;

Morimoto & Doya, 2005; Huang et al., 2017; Kos & Song, 2017; Lin et al., 2017; Pattanaik et al., 2018; Mandlekar et al., 2017). Online DR-RL (He et al., 2025; Liu et al., 2024; Liu & Xu, 2024b; Lu et al., 2024; Ghosh et al., 2025), where the agent directly interacts with the unknown environment but optimizes for the worst-case over some uncertainty set, hence provides a promising approach to overcome the aforementioned issues of online RL and enhance robustness against model mismatches.

Despite its potential, online DR-RL faces two theoretical challenges. The first is due to the *off-target* nature of the objective: training data are generated by nominal dynamics, while robustness is evaluated against worst-case dynamics. The targeted worst-case environment generally differs from the training environment, hence the agent must solve an off-dynamic learning problem (Eysenbach et al., 2020; Liu & Xu, 2024a; Holla, 2021). This can result in an information bottleneck, as samples critical for the target environment may never be observed under the dynamics with which the agent interacts (Lu et al., 2024; Ghosh et al., 2025). Moreover, because the online agent interacts directly with the world, naive exploration that could lead to severe, undesirable consequences is forbidden. This imposes a crucial constraint: the agent must maintain safe and satisfactory performance, even under its worst cases, throughout the entire learning process. Due to these challenges, existing DR-RL mostly assume access to additional data sources, such as a generative model that can freely generate samples (Panaganti & Kalathil, 2022; Xu et al., 2023; Shi et al., 2023), or a comprehensive offline dataset covering the relevant dynamics (Blanchet et al., 2023; Shi & Chi, 2024; Tang et al., 2024; Wang et al., 2024c; Liu & Xu, 2024a; Panaganti et al., 2022; Wang et al., 2024a). Yet in many practical scenarios, such simulators or datasets are unavailable or prohibitively expensive to create, necessitating online DR-RL.

The second challenge is its poor scalability. Most existing DR-RL algorithms are designed for small-scale, tabular problems. Real-world applications, however, often involve vast state-action spaces that render these methods impractical. In standard RL, function approximation techniques (Mnih et al., 2013; Silver et al., 2016; Kober et al., 2013; Li et al., 2016), where a low-dimensional function class is used to approximate the value functions, is the key technique for scaling up to large problems. Yet, its application to DR-RL raises significant theoretical challenges. Due to the inherent model mismatch, the existence of an accurate, low-dimensional approximation of the worst-case value function is not guaranteed. For instance, there may not exist a linear function that properly approximates the worst-case value function (Tamar et al., 2014). Existing attempts to bridge this gap often rely on strong, unverifiable assumptions, such as a small discount factor (Xu & Mannor, 2010; Zhou et al., 2024; Badrinath & Kalathil, 2021) or the environment being modeled as a linear MDP (Ma et al., 2022; Liu & Xu, 2024b;a; Liu et al., 2024; Wang et al., 2024a).

These two gaps naturally lead to one fundamental question: ***Can we develop a sample-efficient online DR-RL algorithm scaling up to large problems, under minimal structural assumptions?***

In this paper, we answer this question by developing an online DR-RL with general function approximations, and deriving convergence guarantees. Our contributions are summarized as follows.

Efficient algorithm design. We develop *Robust Fitted Learning with TV-Divergence Uncertainty Set (RFL-TV)*, the first algorithm for online DR-RL with general function approximation under the total-variation uncertainty set. Our algorithm integrates the optimism principle for efficient exploration within a fitted learning framework. To overcome the challenges of off-dynamic sampling, we introduce a novel functional optimization that reformulates the robust Bellman update. Critically, to manage estimation errors from limited data, we depart from the standard state-action-wise error quantification of tabular UCB methods. Instead, we design a global uncertainty quantifier tailored to our functional optimization, which more effectively captures the aggregate error and guides exploration. This design results in a computationally efficient algorithm suitable for large-scale problems.

Robust coverability. We introduce C_{rcov} (see Definition 3), a new structural measure that captures the inherent statistical difficulty of online DR-RL. This measure quantifies the "information deficit" challenge, namely, how hard it is to optimize for the worst-case model from data generated by a different, potentially more benign, training model. We show that online DR-RL is efficiently learnable if and only if C_{rcov} is finite, establishing it as a fundamental condition for the problem's tractability and a key element in quantifying sample complexity.

Theoretical guarantee. We prove that our RFL-TV finds an ε-optimal robust policy with sample complexity $\tilde{\mathcal{O}}\left(\frac{H^2 \min\{H,\sigma^{-1}\}C_{\mathrm{rcov}}}{\varepsilon^2}\right)$ up to logarithmic factors, where H is the horizon length and σ the uncertainty level. This bound is independent of the state and action space sizes (S, A), confirming our algorithm's scalability beyond tabular methods. The explicit dependence on C_{rcov} further validates our coverability measure as essential for characterizing the complexity of online DR-RL.

2 Related Work

We discuss most related DR-RL works here, and defer the discussion of non-robust RL to Appendix.

Tabular DR-RL: DR-RL is mostly studied under the tabular setting. A substantial body of DR-RL has been developed under the generative-model setting (Clavier et al., 2023; Liu et al., 2022; Panaganti & Kalathil, 2022; Ramesh et al., 2024; Shi et al., 2023; Wang et al., 2023a;b; 2024b; Xu et al., 2023; Yang et al., 2022; 2023; Badrinath & Kalathil, 2021; Li et al., 2022b; Liang et al., 2023), where the agent is assumed to have access to a simulator or with a comprehensive offline dataset (Blanchet et al., 2023; Shi & Chi, 2024; Zhang et al., 2023; Liu & Xu, 2024a; Wang et al., 2024c; Blanchet et al., 2023; Wang et al., 2024a). Recently, limited online DR-RL studies are developed (Dong et al., 2022; Wang & Zou, 2021; Lu et al., 2024; He et al., 2025; Ghosh et al., 2025). The information bottleneck discussed is addressed through adopting some technical assumptions, and sample efficient algorithms are derived. However, all of these works are model-based or value-based, suffering from poor scalability to large-scale problems.

DR-RL with Function Approximation: Existing theoretical DR-RL studies mostly utilize linear function approximation. As discussed, the linear function class may not be complete under the robust Bellman operator, hence no approximation guarantee can be achieved. To address this issue, most studies adopt strong assumptions on the underlying robust MDP, including small discount factor (Xu & Mannor, 2010; Tamar et al., 2014; Xu & Mannor, 2010; Zhou et al., 2024), or that the underlying robust MDP has a linear structure (known as the linear robust MDP) Ma et al. (2022); Liu & Xu (2024b;a); Liu et al. (2024); Wang et al. (2024a); Ma et al. (2022). However, neither of these assumptions can be easily verified in practice. Hence, in this paper, we instead consider a broader function class to bypass these restrictive assumptions. The only work considering general function approximation is Panaganti et al. (2022); however, they consider the offline setting with a globally covered dataset, which is free of the exploration challenge inherent to the online setting.

3 Preliminaries and Problem Formulation

3.1 Distributionally Robust Markov Decision Process (RMDPs).

Distributionally robust RL can be formulated as an episodic finite-horizon RMDP (Iyengar, 2005), represented by $\mathcal{M} := (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$, where the set $\mathcal{S} = \{1, \ldots, S\}$ is the finite state space, $\mathcal{A} = \{1, \ldots, A\}$ is the finite action space, H is the horizon length, $r = \{r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]\}_{h=1}^H$ is the collection of reward functions, 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 arbitrary transition kernel $P_h(\cdot|s_h, a_h) \in \mathcal{P}_h$.

We consider the standard (s, a)-rectangular uncertainty set with divergence ball-structure (Wiesemann et al., 2013). Specifically, there is a *nominal* transition kernel $P^\star = \{P_h^\star\}_{h=1}^H$, where each $P_h^\star : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$[1]. 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 $\mathcal{U}_h^\sigma(s, a) \triangleq \left\{P \in \Delta(\mathcal{S}) : D(P, P_h^\star(\cdot|s, a)) \leq \sigma\right\}$, containing all the transition kernels that differ from $P^\star$ up to some uncertainty level $\sigma \geq 0$, under some probability divergence functions (Iyengar, 2005; Panaganti & Kalathil, 2022; Yang et al., 2022). Specifically, in this paper, we mainly consider uncertainty sets specified by *total-variation (TV)* (Sason & Verdú, 2016), as defined below, and refer to the RMDP defined as an TV-RMDP.

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

$$\mathcal{U}_h^\sigma(s, a) \triangleq \left\{P \in \Delta(\mathcal{S}) : D_{\mathrm{TV}}\left(P, P_h^\star(\cdot|s, a)\right) \leq \sigma\right\}, \tag{1}$$

[1] $\Delta(\cdot)$ denotes the probability simplex over the space.

where for $p, q \in \Delta(\mathcal{S})$, $D_{\mathrm{TV}}(p, q) = \frac{1}{2} \sum_{s' \in \mathcal{S}} |p(s') - q(s')|$ is the TV-divergence.

3.2 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, given any policy π and for each step $h \in [H]$, the *robust value function* and the *robust state-action value function* are defined as the expected accumulative reward under the worst possible transition kernel within the uncertainty set:

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

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

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

The goal of DR-RL is to find the optimal robust policy $\pi^\star := \{\pi_h^\star\}$ that maximizes the robust value function, for some initial state s_1:

$$\pi^\star \triangleq \arg\max_{\pi \in \Pi} V_1^{\pi,\sigma}(s_1), \tag{3}$$

where Π is the set of policies. Such an optimal policy exists and can be obtained as a deterministic policy Iyengar (2005); Blanchet et al. (2023). Moreover, the optimal robust value functions (denoted by $Q_h^{\star,\sigma}, V_h^{\star,\sigma}$), which are the corresponding robust value functions of the optimal policy $\pi^\star$, are shown to be the unique solution to the robust Bellman equations:

$$Q_h^{\star,\sigma}(s,a) = r_h(s,a) + \mathbb{E}_{\mathcal{U}_h^\sigma(s,a)}\left[V_{h+1}^{\star,\sigma}\right], \quad V_h^{\star,\sigma}(s) = \max_{a \in \mathcal{A}} Q_h^{\star,\sigma}(s,a), \tag{4}$$

where $\mathbb{E}_{\mathcal{U}_h^\sigma(s,a)}\left[V_{h+1}^{\star,\sigma}\right] \triangleq \inf_{P_h \in \mathcal{U}_h^\sigma(s,a)} \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\left[V_{h+1}^{\star,\sigma}(s')\right]$.

On the other hand, for any policy π, the corresponding robust value functions also satisfy the following robust Bellman equation for π ((Blanchet et al., 2023, Proposition 2.3)):

$$Q_h^{\pi,\sigma}(s,a) = r_h(s,a) + \mathbb{E}_{\mathcal{U}_h^\sigma(s,a)}\left[V_{h+1}^{\pi,\sigma}\right], \quad V_h^{\pi,\sigma}(s) = \mathbb{E}_{a \sim \pi_h(\cdot|s)}\left[Q_h^{\pi,\sigma}(s,a)\right]. \tag{5}$$

3.3 Online Distributionally Robust RL

In this work, we study distributionally robust RL in an online setting, where the agent's goal is to learn the robust-optimal policy $\pi^\star$ defined in eq. 3 by interacting with the nominal environment $P^\star$ over $K \in \mathbb{N}$ episodes. At the start of episode k, the agent observes the initial state s_1^k, selects a policy π^k based on its history, executes π^k in $P^\star$ to collect a trajectory, and then updates its policy for the next episode. In the online setting, agents cannot freely explore, but instead need to minimize the risk of consequences (under the worst-case) during learning. Hence, the goal to minimize the *cumulative robust regret* over K episodes, defined as

$$\mathrm{Regret}(K) \triangleq \sum_{k=1}^{K} \left[V_1^{\star,\sigma}(s_1^k) - V_1^{\pi^k,\sigma}(s_1^k)\right]. \tag{6}$$

Note that this robust regret extends the regret in standard MDP (Auer et al., 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

$$V_1^{\star,\sigma}(s_1) - V_1^{\widehat{\pi},\sigma}(s_1) \le \varepsilon. \tag{7}$$

4 Robust Bellman Operator with Function Approximation

In this section, we highlight the challenges of online RL and give a step-by-step approach to overcome these challenges.

Functional approximation. When the state–action space is large, learning robust policies from interaction alone is computationally challenging. To address this, we adopt the function approximation technique, where we use a general function class $\mathcal{F} = \{\mathcal{F}_h\}_{h=1}^H$ where $\mathcal{F}_h$ contains some functions $f : \mathcal{S} \times \mathcal{A} \to [0, H]$, to approximate the robust value function $Q_h^{\star,\sigma}$. This function class can be a parametric class with low-dimension parameters, e.g., neural network, to significantly reduce the computation and improve sample efficiency. To ensure effective learning with these function classes, prior work has identified structural conditions that they must satisfy (Russo & Van Roy, 2013; Jiang et al., 2017; Sun et al., 2019; Wang et al., 2020b; Jin et al., 2021; Panaganti et al., 2022). These conditions regulate how the functional class $\mathcal{F}$ interacts with the RMDP dynamics. The most commonly used assumptions are the ***representation conditions***, which require that $\mathcal{F}$ is expressive enough to capture the robust value functions of interest. More specifically, the optimal robust Q-function $Q^{\star,\sigma} \in \mathcal{F}$ (known as realizability) and closure under the robust Bellman operator, namely $\mathcal{T}_h^\sigma \mathcal{F}_{h+1} \subseteq \mathcal{F}_h$ (known as completeness). Following standard studies of function approximation in RL (Jin et al., 2021; Xie et al., 2022; Panaganti et al., 2022; Wang et al., 2019), we adopt the following completeness assumption.

Assumption 1 (Completeness)**.** *For all $h \in [H]$, we have $\mathcal{T}_h^\sigma f_{h+1} \in \mathcal{F}_h$ for all $f_{h+1} \in \mathcal{F}_{h+1}$.*

Per Assumption 1, $\mathcal{F}$ is closed under the robust Bellman operator $\mathcal{T}^\sigma$. Note that, different from standard function approximation RL studies, we do not assume the realizability ($Q^{\star,\sigma} \in \mathcal{F}$). We highlight that realizability may be restricted in RMDPs, for instance, when $\mathcal{F}$ is a linear function class, since the optimal robust value function may not be linear, additional assumptions like linear RMDPs are needed to ensure realizability (Ma et al., 2022; Liu & Xu, 2024b;a; Liu et al., 2024; Wang et al., 2024a; Ma et al., 2022).

Support shifting issue. In RMDPs with a TV-divergence uncertainty set, we face a unique support shifting issue. When the worst-case transition kernel P^ω and the nominal kernel $P^\star$ have different support, states that will be visited under the worst-case may never be visited under the nominal kernel, thus the agent cannot get samples from these states, resulting in an information bottleneck. Notably, the sample complexity of RMDPs with this issue can be exponentially large (Lu et al., 2024). To overcome this challenge, we follow prior work and adopt a standard fail-states assumption (Lu et al., 2024; Liu et al., 2024; Liu & Xu, 2024b; Panaganti et al., 2022) to enable sample-efficient robust RL through interactive data collection.

Assumption 2 (Failure States)**.** *For a TV-RMDP, there exists a set of failure states $\mathcal{S}_F \subseteq \mathcal{S}$, such that $r_h(s, a) = 0$, and $P_h^\star(s'|s, a) = 0$, $\forall a \in \mathcal{A}, \forall s \in \mathcal{S}_F, \forall s' \notin \mathcal{S}_F$.*

Note that this issue does not exist in offline or generative model settings, as the coverage assumption directly ensures the inclusion of the worst-case kernel support.

To better understand the necessity of this assumption, we introduce an intrinsic metric based on visitation measures in both the nominal and the worst-case environments as follows.

Definition 2 (Visitation measure (He et al., 2025))**.** Under TV-RMDP, for any policy π, we denote the worst transition kernel by $P_h^{\omega,\pi}(\cdot|s, a) \triangleq \arg\min_{P_h \in \mathcal{U}_h^\sigma(s,a)} \mathbb{E}_{P_h}[V_{h+1}^{\pi,\sigma}](s, a)$. Furthermore, at step $h \in [H]$, we define $d_h^\pi(\cdot)$ as the visitation measure on $\mathcal{S}$ induced by the policy π under $P^{\omega,\pi}$, and $\mu_h^\pi(\cdot)$ as the visitation measure on $\mathcal{S}$ induced by the policy π under $P^\star$.

Inspired by offline learning (Agarwal et al., 2019; Chen & Jiang, 2019; Wang et al., 2020a; Xie et al., 2021), we further introduce a term to capture the ratio of the visitation measure between the nominal and worst-transition kernels.

Definition 3 (Robust Coverability)**.** Under Definition 2, we define

$$C_{\text{rcov}} := \sup_{\pi \in \Pi, h \in [H]} \left\| \frac{d_h^\pi}{\mu_h^\pi} \right\|_\infty ,$$

as the maximum ratio between the worst-case visitation measure and the nominal visitation measure.

When $C_{\text{rcov}} = \infty$, there exists some state that is visited under the worst-case kernel, but not under the nominal kernel. Thus no data can be obtained for that state, resulting in the support shifting issue. As illustrated in (He et al., 2025), online learning algorithm is efficient only if the coverability measure $C_{\text{rcov}} < \infty$, which, however, does not generally hold in TV cases. However, we show that the failure state Assumption 2 guarantees the finiteness of the robust coverability, thereby providing a necessary condition for efficient online learning algorithms.

Empirical robust Bellman operator and functional optimization. Given the fact that the robust value function is the fixed point of the robust Bellman operator in equation 5, finding the optimal robust policy can be reduced to finding the fixed point of the robust Bellman operator. Since the operator involves an optimization over the S-dimensional uncertainty set which can be inefficient, it is shown that the worst-case operator $\mathbb{E}_{\mathcal{U}^\sigma_h(s,a)}[\cdot]$ has an equivalent duality:

Proposition 1. *Consider an TV-RMDP with TV-uncertainty set $\mathcal{U}^\sigma(P^\star)$ as specified in eq. 1. Let $\mathcal{T}_h$ denote the Bellman operator for layer h. Then, for any $f : \mathcal{S} \times \mathcal{A} \to [0, H]$, the robust Bellman operator $\mathcal{T}$ can be equivalently written as*

$$[\mathcal{T}^\sigma f](s,a) = r(s,a) - \inf_{\eta \in [0, 2H/\sigma]} \left\{ \mathbb{E}_{s' \in P^\star_h(s,a)} \left[\left(\eta - \max_{a'} f(s', a') \right)_+ \right] + \left(\frac{\sigma}{2} - 1 \right) \eta \right\}. \quad (8)$$

Note that the inner optimization problem in eq. 8 is convex in η and only depends on the nominal kernel, hence it can be solved efficiently and empirically. Thus one straightforward approach is to empirically estimate the operator based on this duality and find the fixed point. In fact, all tabular approaches follow this direction. However, this approach becomes infeasible under the large-scale problem with function approximation.

The main difficulty is that Eq. 8 is defined over all (s, a) pairs, with each requiring a separate optimization. Solving all of these to estimate the operator is clearly infeasible, even for moderately sized state–action spaces. Another challenge is that, even if we want to empirically solve the optimization in eq. 8 for some (s, a)-pair, the direct plug-in estimator will be biased. Specifically, $\mathbb{E}[\inf\{\mathbb{E}_{\hat{P}}[\cdot]\}] \neq \inf\{\mathbb{E}_P[\cdot]\}$, due to the non-linearity. To construct an unbiased estimator, techniques like Multi-level Monte-Carlo are introduced (Liu et al., 2022; Wang et al., 2023c), which are also inefficient for large-scale problems.

To address these issues and construct an efficient empirical solution, inspired by (Panaganti et al., 2022), we reformulate the state-action wise optimization as a functional optimization problem as follows. We consider the probability space $(\mathcal{S} \times \mathcal{A}, \Sigma(\mathcal{S} \times \mathcal{A}), \mu)$ and let $\mathcal{L}^1(\mu)$ be the set of all absolutely integrable functions defined on this space.[2]. Under Assumption 2, for any given function $f : \mathcal{S} \times \mathcal{A} \to [0, H]$, we define a dual loss function $\text{Dual}_{loss}(\cdot; f)$ based on the duality in eq. 8 as

$$\text{Dual}_{loss}(g; f) = \mathbb{E}_{(s,a)\sim\mu}\big[\mathbb{E}_{s' \sim P^\star_{s,a}}[(g(s,a) - \max_{a'} f(s', a'))_+] - (1-\sigma) g(s,a)\big], \forall g \in \mathcal{L}^1. \quad (9)$$

In the following lemma, we show that the scalar optimization over η for each (s, a) pair in eq. 8 can be replaced by a single functional optimization w.r.t. the loss function Dual_{loss}.

Lemma 1 (Equivalence Between Pointwise and Functional Minimization of the Dual Loss (Panaganti et al., 2022))**.** *Let Dual_{loss} be the dual loss function defined in eq. 9. Then, for any function $f : \mathcal{S} \times \mathcal{A} \to [0, H]$, we have that*

$$\inf_{g \in \mathcal{L}^1(\mu)} \text{Dual}_{loss}(g; f) = \mathbb{E}_{(s,a)\sim\mu} \left[\inf_{\eta \in [0, 2H/\sigma]} \left\{ \mathbb{E}_{s' \sim P^\star_{s,a}} \left[\left(\eta - \max_{a'} f(s', a') \right)_+ \right] - (1-\sigma)\eta \right\} \right].$$

On the right-hand side of the equation in Lemma 1, the minimization over η is performed pointwise for each (s, a) and minimization is inside the expectation $\mathbb{E}_{(s,a)\sim\mu}[\cdot]$, whereas, on the left-hand side, there is a single minimization over a function $g \in \mathcal{L}^1$ taken outside the expectation. The equivalence follows from the interchange rule for integral functionals by following the result of (Rockafellar & Wets, 1998, Theorem 14.60). Moreover, we consider $\mathcal{L}^1$ as a decomposable space, which allows us to assemble pointwise minimizers into a measurable, integrable selector g that attains the same objective value, thereby justifying the move from pointwise to functional optimization. This interchange is standard in distributionally robust optimization (Shapiro, 2017; Duchi & Namkoong, 2021).

[2]In other words, $\mathcal{L}^1(\mu)$ is the set of all functions $g : \mathcal{S} \times \mathcal{A} \to [0, 2H/\sigma] \subset \mathbb{R}$, such that $\|g\|_{1,\mu}$ is finite.

A related result was first derived in (Panaganti et al., 2022, Lemma 1) for an offline setting under a global coverage assumption. In contrast, our algorithm sets this distribution to be the visitation measure induced by updated policies during interactions, making it inherently non-stationary. Whereas the offline case assumes a fixed dataset distribution, our setting introduces a key analytical challenge: handling this time-varying equivalence.

We then construct our empirical operator based on this functional optimization. Given any dataset $\mathcal{D}$, we define the empirical dual loss function $\widehat{\mathrm{Dual}}_{loss}$ corresponding to the true dual loss Dual_{loss} as

$$\widehat{\mathrm{Dual}}_{loss}(g; f) = \sum_{(s,a,s')\sim\mathcal{D}} \big((g(s,a) - \max_{a'} f(s',a'))_+ - (1-\sigma)g(s,a)\big), \tag{10}$$

which is an empirical estimation of $\mathrm{Dual}_{loss}(g; f)$. For a fixed f, we can find an approximately optimal dual function by minimizing the empirical dual loss, i.e., $\inf_{g\in\mathcal{L}^1} \widehat{\mathrm{Dual}}_{loss}(g; f)$. Note that the solution to this empirical loss can be biased (due to the non-linearity), however, we will show later that this bias and the overall error can be controlled.

Instead of optimizing over $\mathcal{L}^1$, we follow (Panaganti et al., 2022) and use another function class $\mathcal{G} = \{g : \mathcal{S}\times\mathcal{A} \to [0, 2H/\sigma]\}$ to approximate the dual variable functions. Thus, in the optimization problem, instead of taking the infimum over $\mathcal{L}^1$, we will take the infimum over all $\mathcal{G}$. For this to be meaningful, $\mathcal{G}$ should have sufficient representation power. In particular, the result in Lemma 1 should hold approximately even if we replace the infimum over $\mathcal{L}^1$ with infimum over $\mathcal{G}$. We thus further make the following assumption on realizability of $\mathcal{G}$:

Assumption 3. *(Panaganti et al., 2022) For all $f \in \mathcal{F}$ and any π, there exists a uniform constant ξ_{dual} such that $\inf_{g\in\mathcal{G}} \mathrm{Dual}_{loss}(g; f) - \inf_{g\in\mathcal{L}^1(\mu^\pi)} \mathrm{Dual}_{loss}(g; f) \le \xi_{\mathrm{dual}}$.*

With this assumption, we can then find an approximate value of $\mathcal{T}^\sigma(f)$ by first minimizing the empirical loss $\widehat{\mathrm{Dual}}_{loss}(g; f)$ over $\mathcal{G}$:

$$\widehat{g}_f = \arg\min_{g\in\mathcal{G}} \widehat{\mathrm{Dual}}_{loss}(g; f), \tag{11}$$

and applying the operator $\mathcal{T}^\sigma_{\widehat{g}_f}$ to f, where

$$(\mathcal{T}^\sigma_g f)(s,a) \triangleq r(s,a) - \mathbb{E}_{s'\sim P^\star_{s,a}}[(g(s,a) - \max_{a'} f(s',a'))_+]\big) - (1-\sigma)g(s,a). \tag{12}$$

We then show that our construct results in a small approximation error.

Lemma 2. *For a policy π, let $\mu^\pi_h(\cdot)$ be the visitation measure on $\mathcal{S}\times\mathcal{A}$ induced by the policy π under $P^\star$. Then, given a dataset $\mathcal{D}$ collected by π, with probability at least $1-\delta$, it holds that*

$$\sup_{f\in\mathcal{F}} \|\mathcal{T}^\sigma f - \mathcal{T}^\sigma_{\widehat{g}_f} f\|_{1,\mu^\pi} = \mathcal{O}\left(H\min\{H, 1/\sigma\}\sqrt{\frac{2\log(8|\mathcal{G}||\mathcal{F}|/\delta)}{|\mathcal{D}|}} + \xi_{\mathrm{dual}}\right). \tag{13}$$

Our result hence implies that the error of our empirical functional optimization can be controlled. Moreover, the error bound we obtained is w.r.t. $\|\cdot\|_{1,\mu^\pi}$-norm, instead of state-action pair wise, which will later be used to construct our global error quantification term in our algorithm design.

5 Robust Fitting Learning Algorithm

In this section, we utilize our previous constructions and propose our Robust Fitted Learning (RFL) algorithm in Algorithm 1.

Our algorithm follows the standard fitting learning structure. In each step h, we will construct a confidence set $\mathcal{F}^{(k)}$ (Line 9) based on the fitted error under the robust Bellman operator to ensure the inclusion of $Q^{\star,\sigma} \in \mathcal{F}^{(k)}$. As discussed, we utilize our functional optimization based loss function and the error bound in Lemma 2 to construct the set. Namely, given a function f, we first solve the dual-variable approximation through the empirical functional optimization loss as

$$\widehat{g}_f \triangleq \arg\min_{g\in\mathcal{G}} \sum_{(s,a,s')\in\mathcal{D}^{(k)}_h} \Big(g(s,a) - \max_{a'\in\mathcal{A}} f(s',a')\Big)_+ - (1-\sigma)g(s,a). \tag{14}$$

Algorithm 1: Robust Fitted Learning with TV-Divergence Uncertainty Set (RFL-TV)

Input: Function class $\mathcal{F}$, Dual Function class $\mathcal{G}$, confidence width $\beta > 0$, uncertainty level $\sigma > 0$.
Initialize: $\mathcal{F}^{(0)} \leftarrow \mathcal{F},\ \mathcal{D}_h^{(0)} \leftarrow \emptyset\ \forall h \in [H]$
for episode $k = 1, 2, \ldots, K$ **do**
  Set $f^{(k)} \leftarrow \arg\max_{f \in \mathcal{F}^{(k-1)}} f(s_1, \pi_1^f(s_1))$ and $\pi^{(k)} \leftarrow \pi^{f^{(k)}}$
  Execute $\pi^{(k)}$ and obtain a trajectory $(s_1^{(k)}, a_1^{(k)}, r_1^{(k)}), \ldots, (s_H^{(k)}, a_H^{(k)}, r_H^{(k)})$
  Update dataset: $\mathcal{D}_h^{(k)} \leftarrow \mathcal{D}_h^{(k-1)} \cup \{(s_h^{(k)}, a_h^{(k)}, s_{h+1}^{(k)})\}\ \forall h \in [H]$
  $\mathcal{F}_H^{(k)} \leftarrow \{0\}$
  for $h = H-1, ..., 1$ **do**
    Update the confidence set, with notations defined in equation 14 and equation 15:

$$\mathcal{F}_h^{(k)} \leftarrow \left\{ f \in \mathcal{F}_h : L_h^{(k)}(f_h, f_{h+1}, \widehat{g}_{f_{h+1}}) - \min_{f'_h \in \mathcal{F}_h} L_h^{(k)}(f'_h, f_{h+1}, \widehat{g}_{f_{h+1}}) \le \beta, \forall f_{h+1} \in \mathcal{F}_{h+1}^{(k)} \right\}$$

  end for
end for
Output: $\bar{\pi} = \text{unif}(\pi^{(1:K)})$. `For PAC guarantee only.`

We further capture the empirical robust Bellman error through our functional optimization as:

$$L_h^{(k)}(f', f, g) \triangleq \sum_{(s,a,r,s') \in \mathcal{D}_h^{(k)}} \left\{ f'(s,a) - r - \left(g(s,a) - \max_{a' \in \mathcal{A}} f(s', a') \right)_+ + (1-\sigma) g(s,a) \right\}^2. \tag{15}$$

Notably, due to the large-scale of the problem, we construct the confidence set of function classes in a *global* fashion that entails optimizing over f_h for all steps $h \in [H]$ simultaneously (Zanette et al., 2020), instead of constructing error qualifications for each state-action pair as in tabular UCB approaches. More specifically, the confidence set is constructed by considering all the functions that not only minimize the squared robust Bellman error on the collected transition data $\mathcal{D}_h^{(k)}$ in terms of the dual variable function, but also any function whose loss is only slightly larger than the optimal loss over the functional class $\mathcal{F}_h$. We will later design an error quantification error β, to ensure that $Q^{\star,\sigma} \in \mathcal{F}^{(k)}$ with high probability. With the function confidence set which contains $Q^{\star,\sigma}$, we then adopt the optimism principle and choose $\pi^{(k)} = \pi^{f^{(k)}}$ based on the robust value function $f^{(k)} \in \mathcal{F}^{(k)}$ with the most optimistic estimate $f_1(s_1, \pi_1^{(k)}(s_1))$ for the total reward. This will ensure the optimism of our algorithm, and balance the exploration and exploitation.

We highlight that our dataset $\mathcal{D}_h^{(k)}$ is collected under different policies over episodes, thus there does not exist any single policy π such that $\mathcal{D}_h^{(k)} \sim \mu^\pi$. Hence the error quantification we derived in Lemma 2 cannot be directly applied in our analysis. However, as we will show in the next section, we can derive an error quantification and the associated analysis utilizing the robust coverability term defined in Definition 3. Note that this issue does not exist in the offline robust RL setting (Panaganti et al., 2022), as the offline dataset therein is generated by a fixed distribution, whereas our dataset is non-stationary and time-varying, thus the offline analysis is not applicable in our setting.

6 Theoretical Guarantees

We then develop the theoretical guarantees of our algorithm.

Theorem 1. *For any $\delta \in (0,1]$, we set $\beta = \mathcal{O}\Big(\big(H \min\{H, 1/\sigma\}\big) \log\Big(\frac{KH\,|\mathcal{F}||\mathcal{G}|}{\delta}\Big)\Big)$ in RFL-TV. Then under Assumption 1, 2, and 3, there exists an absolute constant c such that with probability at*

least $1-\delta$*, it holds that*[3]

$$\mathrm{Regret}(K) \le \mathcal{O}\Bigg(H\sqrt{C_{\mathrm{rcov}} \cdot H\min\{H, 1/\sigma\}\log\Big(\tfrac{KH\,|\mathcal{F}||\mathcal{G}|}{\delta}\Big)K\log K} + C_{\mathrm{rcov}}\xi_{\mathrm{dual}}\Bigg). \quad (16)$$

As mentioned, analysis of offline robust RL with function approximation (Panaganti et al., 2022) relies on a static distribution which enjoys a global concentratability that covers all distributions induced by all policies and kernels in the uncertainty set. Such an assumption is significantly strong and enables the analysis of the functional error and sample complexity based on this single distribution. However, in our online setting, we do not have such a dataset and we need to explore the environment while maintaining a low regret. We also need to tackle the updated data distribution μ^{π_k} over episodes and the mismatch between the worst-case and nominal kernels, which invalidates the global coverage assumption and requires an episode-specific analysis. To address these issues, we utilize the robust coverability C_{rcov} as a uniform bound of the distribution ratio over episodes, and use it to capture the functional optimization errors and regret bound. Notably, our robust coverability is strictly weaker than the global concentratability in (Panaganti et al., 2022), and our analysis reveals that such a local concentratability quantity is sufficient for sample-efficient exploration.

As an immediate corollary, we obtain the sample complexity for learning an ε-optimal policy with RFL-TV by applying a standard online-to-batch conversion (Cesa-Bianchi et al., 2001).

Corollary 1 (Sample Complexity)**.** *Under the same setup in Theorem 1, with probability at least* $1-\delta$*, the sample-complexity of RFL-TV to obtain an* ε*-optimal robust policy is*

$$T = KH = \mathcal{O}\Bigg(\frac{H^2\min\{H, \sigma^{-1}\}C_{\mathrm{rcov}}\log\Big(\frac{T\,|\mathcal{F}||\mathcal{G}|}{\delta}\Big)\log\Big(\frac{T}{H}\Big)}{\varepsilon^2} + \frac{C_{\mathrm{rcov}}\xi_{\mathrm{dual}}}{\varepsilon}\Bigg) \quad (17)$$

Our result implies that, RFL-TV finds an ε-optimal robust policy within polynomial interactive samples in H and ε^{-1}, with logarithmic dependence on the functional class cardinality. The dependence of these terms match or improve the ones under the tabular setting (He et al., 2025), and hence our algorithm is sample efficient. Moreover, the result does not depend on S, A, implying enhanced scalability to large-scale problems. We note that an algorithm for online DR-RL without function approximation was recently proposed in (Shazman et al., 2025). However, its complexity scales as H^3, quadratically in σ, and linearly in the action number A. In contrast, our analysis accommodates general function classes, replacing the bound from quadratic to linear dependence on σ via $\min\{H, 1/\sigma\}$, and shifts the dependence from action space to the structural term C_{rcov}, yielding a better bound and enhanced scalability to large problems.

Remark 1. *We developed our results in terms of robust coverability, a notion also used and studied in non-robust learning (Xie et al., 2022). There is also a line of work in online RL that employs complexity measures such as Bellman rank (Jiang et al., 2017; Du et al., 2021) and BE dimension (Jin et al., 2021), and we expect our analysis could similarly be adapted to these notions. However, as in non-robust RL, robust coverability measure provides a more faithful and often strictly weaker condition for sample efficiency than the BE dimension or Bellman/Bilinear rank. For example, there exist MDPs with* $C_{\mathrm{cov}} = O(1)$ *but large BE dimension or Bellman/Bilinear ranks (Xie et al., 2022). Hence, we adopt robust coverability as our complexity measure to obtain a tighter regret bound.*

7 Conclusion

In this work, we introduced RFL-TV, an online DR-RL algorithm with general function approximation under TV-divergence uncertainty set. The algorithm implements a fitted robust Bellman update via a functional optimization and replaces state-action bonuses with a global uncertainty quantifier that more effectively guides exploration. We also identified robust coverability C_{rocv} as the structural condition that governs learnability, yielding sharp, scalable sample-efficiency guarantees. We further developed a regret bound of our algorithm that does not scale with problem scales, implying the efficiency and scalability of our method. Our algorithm hence stands for the first online DR-RL algorithm for large scale problems, with minimum structural assumptions. A future direction will be to extend our functional optimization and algorithm design to other general f-divergence (Yang et al., 2022) uncertainty sets.

[3]We assume for simplicity that $|\mathcal{F}|, |\mathcal{G}| < \infty$, but our result can be directly extended to the general infinite case with a standard finite coverage technique (Xie et al., 2022; Panaganti et al., 2022).

REFERENCES

Mohammed Amin Abdullah, Hang Ren, Haitham Bou Ammar, Vladimir Milenkovic, Rui Luo, Mingtian Zhang, and Jun Wang. Wasserstein Robust Reinforcement Learning. *arXiv preprint arXiv:1907.13196*, 2019.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1638–1646. PMLR, 2014.

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and Algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

Christopher G Atkeson and Jun Morimoto. Nonparametric Representation of Policies and Value Functions: A Trajectory-Based Approach. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 1643–1650, 2003.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-Optimal Regret Bounds for Reinforcement Learning. *Advances in neural information processing systems*, 21, 2008.

Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pp. 511–520. PMLR, 2021.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with Large Scale Deep Reinforcement Learning. *arXiv preprint arXiv:1912.06680*, 2019.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training Diffusion Models with Reinforcement Learning. *arXiv preprint arXiv:2305.13301*, 2023.

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. 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, 2023.

Yuanjiang Cao, Quan Z Sheng, Julian McAuley, and Lina Yao. Reinforcement learning for generative ai: A survey. *arXiv preprint arXiv:2308.14328*, 2023.

Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li. Survey on Large Language Model-Enhanced Reinforcement Learning: Concept, Taxonomy, and Methods. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Nicoló Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the Generalization Ability of On-Line Learning Algorithms. *Advances in neural information processing systems*, 14, 2001.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1042–1051. PMLR, 2019.

Pierre Clavier, Erwan Le Pennec, and Matthieu Geist. Towards Minimax Optimality of Model-based Robust Reinforcement Learning. *arXiv preprint arXiv:2302.05372*, 2023.

Jing Dong, Jingwei Li, Baoxiang Wang, and Jingzhao Zhang. Online policy optimization for robust mdp. *arXiv preprint arXiv:2209.13841*, 2022.

Simon S Du, Sham M Kakade, Jason D Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021.

Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. Guiding Pretraining in Reinforcement Learning with Large Language Models. In *Proc. International Conference on Machine Learning (ICML)*, pp. 8657–8677. PMLR, 2023.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Benjamin Eysenbach, Swapnil Asawa, Shreyas Chaudhari, Sergey Levine, and Ruslan Salakhutdinov. Off-Dynamics Reinforcement Learning: Training for Transfer with Domain Classifiers. *arXiv preprint arXiv:2006.13916*, 2020.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Debamita Ghosh, George K Atia, and Yue Wang. Provably near-optimal distributionally robust reinforcement learning in online settings. *arXiv preprint arXiv:2508.03768*, 2025.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*, 2014.

Yiting He, Zhishuai Liu, Weixin Wang, and Pan Xu. Sample Complexity of Distributionally Robust Off-Dynamics Reinforcement Learning with Online Interaction. In *Forty-second International Conference on Machine Learning*, 2025.

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

Linfang Hou, Liang Pang, Xin Hong, Yanyan Lan, Zhiming Ma, and Dawei Yin. Robust reinforcement learning with Wasserstein constraint. *arXiv preprint arXiv:2006.00945*, 2020.

Jiachen Hu, Han Zhong, Chi Jin, and Liwei Wang. Provable Sim-to-real Transfer in Continuous Domain with Partial Observations. *arXiv preprint arXiv:2210.15598*, 2022.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial Attacks on Neural Network Policies. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

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

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 2024.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1704–1713. PMLR, 2017.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep Reinforcement Learning for Autonomous Driving: A Survey. *IEEE transactions on intelligent transportation systems*, 23(6):4909–4926, 2021.

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

Jernej Kos and Dawn Song. Delving into adversarial attacks on deep policies. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Gene Li, Pritish Kamath, Dylan J Foster, and Nati Srebro. Understanding the eluder dimension. *Advances in Neural Information Processing Systems*, 35:23737–23750, 2022a.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

Yan Li, Guanghui Lan, and Tuo Zhao. First-order policy optimization for robust markov decision process. *arXiv preprint arXiv:2209.10579*, 2022b.

Zhipeng Liang, Xiaoteng Ma, Jose Blanchet, Jiheng Zhang, and Zhengyuan Zhou. Single-trajectory distributionally robust reinforcement learning. *arXiv preprint arXiv:2301.11721*, 2023.

Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of Adversarial Attack on Deep Reinforcement Learning Agents. In *Proc. International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 3756–3762, 2017.

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

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

Zhishuai Liu, Weixin Wang, and Pan Xu. Upper and Lower Bounds for Distributionally Robust Off-Dynamics Reinforcement Learning. *arXiv preprint arXiv:2409.20521*, 2024.

Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust Q-learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 13623–13643. PMLR, 2022.

Miao Lu, Han Zhong, Tong Zhang, and Jose Blanchet. Distributionally Robust Reinforcement Learning with Interactive Data Collection: Fundamental Hardness and Near-Optimal Algorithm. *The Thirty-eighth Annual Conference on Neural Information Processing Systemss*, 2024.

Xiaoteng Ma, Zhipeng Liang, Jose Blanchet, Mingwen Liu, Li Xia, Jiheng Zhang, Qianchuan Zhao, and Zhengyuan Zhou. Distributionally Robust Offline Reinforcement Learning with Linear Function Approximation. *arXiv preprint arXiv:2209.06620*, 2022.

Ajay Mandlekar, Yuke Zhu, Animesh Garg, Li Fei-Fei, and Silvio Savarese. Adversarially Robust Policy Learning: Active Construction of Physically-Plausible Perturbations. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3932–3939. IEEE, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Jun Morimoto and Kenji Doya. Robust Reinforcement Learning. *Neural computation*, 17(2):335–359, 2005.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Kishan Panaganti and Dileep Kalathil. Sample Complexity of Robust Reinforcement Learning with a Generative Model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust Reinforcement Learning using Offline Data. *Advances in neural information processing systems*, 35:32211–32224, 2022.

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust deep reinforcement learning with adversarial attacks. In *Proc. International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2040–2042, 2018.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust Adversarial Reinforcement Learning. In *International conference on machine learning*, pp. 2817–2826. PMLR, 2017.

Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. EPOpt: Learning Robust Neural Network Policies Using Model Ensembles. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

Shyam Sundhar Ramesh, Pier Giuseppe Sessa, Yifan Hu, Andreas Krause, and Ilija Bogunovic. Distributionally robust model-based reinforcement learning with large state spaces. In *International Conference on Artificial Intelligence and Statistics*, pp. 100–108. PMLR, 2024.

R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

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

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

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

Tamir Shazman, Idan Lev-Yehudi, Ron Benchetit, and Vadim Indelman. Online robust planning under model uncertainty: A sample-based approach. *arXiv preprint arXiv:2509.10162*, 2025.

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

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The Curious Price of Distributional Robustness in Reinforcement Learning with a Generative Model. *Advances in Neural Information Processing Systems*, 36:79903–79917, 2023.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust MDPs using function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 181–189. PMLR, 2014.

Cheng Tang, Zhishuai Liu, and Pan Xu. Robust Offline Reinforcement Learning with Linearly Structured f-Divergence Regularization. *arXiv preprint arXiv:2411.18612*, 2024.

Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding Reinforcement Learning-Based Fine-Tuning of Diffusion Models: A Tutorial and Review. *arXiv preprint arXiv:2407.13734*, 2024.

Eugene Vinitsky, Yuqing Du, Kanaad Parvate, Kathy Jang, Pieter Abbeel, and Alexandre Bayen. Robust reinforcement learning using adversarial populations. *arXiv preprint arXiv:2008.01825*, 2020.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. Starcraft II: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

He Wang, Laixi Shi, and Yuejie Chi. Sample Complexity of Offline Distributionally Robust Linear Markov Decision Processes. *arXiv preprint arXiv:2403.12946*, 2024a.

Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. 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*, pp. 2447–2456, 2018.

Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020a.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020b.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A Finite Sample Complexity Bound for Distributionally Robust Q-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3370–3398. PMLR, 2023a.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. On the foundation of distributionally robust reinforcement learning. *arXiv preprint arXiv:2311.09018*, 2023b.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample Complexity of Variance-Reduced Distributionally Robust Q-Learning. *Journal of Machine Learning Research*, 25(341):1–77, 2024b.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Yue Wang and Shaofeng Zou. Online Robust Reinforcement Learning with Model Uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

Yue Wang, Alvaro Velasquez, George K Atia, Ashley Prater-Bennette, and Shaofeng Zou. Model-free robust average-reward reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 36431–36469. PMLR, 2023c.

Yue Wang, Zhongchang Sun, and Shaofeng Zou. A Unified Principle of Pessimism for Offline Reinforcement Learning under Model Mismatch. *Advances in Neural Information Processing Systems*, 37:9281–9328, 2024c.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov Decision Processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2505–2513, 2010.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved Sample Complexity Bounds for Distributionally Robust Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 9728–9754. PMLR, 2023.

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

Wenhao Yang, Han Wang, Tadashi Kozuno, Scott M Jordan, and Zhihua Zhang. Robust markov decision processes without model estimation. *arXiv preprint arXiv:2302.01248*, 2023.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020.

Daochen Zha, Jingru Xie, Wenye Ma, Sheng Zhang, Xiangru Lian, Xia Hu, and Ji Liu. DouZero: Mastering DouDizhu with Self-Play Deep Reinforcement Learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 12333–12344. PMLR, 2021.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason D Lee. Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634*, 2022.

Runyu Zhang, Yang Hu, and Na Li. Soft Robust MDPs and Risk-Sensitive MDPs: Equivalence, Policy Gradient, and Sample Complexity. *arXiv preprint arXiv:2306.11626*, 2023.

Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Ruida Zhou, Tao Liu, Min Cheng, Dileep Kalathil, PR Kumar, and Chao Tian. Natural actor-critic for robust reinforcement learning with function approximation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2024.

A USE OF LARGE LANGUAGE MODELS

We used ChatGPT strictly as a general-purpose assist tool for typesetting and language polishing. In particular, it helped with (i) grammar, style, and readability improvements, and (ii) LaTeX formatting tasks such as managing algorithm placement, cleaning BibTEX entries and citation styles, and resolving compile issues (e.g., Type-3 font warnings and package conflicts).

All ideas, derivations, and final claims were developed, verified, and validated by the authors. The authors take full responsibility for the content of this paper.

B RELATED WORKS: NON-ROBUST RL WITH FUNCTIONAL APPROXIMATION

Function approximation has been widely studied in non-robust RL. While extensive studies are developed for offline RL with general function approximation, e.g., (Zhan et al., 2022; Jiang & Xie, 2024; Wang et al., 2020a), we mainly discuss online RL here, which requires the agent to explore while learning actively.

A foundational direction is the development of complexity measures that capture when online RL with function approximation is tractable. The Eluder dimension (Li et al., 2022a; Russo & Van Roy, 2013) provides a measure of the sequential complexity of a function class. Online RL algorithms have been developed that use optimism based on confidence sets constructed around the true value function, and the size of these confidence sets and the magnitude of the exploration bonus are constructed based on the Eluder dimension (Wang et al., 2020b).

Since the Eluder dimension merely captures the complexity of the function class in isolation, other measures have been proposed that capture the interaction between $\mathcal{F}$ and the MDP dynamics. Bellman rank (Jiang et al., 2017) and Witness rank (Sun et al., 2019) are later then developed to capture these interactions, and are later unified by the Bellman–Eluder dimension Jin et al. (2021). It directly measures the complexity relevant to value-based RL, i.e., the difficulty of learning to minimize Bellman errors.

More recently, attention has turned to coverage conditions as the key lens for understanding learnability in online RL. Xie et al. (2022) introduced the notion of coverability, which provides a sharp characterization of when exploration with function approximation is sample-efficient. Their results demonstrate that coverability is both necessary and sufficient, thereby subsuming earlier assumptions such as concentrability or bounded Bellman rank. Complementary hardness results (Foster et al., 2021; Du et al., 2021) show that, without such structural or coverage conditions, online RL in rich-observation environments may require exponentially many samples, highlighting the limits of tractability.

Our work situates itself in this online regime, explicitly addressing exploration rather than assuming exploratory data. However, the non-robust guarantees above do not transfer directly to our robust setting. Robust RL replaces a single nominal kernel with an uncertainty set and a worst-case Bellman operator, which breaks several conveniences used by non-robust analyses: (i) Bellman errors are non-linear and invalidates the usual variance-style error accounting: In non-robust RL, the kernel is fixed so the Bellman error can be captured through standard concentration inequalities; However, in robust case, the error propagation requires "functional transfer" between value functions and the dual variables to be quantified; (ii) Confidence sets and bonuses must control both sampling noise and adversarial model shift induced by the worst-case kernel: In non-robust RL, the confidence set only considers data limitations, whereas we additionally consider the uncertainties from the uncertainty set; (iii) Since the mismatch between the nominal and the worst-case kernels, our analysis requires additional structural notions (e.g., coverability) to capture such mismatches. We thus develop new concentration arguments that commute with the supremum over models, and new pessimism/optimism couplings to control duality gaps. In short, our robust online RL introduces adversarial model coupling and functional transfer effects that require genuinely different analysis and algorithmic design, which are not directly achievable from the non-robust studies.

C Proof of Theorem 1

Proof. We will now prove Theorem 1. To prove this, we first highlight the role of robust coverability, as defined in Definition 3, in limiting the complexity of exploration.

- **Equivalence between robust coverability and cumulative visitation.** A key idea underlying the proof of Theorem 1 is the equivalence between robust coverability and a quantity we term *cumulative visitation* under the worst-transition kernel P^ω as defined in Definition 2. We define the cumulative visitation as given below:

  Definition 4 (Cumulative Visitation)**.** We define the cumulative visitation at step h as

  $$C_h^{\mathrm{cv}} := \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sup_{\pi\in\Pi} d_h^{\pi,P^\omega}(s,a). \tag{18}$$

  The cumulative visitation C_h^{cv} reflects the variation in visitation probabilities under the worst-kernel for policies in the class Π. More specifically, it captures the total worst-case probability mass that policies in Π can allocate across the state-action space, under all admissible transition kernels. When this quantity is low, it indicates that policies in Π largely overlap in the regions they visit, limiting exploration complexity. Conversely, a high value implies that policies can spread mass across disjoint state-action pairs, making exploration harder. By Lemma T.3, we have

  $$C_{\mathrm{rcov}} \;=\; \max_{h\in[H]} C_h^{\mathrm{cv}}. \tag{19}$$

- **Relate Regret to Robust Average Bellman Error:** According to Assumption 1, we can guarantee $f^{(k)}$ is optimistic. Based on this optimistic algorithm, we will now relate the regret to the robust average Bellman error under the learner's sequence of policies.

  For any Markov kernel $Q = \{Q_h(\cdot \mid s,a)\}_{h=1}^H \in \mathcal{P}$ and by the definition of the occupancy measure of (s_h,a_h) as $d_h^{\pi^f,Q}$ induced by π_f and Q, we define the robust average Bellman error at level h by

  $$\varepsilon_{TV}^{\sigma}(f,\pi^f,h;Q) := \mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^f,Q}}\Big[f_h(s_h,a_h) - [\mathcal{T}_h^{\sigma} f_{h+1}](s_h,a_h)\Big]. \tag{20}$$

  By applying Lemma K.1 and by denoting $d^{\pi^{f^{(k)}},P^\omega} := d^{(k),P^\omega}$, we can relate regret to the robust average Bellman error as

$$\begin{aligned}
\mathrm{Regret}(K) &\le \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\Big[f_h^{(k)}(s_h,a_h) - [\mathcal{T}_h^{\sigma} f_{h+1}^{(k)}](s_h,a_h)\Big],\\
&= \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\Big[f_h^{(k)}(s_h,a_h) - \Big[\mathcal{T}^{\sigma}_{\widehat{g}_{f_{h+1}^{(k)}},h} f_{h+1}^{(k)}\Big](s_h,a_h)\\
&\qquad + \Big[\mathcal{T}^{\sigma}_{\widehat{g}_{f_{h+1}^{(k)}},h} f_{h+1}^{(k)}\Big](s_h,a_h) - [\mathcal{T}_h^{\sigma} f_{h+1}^{(k)}](s_h,a_h)\Big],\\
&= \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\Big[f_h^{(k)}(s_h,a_h) - \Big[\mathcal{T}^{\sigma}_{\widehat{g}_{f_{h+1}^{(k)}},h} f_{h+1}^{(k)}\Big](s_h,a_h)\Big]\\
&\qquad + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\Big[\Big[\mathcal{T}^{\sigma}_{\widehat{g}_{f_{h+1}^{(k)}},h} f_{h+1}^{(k)}\Big](s_h,a_h) - [\mathcal{T}_h^{\sigma} f_{h+1}^{(k)}](s_h,a_h)\Big]\\
&= \mathrm{I} + \mathrm{II},
\end{aligned} \tag{21}$$

where we denote

$$\mathrm{I} := \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\left[f_h^{(k)}(s_h,a_h) - \left[\mathcal{T}^{\sigma}_{\widehat{g}_{f_{h+1}^{(k)}},h} f_{h+1}^{(k)}\right](s_h,a_h)\right]. \tag{22}$$

$$\mathrm{II} := \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\left[\left[\mathcal{T}^{\sigma}_{\widehat{g}_{f_{h+1}^{(k)}},h} f_{h+1}^{(k)}\right](s_h,a_h) - [\mathcal{T}_h^{\sigma} f_{h+1}^{(k)}](s_h,a_h)\right]. \tag{23}$$

- **Bound of II via Robust Coverability:** To bound II, let us define $\Delta_{k,h}$ as

$$\Delta_{k,h}(s,a) := \left[\mathcal{T}^{\sigma}_{\widehat{g}_{f_{h+1}^{(k)}},h} f_{h+1}^{(k)}\right](s,a) - \left[\mathcal{T}_h^{\sigma} f_{h+1}^{(k)}\right](s,a).$$

Then, II can be written as

$$\mathrm{II} := \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\left[\Delta_{k,h}(s_h,a_h)\right]. \tag{24}$$

To bound the term II, we follow the following steps:

Step 1: Density ratio control. By Holder's inequality and using hte fact that $\mathbb{E}[X] \le \mathbb{E}[|X|]$, for any $\mu_h^\pi \in \Delta(\mathcal{S}\times\mathcal{A})$, we get

$$\mathbb{E}_{d_h^{(k),P^\omega}}[\Delta_{k,h}] \;\le\; \left\|\frac{d_h^{(k),P^\omega}}{\mu_h^\pi}\right\|_\infty \|\Delta_{k,h}\|_{1,\mu_h^\pi}, \tag{25}$$

where $\|\phi\|_{1,\mu^\pi} := \sum_{s,a}\mu^\pi(s,a)|\phi(s,a)|$. According to Definition 3, we have

$$\left\|\frac{d_h^{(k),P^\omega}}{\mu_h^\pi}\right\|_\infty \;\le\; C_{\mathrm{rcov}}. \tag{26}$$

Step 2: Apply Lemma K.3. By Lemma K.3, applied with μ_h^π and $f = f_{h+1}^{(k)}$ and by the choice of ξ_{dual} as ξ_{dual}/KH, and using a union bound over (k,h), we obtain

$$\|\Delta_{k,h}\|_{1,\mu_h^\pi} = \mathcal{O}\left(\frac{H}{\sigma}\sqrt{\frac{2\log\big(8|\mathcal{G}||\mathcal{F}|KH/\delta\big)}{|\mathcal{D}_h^{(k)}|}} + \frac{\xi_{\mathrm{dual}}}{KH}\right). \tag{27}$$

Step 3: Combine bounds. Hence, by combining eq. 25, eq. 26 and eq. 27, we get

$$\mathbb{E}_{d_h^{(k),P^\omega}}[\Delta_{k,h}] = \mathcal{O}\left(C_{\mathrm{rcov}}\frac{H}{\sigma}\sqrt{\frac{2\log\big(8|\mathcal{G}||\mathcal{F}|KH/\delta\big)}{|\mathcal{D}_h^{(k)}|}} + C_{\mathrm{rcov}}\frac{\xi_{\mathrm{dual}}}{KH}\right). \tag{28}$$

Step 4: Summing over $k,h \in [K]\times[H]$. Summing the bound in eq. 28 over $k\in[K]$ and $h\in[H]$ yields the desired result:

$$\mathrm{II} = \mathcal{O}\left(C_{\mathrm{rcov}}\frac{H}{\sigma}\sqrt{2\log\Big(\tfrac{8|\mathcal{G}||\mathcal{F}|KH}{\delta}\Big)}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{|\mathcal{D}_h^{(k)}|}} + C_{\mathrm{rcov}}\xi_{\mathrm{dual}}\right). \tag{29}$$

Step 5: Final Bound of II. By the update rule of RFL-TV, we have

$$\mathcal{D}_h^{(k)} \;\leftarrow\; \mathcal{D}_h^{(k-1)} \cup \{(s_h^{(k)}, a_h^{(k)}, s_{h+1}^{(k)})\} \qquad \forall\, h\in[H].$$

Therefore, in each episode k, exactly one sample appended to each step h in the dataset, hence $|\mathcal{D}_h^{(k)}| = |\mathcal{D}_h^{(0)}| + k = k$.

Since, $f(k) = k^{-1/2}$ is decreasing on $[1, \infty)$ and $f(1) = 1$, the term $\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{\sqrt{|\mathcal{D}_h^{(k)}|}}$ in eq. 29 can be bounded by the following intergral, as

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{\sqrt{|\mathcal{D}_h^{(k)}|}} = \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{\sqrt{k}} \leq H\left(1 + \int_1^K \frac{dx}{\sqrt{x}}\right) = 2H\sqrt{K} - H \leq 2H\sqrt{K}. \tag{30}$$

Applying eq. 30 in eq. 29, we get the final bound as

$$\mathrm{II} = \mathcal{O}\left(C_{\mathrm{rcov}} \frac{H^2}{\sigma} \sqrt{2K \log\left(\frac{8|\mathcal{G}||\mathcal{F}|KH}{\delta}\right)} \; + \; C_{\mathrm{rcov}} \xi_{\mathrm{dual}}\right). \tag{31}$$

- **Bound of I via Robust Coverability:** Before we bound I, we first define the robust Bellman error w.r.t. $\mathcal{T}_g^\sigma f$ as

$$\delta_h^{(k)}(\cdot, \cdot) := f_h^{(k)}(\cdot, \cdot) - \left[\mathcal{T}^\sigma_{\widehat{g}_{f_{h+1}^{(k)}}, h} f_{h+1}^{(k)}\right](\cdot, \cdot). \tag{32}$$

Then, I can be written as

$$\mathrm{I} := \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{(s_h, a_h) \sim d_h^{(k), P^\omega}} \left[\delta_h^{(k)}(s_h, a_h)\right]. \tag{33}$$

We denote the expected number of times of visiting (s, a) before episode k under the worst-transition kernel P^ω as $\widetilde{d}_h^{(k)} \equiv d_h^{\pi^{f^{(k)}}}$, and is defined as

$$\widetilde{d}_h^{(k)}(s, a) \; := \; \sum_{i=1}^{k-1} d_h^{(i), P^\omega}(s, a). \tag{34}$$

That is, $\widetilde{d}_h^{(k)}$ is the unnormalized average of all state visitations encountered prior to episode k, and μ_h^π is the visitation measure under nominal-kernel $P^\star$ for step h. Throughout the proof, we perform a slight abuse of notation and write

$$\mathbb{E}_{\widetilde{d}_h^{(k)}}[f] \; := \; \sum_{i=1}^{k-1} \mathbb{E}_{d_h^{(i), P^\omega}}[f] \quad \text{for any function } f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}.$$

Step 1: Robust optimism. Under the Assumption 1 and the construction of the confidence set $\mathcal{F}^{(k)}$, the following Lemma K.2, will guarantee that with probability at least $1 - \delta$, for all $k \in [K]$:

$$Q^{\star, \sigma} \in \mathcal{F}^{(k)} \quad \text{and} \quad \sum_{(s,a)} \widetilde{d}_h^{(k)}(s, a) \left(\delta_h^{(k)}(s, a)\right)^2 \; \leq \; \mathcal{O}(\beta). \tag{35}$$

Step 1: Conservative Burn-in Phase Construction. We introduce the notion of a "burn-in" phase for each state–action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ by defining

$$\tau_h(s, a) \; = \; \min\Big\{ t \;\Big|\; \widetilde{d}_h^{(t)}(s, a) \; \geq \; C_{\mathrm{rcov}} \cdot \mu_h^\pi(s, a) \Big\}, \tag{36}$$

which captures the earliest time at which (s, a) has been explored sufficiently; we refer to $k < \tau_h(s, a)$ as the burn-in phase for (s, a). In other words, $\tau_h(s, a)$ guarantees that no matter which kernel in the uncertainty set we are facing, the state–action pair (s, a) has received enough coverage.

Going forward, let $h \in [H]$ be fixed. We decompose regret into contributions from the burn-in phase for each state–action pair, and contributions from pairs which have been

explored sufficiently and reached a stable phase "stable phase":

$$I = \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d_h^{(k),P^\omega}}\Big[\delta_h^{(k)}(s,a)\,\mathbb{I}\{k<\tau_h(s,a)\}\Big]}_{\text{conservative burn-in phase}} \tag{37}$$

$$+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d_h^{(k),P^\omega}}\Big[\delta_h^{(k)}(s,a)\,\mathbb{I}\{k\geq\tau_h(s,a)\}\Big]}_{\text{stable phase}}\,. \tag{38}$$

We will not show that every state–action pair leaves the conservative burn-in phase. Instead, we use robust coverability to argue that the contribution from pairs that have not left this phase is small on average. In particular, we use that $|\delta_h^{(k)}| \leq [0, c_3 H/\sigma]$ to bound the factor, as follows

$$\mathbb{E}_{(s,a)\sim d_h^{(k),P^\omega}}\Big[\delta_h^{(k)}(s,a)\,\mathbb{I}\{k<\tau_h(s,a)\}\Big] \leq c_3\frac{H}{\sigma}\sum_{s,a} d_h^{(k),P^\omega}(s,a)\mathbb{I}\{k<\tau_h(s,a)\}. \tag{39}$$

Plugging eq. 39 in the conservative burn-in phase term of eq. 37, we get

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d_h^{(k),P^\omega}}\Big[\delta_h^{(k)}(s,a)\,\mathbb{I}\{k<\tau_h(s,a)\}\Big]\\
&\overset{(a)}{\leq} c_3\frac{H}{\sigma}\sum_{k=1}^{K}\sum_{h=1}^{H} d_h^{(k),P^\omega}(s,a)\mathbb{I}\{k<\tau_h(s,a)\}\\
&= c_3\frac{H}{\sigma}\sum_{h=1}^{H}\sum_{s,a}\sum_{k<\tau_h(s,a)} d_h^{(k),P^\omega}(s,a)\\
&\overset{(b)}{=} c_3\frac{H}{\sigma}\sum_{h=1}^{H}\sum_{s,a}\widetilde{d}_h^{\tau_h(s,a)}(s,a)\\
&= c_3\frac{H}{\sigma}\sum_{h=1}^{H}\sum_{s,a}\Big\{\widetilde{d}_h^{\tau_h(s,a)-1}(s,a) + d_h^{\tau_h(s,a)-1,P^\omega}(s,a)\Big\}\\
&\overset{(c)}{\leq} c_3\frac{H}{\sigma}\sum_{h=1}^{H}\sum_{s,a}\Big\{2C_{\mathrm{rcov}}\,\mu_h^\pi(s,a)\Big\}\\
&\overset{(d)}{=} c_3\frac{H^2}{\sigma}C_{\mathrm{rcov}}.
\end{aligned} \tag{40}$$

The ineq. (a) is due to the fact $\sup_P\sum_x g_x(P) \leq \sum_x \sup_P g_x(P)$; the equality (b) is by the definition of $\widetilde{d}_h^{\tau_h(s,a)}(s,a)$ by eq. 34; ineq. (c) is due to eq. 36 and by the fact $d_h^{\tau_h(s,a)-1,P^\omega}(s,a) \leq C_{\mathrm{rcov}}\,\mu_h^\pi(s,a)$.

For the stable phase, we apply change-of-measure as follows:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d_h^{(k),P^\omega}}\big[\delta_h^{(k)}(s,a)\,\mathbb{I}\{k\ge\tau_h(s,a)\}\big]\\
&=\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a} d_h^{(k),P^\omega}(s,a)\left(\frac{\widetilde{d}_h^{(k)}(s,a)}{\widetilde{d}_h^{(k)}(s,a)}\right)^{1/2}\delta_h^{(k)}(s,a)\,\mathbb{I}\{k\ge\tau_h(s,a)\}\\
&\le\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a} d_h^{(k),P^\omega}(s,a)\left(\frac{\widetilde{d}_h^{(k)}(s,a)}{\widetilde{d}_h^{(k)}(s,a)}\right)^{1/2}\delta_h^{(k)}(s,a)\,\mathbb{I}\{k\ge\tau_h(s,a)\}\\
&\le\sum_{h=1}^{H}\underbrace{\left(\sum_{k=1}^{K}\sum_{s,a}\frac{\big(\mathbb{I}\{t\ge\tau_h(x,a)\},d_h^{(k),P^\omega}(s,a)\big)^2}{\widetilde{d}_h^{(k)}(s,a)}\right)^{1/2}}_{\text{(A): extrapolation error}}\underbrace{\left(\sum_{k=1}^{K}\sum_{s,a}\tilde{d}_h^{(k)}(s,a)\left(\delta_h^{(k)}(s,a)\right)^2\right)^{1/2}}_{\text{(B): in-sample squared Bellman error}},
\end{aligned}\tag{41}$$

where the last inequality is Cauchy–Schwarz.

Using part (b) of Lemma K.2, we bound the in-sample error (B) by

$$\text{(B)}\ \le\ \mathcal{O}(\sqrt{\beta K}).\tag{42}$$

Bounding the extrapolation error using robust coverability. We control the extrapolation error (A) via robust coverability. We use the following scalar variant of the elliptic potential lemma of (Lattimore & Szepesvári, 2020) (proved in (Xie et al., 2022, Lemma 4)).

We bound (A) on a per-state basis and invoke robust coverability (and the equivalence to cumulative visitation) so that potentials from different (s,a) pairs aggregate well. From the definition of τ_h in eq. 36, for all $t\ge\tau_h(s,a)$ we have $\widetilde{d}_h^{(k)}(s,a)\ge C_{\mathrm{rcov}}\mu_h^\pi(s,a)$, which implies $\widetilde{d}_h^{(k)}(s,a)\ge\frac{1}{2}\Big(\widetilde{d}_h^{(k)}(s,a)+C_{\mathrm{rcov}}\mu_h^\pi(s,a)\Big)$. Thus,

$$\begin{aligned}
(A)&=\sqrt{\sum_{k=1}^{K}\sum_{s,a}\frac{\big(\mathbb{I}\{k\ge\tau_h(s,a)\}d_h^{(k),P^\omega}(s,a)\big)^2}{\widetilde{d}_h^{(k)}(s,a)}}\\
&\le\sqrt{2\sum_{k=1}^{K}\sum_{s,a}\frac{d_h^{(k),P^\omega}(s,a)\cdot d_h^{(k),P^\omega}(s,a)}{\widetilde{d}_h^{(k)}(s,a)+C_{\mathrm{rcov}}\cdot\mu_h^\pi(s,a)}}\\
&\le\sqrt{2\sum_{k=1}^{K}\sum_{s,a}\max_{\ell\in[K]}d_h^{(l),P^\omega}(s,a)\frac{d_h^{(k),P^\omega}(s,a)}{\widetilde{d}_h^{(k)}(s,a)+C_{\mathrm{rcov}}\cdot\mu_h^\pi(s,a)}}\\
&\le\sqrt{2\left(\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\frac{d_h^{(k),P^\omega}(s,a)}{\widetilde{d}_h^{(k)}(s,a)+C_{\mathrm{rcov}}\cdot\mu_h^\star(s,a)}\right)\left(\sum_{s,a}\max_{l\in[K]}d_h^{(l),P^\omega}(s,a)\right)}\\
&\le\ \mathcal{O}\big(\sqrt{C_{\mathrm{rcov}}\log K}\big),
\end{aligned}\tag{43}$$

where the last line uses Lemma T.5 and Lemma T.3.

To conclude, substitute eq. 42 and eq. 43 into eq. 41 to obtain

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d_h^{(k),P^\omega}}\big[\delta_h^{(k)}(s,a)\,\mathbb{I}\{k\ge\tau_h(s,a)\}\big]\ \le\ \mathcal{O}\Big(H\sqrt{C_{\mathrm{rcov}}\cdot\beta K\log K}\Big).\tag{44}$$

By applying eq. 40 and eq. 44 in eq. 37, we get

$$I\le\mathcal{O}\left(\frac{H^2}{\sigma}C_{\mathrm{rcov}}+H\sqrt{C_{\mathrm{rcov}}\cdot\beta K\log K}\right).\tag{45}$$

Therefore, by applying eq. 45 and eq. 31 in eq. 21, we get

$$\mathrm{Regret}(K) \leq \mathcal{O}\Bigg(\frac{H^2}{\sigma}C_{\mathrm{rcov}} + H\sqrt{C_{\mathrm{rcov}}\cdot\beta K\log K} + C_{\mathrm{rcov}}\,\frac{H^2}{\sigma}\sqrt{2K\log\Big(\tfrac{8|\mathcal{G}||\mathcal{F}|KH}{\delta}\Big)} \;+\; C_{\mathrm{rcov}}\xi_{\mathrm{dual}}\Bigg).$$

This concludes the proof of Theorem 1. □

C.1 Key Lemmas

Lemma K.1 (Robust Value function error decomposition)**.** *Consider an RMDP using the TV-divergence uncertainty set as defined in eq. 1 where we define* $V^f := \mathbb{E}[f_1(s_1, \pi_1^f(s_1))]$ *and* $V^{\pi^f,Q} := \mathbb{E}_{a_{1:H}\sim\pi^f, s_{h+1}\sim Q_h}\Big[\sum_{h=1}^H r_h(s_h,a_h)\Big]$*. Then, under Assumption 1 and Definition 2, we define the robust average Bellman error* $\varepsilon^\sigma_{TV}(f,\pi^f,h;P^\omega)$ *as given in eq. 20. Then, we can bound the regret as given in eq. 6 as,*

$$\mathrm{Regret}(K) \leq \sum_{k=1}^K\sum_{h=1}^H \varepsilon^\sigma_{TV}(f^{(k)}, \pi^{f^{(k)}}, h; P^\omega). \tag{46}$$

Proof. Fix any kernel $Q \in \mathcal{P}$. Let us denote $\psi^f(s') := \max_{a'\in\mathcal{A}} f(s',a')$. By definition of $\mathcal{T}_h^\sigma f$ in eq. 8, we get

$$[\mathcal{T}_h^\sigma f_{h+1}](s,a) = r_h(s,a) + \inf_{P\in\mathcal{U}_h^\sigma(s,a)} \mathbb{E}_P\Big[\psi^f_{h+1}\Big] \;\leq\; r_h(s,a) + \mathbb{E}_{s'\sim Q_h(\cdot|s,a)}[\psi^f_{h+1}(s')]. \tag{47}$$

Thus, from eq. 47 we get

$$f_h(s,a) - [\mathcal{T}_h^\sigma f_{h+1}](s,a) \;\geq\; f_h(s,a) - r_h(s,a) - \mathbb{E}_{s'\sim Q_h}[\,\psi^f_{h+1}(s')\,]. \tag{48}$$

Taking expectation under $d_h^{\pi^f,Q}$ and summing over h gives

$$\sum_{h=1}^H \varepsilon^\sigma_{\mathrm{TV}}(f,\pi^f,h;Q) \;\geq\; \sum_{h=1}^H \mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^f,Q}}\Big[f_h(s_h,a_h) - r_h(s_h,a_h) - \mathbb{E}_{Q_h}[\psi^f_{h+1}]\Big]. \tag{49}$$

The right-hand side of eq. 49 follows the same proof-lines as in (Jiang et al., 2017, Lemma 1), yielding

$$\sum_{h=1}^H \varepsilon^\sigma_{\mathrm{TV}}(f,\pi^f,h;Q) \;\geq\; V^f - V^{\pi^f,Q}. \tag{50}$$

Finally, if Q is a worst–case kernel for π^f, i.e., $Q \equiv P^\omega$ then for each (s,a,h),

$$\mathbb{E}_{s'\sim P_h^\omega(\cdot|s,a)}[\psi^f_{h+1}(s')] := \mathbb{E}_{s'\sim Q_h(\cdot|s,a)}[\psi^f_{h+1}(s')] = \inf_{P\in\mathcal{U}_h(s,a)} \mathbb{E}_P[\psi^f_{h+1}(s')],$$

so the inequality becomes equality. In this case,

$$\sum_{h=1}^H \varepsilon^\sigma_{\mathrm{TV}}(f,\pi^f,h;Q) = V^f - V^{\pi^f,P^\omega}.$$

Now, under the worst-transition kernel P^ω, we have $V_1^{\pi^{(k)},\sigma}(s_1) = V_1^{\pi^{(k)},P^\omega}(s_1)$. Furthermore, according to Assumption 1, we can guarantee that $f^{(k)}$ is optimistic in episode k. Using these fact, we can say that $V_h^{\star,\sigma}(s) \leq V_h^{f^{(k)}}(s)$. Therefore, we can write

$$\begin{aligned}
\mathrm{Regret}(K) &= \sum_{k=1}^K V_1^{\star,\sigma}(s_1) - V_1^{\pi^{(k)},\sigma}(s_1) \\
&\leq \sum_{k=1}^K V_1^{f^{(k)}}(s_1) - V_1^{\pi^{(k)},P^\omega}(s_1) \\
&\leq \sum_{k=1}^K\sum_{h=1}^H \varepsilon^\sigma_{\mathrm{TV}}(f^{(k)},\pi^{f^{(k)}},h;P^\omega) \qquad \text{[By eq. 50]}.
\end{aligned}$$

This concludes the proof of Lemma K.1. □

Lemma K.2. *Suppose Assumption 1 holds. Then if $\beta > 0$ is selected as in Theorem 1, then with probability at least $1-\delta$, for all $k \in [K]$, RFL-TV satisfies*

(a) $Q^{\star,\sigma} \in \mathcal{F}^{(k)}$.

(b) $\sum_{(s,a)} \widetilde{d}_h^{(k)}(s,a)\left(\delta_h^{(t)}(s,a)\right)^2 \le \mathcal{O}(\beta)$.

Proof. The proof follows the same structure as the non-robust argument (Jin et al., 2021, Lemma 39 and 40) and (Xie et al., 2022, Lemma 15) (martingale concentration via Freedman's inequality plus a finite cover of the functional class), with two robust-specific ingredients: (i) the dual scalar representation of the TV worst-case expectation and (ii) the use of the dual pointwise integrand as a sample target. We derive the complete proof as follows.

☞ *Proof of ineq. (b)* To show ineq. (b), we will focus on the proof-lines of (Jin et al., 2021, Lemma 39) and (Xie et al., 2022, Lemma 15 (2)). We first fix (k,h,f) tuple, where an episode k we consider a function $f^{(k)} = \{f_1^{(k)}, \dots, f^{(k)}{}_H\} \in \mathcal{F}$. Let us denote $\psi^k(s) := \psi^f_{f^k_{h+1}}(s)$ such that $\psi^k(s_{h+1}) := f^{(k)}_{h+1}(s_{h+1}, \pi^{(k)}_{h+1}(s_{h+1}))$, and we assume $\|f\|_\infty, \|\psi^f\|_\infty \le H$ (this is the boundedness assumption used throughout). We consider the filtration induced as

$$\mathcal{H}_h^{(k)} = \{s_1^i, a_1^i, r_1^i, \dots, s_H^i\}_{i=1}^{k-1} \bigcup \{s_1^k, a_1^k, r_1^k, \dots, s_h^k, a_h^k\}$$

as the filtration containing the history up to the episode k at step h.

We obtain $\widehat{g}_{f_h} \in [0, 2H/\sigma]$ as a measurable minimizer of eq. 10 that satisfies Assumption 3. For the trajectory of episode k, we define

$$Z_h^{(k)}(f, \widehat{g}_f) := \left(\widehat{g}_{f^{(k)}_{h+1}}(s_h^k, a_h^k) - \psi^{f^{(k)}}_{h+1}(s_h^k, a_h^k)\right)_+ - (1-\sigma)\widehat{g}_{f^{(k)}_{h+1}}(s_h^k, a_h^k), \tag{51}$$

such that $\left|Z_h^{(k)}(f, \widehat{g}_f)\right| \le 5H/\sigma$ and

$$\mathbb{E}\left[Z_h^{(k)}(\widehat{g}_f, f)\middle|\mathcal{H}_h^{(k)}\right] = \left[\mathcal{T}^\sigma_{\widehat{g}_{f^{(k)}_{h+1}}, h} f^{(k)}_{h+1}\right](s_h^k, a_h^k) - r_h^{(k)}(s_h^k, a_h^k). \tag{52}$$

For each episode k and step h, we define the martingale difference as

$$\begin{aligned} X_h^{(k)}(f, \widehat{g}_f) := &\left(f_h^{(k)}(s_h^k, a_h^k) - r_h^{(k)}(s_h^k, a_h^k) - Z_h^{(k)}(f^{(k)}, \widehat{g}_{f^{(k)}})\right)^2 \\ &- \left(\left[\mathcal{T}^\sigma_{\widehat{g}_{f^{(k)}_{h+1}}, h} f^{(k)}_{h+1}\right](s_h^k, a_h^k) - r_h^{(k)}(s_h^k, a_h^k) + Z_h^{(k)}(f^{(k)}, \widehat{g}_{f^{(k)}})\right)^2, \end{aligned} \tag{53}$$

such that we have $\left|X_h^{(k)}(f, \widehat{g}_f)\right| \le c_1\left(H \min\{H, 1/\sigma\}\right)^2$, where $c_1 > 0$ is an absolute constant. Moreover,

$$\begin{aligned} \mathbb{E}\left[X_h^{(k)}(f, \widehat{g}_f)\middle|\mathcal{H}_h^{(k)}\right] &= \left(\delta_h^{(k)}(s_h^k, a_h^k)\right)^2 \\ \mathrm{Var}\left[X_h^{(k)}(f, \widehat{g}_f)\middle|\mathcal{H}_h^{(k)}\right] &\le c_2\left(H \min\{H, 1/\sigma\}\right)^2 \mathbb{E}\left[X_h^{(k)}(f, \widehat{g}_f)\middle|\mathcal{H}_h^{(k)}\right], \end{aligned} \tag{54}$$

where $c_1, c_2 > 0$ are absolute constants.

Therefore, by Freedman's inequality as given Lemma T.4, we can write

$$\left|\sum_{k=1}^K \left(X_h^{(k)}(f, \widehat{g}_f) - \mathbb{E}\left[X_h^{(k)}(f, \widehat{g}_f)\right]\right)\middle|\mathcal{H}_h^{(k)}\right| \le \mathcal{O}\left(\sqrt{\log(1/\delta)\sum_{k=1}^K \mathbb{E}\left[X_h^{(k)}(f, \widehat{g}_f)\middle|\mathcal{H}_h^{(k)}\right]} + \log(1/\delta)\right). \tag{55}$$

Now, let us consider $\mathcal{X}_\rho$ be the ρ-cover of $\mathcal{F} \bigcup \mathcal{G}$. Now taking a union bound for all $(k, h, \phi) \in [K] \times [H] \times \mathcal{X}_\rho$, and following the same proof-lines as in (Jin et al., 2021, Lemma 39), we get

$$\sum_{t<k} \mathbb{E}\Big[\Big(\delta_h^{(t)}(s_h, a_h)\Big)^2 \Big| \mathcal{H}_h^{(t)}\Big] \leq \mathcal{O}(\beta), \tag{56}$$

where $\beta = \mathcal{O}\Big(\Big(H \min\{H, 1/\sigma\}\Big) \log\Big(\frac{KH\,|\mathcal{F}||\mathcal{G}|}{\delta}\Big)\Big)$.

Therefore, eq. 56 concludes that $\sum_{t<k} \mathbb{E}_{(s,a)\sim d_h^{(t),P^\omega}(s,a)} \left[\delta_h^{(t)}(s,a)\right)^2 \leq \mathcal{O}(\beta)$.

By the definition of visitation measures, we have

$$\begin{aligned}
\sum_{(s,a)} \widetilde{d}_h^{(k)}(s,a)\, \delta_h^{(t)}(s,a)^2 &\overset{(a)}{=} \sum_{t<k} \sum_{(s,a)} d_h^{(t),P^\omega}(s,a)\, \delta_h^{(t)}(s,a)^2 \\
&= \sum_{t<k} \mathbb{E}_{(s,a)\sim d_h^{(t),P^\omega}} \left[\delta_h^{(t)}(s,a)^2\right] \\
&\overset{(b)}{\leq} \mathcal{O}(\beta),
\end{aligned} \tag{57}$$

where (a) is by the definition of $\widetilde{d}_h^{(k)}(s,a)$ given by equation 34, and (b) is using equation 56.

☞ *Proof of ineq. (a)* To show ineq. (a), we will focus on the proof-lines of (Jin et al., 2021, Lemma 40) and (Xie et al., 2022, Lemma 15 (1)). Fix (k, h, f) and follow the same notation as mentioned in the proof lines of the inequality (b), we define

$$\begin{aligned}
W_h^{(t)}(f, \widehat{g}_f) := &\left(f_h^{(t)}(s_h^t, a_h^t) - r_h^{(t)}(s_h^t, a_h^t) - Z_h^{(t)}(f^{(t)}, \widehat{g}_{f^{(t)}})\right)^2 \\
&- \left(Q_h^{\star,\sigma}(s_h^t, a_h^t) - r_h^{(t)}(s_h^t, a_h^t) + Z_h^{(t)}(f^{(t)}, \widehat{g}_{f^{(t)}})\right)^2, \quad \text{for } 1 \leq t \leq k.
\end{aligned}$$

As in eq. 54, $\mathbb{E}\Big[W_h^{(t)}(f, \widehat{g}_f) \mid \mathcal{H}_h^{(t)}\Big] = \Big(f_h^{(t)}(s_h^t, a_h^t) - Q_h^{\star,\sigma}(s_h^t, a_h^t)\Big)^2$ where $\mathcal{H}_h^{(t)}$ be the filtration induced by $\{s_1^i, a_1^i, r_1^i, \ldots, s_H^i\}_{i=1}^{t-1} \bigcup \{s_1^t, a_1^t, r_1^t, \ldots, s_h^t, a_h^t\}$. Similarly, we can verify that $|W_h^{(t)}(f, \widehat{g}_f)| \leq c_1\Big(H \min\{H, 1/\sigma\}\Big)^2$ and $\mathrm{Var}\Big[W_h^{(t)}(f, \widehat{g}_f) \mid \mathcal{H}_h^{(t)}\Big] \leq c_2\Big(H \min\{H, 1/\sigma\}\Big)^2 E\Big[W_h^{(t)}(f, \widehat{g}_f) \mid \mathcal{H}_h^{(t)}\Big]$. Now, following the proof-lines of (Jin et al., 2021, Lemma 40), and applying Freedman's ineq. (Lemma T.4 and a cover of $\mathcal{G}$ yields, w.p. $1-\delta$, we get

$$\begin{aligned}
&\sum_{t=1}^{k-1} \Big[Q_h^{\star,\sigma}(s_h^t, a_h^t) - r_h^t(s_h^t, a_h^t) - Q_{h+1}^{\star,\sigma}(s_{h+1}^t, \pi_{h+1}^{Q^{\star,\sigma}}(s_{h+1}^t))\Big]^2 \\
&\leq \sum_{t=1}^{k-1} \Big[f_h^{(t)}(s_h^t, a_h^t) - r_h^t(s_h^t, a_h^t) - Q_{h+1}^{\star,\sigma}(s_{h+1}^t, \pi_{h+1}^{Q^{\star,\sigma}}(s_{h+1}^t))\Big]^2 + \mathcal{O}(\beta).
\end{aligned}$$

Finally, by recalling the definition of $\mathcal{F}^{(k)}$, we conclude that with probability at least $1-\delta$, $Q^{\star,\sigma} \in \mathcal{F}^{(k)}$ for all $k \in [K]$.

This concludes the proof of Lemma K.2. □

Lemma K.3 (Dual Optimization Error Bound)**.** *Let $\widehat{g}_f$ be the dual optimization parameter obtained from eq. 10 for the state-action value function f and let $\mathcal{T}_g^\sigma$ be as defined in eq. 8. Then, under Definition 2, with probability at least $1-\delta$, we have*

$$\sup_{f\in\mathcal{F}} \|\mathcal{T}^\sigma f - \mathcal{T}_{\widehat{g}_f}^\sigma f\|_{1,\mu^\pi} = \mathcal{O}\Big(H \min\{H, 1/\sigma\} \sqrt{\frac{2\log(8|\mathcal{G}||\mathcal{F}|/\delta)}{|\mathcal{D}|}} + \xi_{\text{dual}}\Big). \tag{58}$$

Proof. Fix an $f \in \mathcal{F}$. We will also invoke union bound for the supremum here. We recall from (8) that $\widehat{g}_f = \arg\min_{g\in\mathcal{G}} \widehat{\text{Dual}}_{loss}(g; f)$. From the robust Bellman equation, we directly obtain

$$\begin{aligned}
\|\mathcal{T}^\sigma f - \mathcal{T}^\sigma_{\widehat{g}_f} f\|_{1,\mu^\pi} &= \mathbb{E}_{(s,a)\sim\mu^\pi}\Bigg[\mathbb{E}_{s'\sim P^\star_{s,a}}\left[\left(\widehat{g}_f(s,a) - \max_{a'} f(s',a')\right)_+ - (1-\sigma)\widehat{g}_f(s,a)\right] \\
&\qquad - \inf_{\eta\in[0,2H/\sigma]}\left(\mathbb{E}_{s'\sim P^\star_{s,a}}\left[\left(\eta - \max_{a'} f(s',a')\right)_+ - (1-\sigma)\eta\right]\right)\Bigg]. \\
&\overset{(a)}{\leq} \mathbb{E}_{(s,a)\sim\mu^\pi}\mathbb{E}_{s'\sim P^\star_{s,a}}\left[\left(\widehat{g}_f(s,a) - \max_{a'} f(s',a')\right)_+ - (1-\sigma)\widehat{g}_f(s,a)\right] \\
&\qquad - \mathbb{E}_{(s,a)\sim\mu^\pi}\left[\inf_{\eta\in[0,2H/\sigma]}\left(\mathbb{E}_{s'\sim P^\star_{s,a}}\left[\left(\eta - \max_{a'} f(s',a')\right)_+ - (1-\sigma)\eta\right]\right)\right].
\end{aligned}$$

Moreover,

$$\begin{aligned}
\|\mathcal{T}^\sigma f - \mathcal{T}^\sigma_{\widehat{g}_f} f\|_{1,\mu^\pi} &\overset{(b)}{\leq} \mathbb{E}_{(s,a)\sim\mu^\pi, s'\sim P^\star_{s,a}}\left[\left(\widehat{g}_f(s,a) - \max_{a'} f(s',a')\right)_+ - (1-\sigma)\widehat{g}_f(s,a)\right] \\
&\qquad - \inf_{g\in\mathcal{L}^1}\mathbb{E}_{(s,a)\sim\mu^\pi, s'\sim P^\star_{s,a}}\left[\left(g(s,a) - \max_{a'} f(s',a')\right)_+ - (1-\sigma)g(s,a)\right]\Bigg)\Bigg]. \\
&= \Bigg(\mathbb{E}_{(s,a)\sim\mu^\pi, s'\sim P^\star_{s,a}}\left[\left(\widehat{g}_f(s,a) - \max_{a'} f(s',a')\right)_+ - (1-\sigma)\widehat{g}_f(s,a)\right] \\
&\qquad - \inf_{g\in\mathcal{G}}\mathbb{E}_{(s,a)\sim\mu^\pi, s'\sim P^\star_{s,a}}\left[\left(g(s,a) - \max_{a'} f(s',a')\right)_+ - (1-\sigma)g(s,a)\right]\Bigg) \\
&\qquad + \Bigg(\inf_{g\in\mathcal{G}}\mathbb{E}_{(s,a)\sim\mu^\pi, s'\sim P^\star_{s,a}}\left[\left(g(s,a) - \max_{a'} f(s',a')\right)_+ - (1-\sigma)g(s,a)\right] \\
&\qquad\qquad - \inf_{g\in\mathcal{L}^1}\mathbb{E}_{(s,a)\sim\mu^\pi, s'\sim P^\star_{s,a}}\left[\left(g(s,a) - \max_{a'} f(s',a')\right)_+ - (1-\sigma)g(s,a)\right]\Bigg). \\
&\overset{(c}{\leq} \Bigg(\mathbb{E}_{(s,a)\sim\mu^\pi, s'\sim P^\star_{s,a}}\left[\left(\widehat{g}_f(s,a) - \max_{a'} f(s',a')\right)_+ - (1-\sigma)\widehat{g}_f(s,a)\right] \\
&\qquad - \inf_{g\in\mathcal{G}}\mathbb{E}_{(s,a)\sim\mu^\pi, s'\sim P^\star_{s,a}}\left[\left(g(s,a) - \max_{a'} f(s',a')\right)_+ - (1-\sigma)g(s,a)\right]\Bigg) + \xi_{\text{dual}}. \\
&\overset{(d)}{\leq} \frac{4H(2-\sigma)}{\sigma}\sqrt{\frac{2\log(|\mathcal{G}|)}{|\mathcal{D}|}} + \frac{25H}{\sigma}\sqrt{\frac{2\log(8/\delta)}{|\mathcal{D}|}} + \xi_{\text{dual}}.
\end{aligned}$$

The inequality (a) follows since $\inf_g h(g) \leq h(\hat{g}_f)$, wher we denote $h(g) := \mathbb{E}_{s'\sim P^\star_{s,a}}\big((g - \max_{a'} f(s',a'))_+ - (1-\sigma)g\big)$; (b) follows from Lemma T.2; (c) follows from the approximate dual realizability assumption (Assumption 3).

For (d), we consider the loss function $l(g,(s,a,s')) = (g(s,a) - \max_{a'} f(s',a'))_+ - (1-\sigma)g(s,a)$ and dataset $\mathcal{D}$. Note that $|l(g,(s,a,s'))| \leq 5H/\sigma$ (since $f \in \mathcal{F}$ and $g \in \mathcal{G}$). Now, we can apply the empirical risk minimization result (11) in Lemma 3 to get (d), where $R(\cdot)$ is the Rademacher complexity.

Finally, (e) follows from eq. 60 in Lemma T.1 when combined with the facts that $l(g,(s,a,s'))$ is $(2-\sigma)$-Lipschitz in g and $g(s,a) \leq 2H/\sigma$, since $g \in \mathcal{G}$.

With union bound, with probability at least $1-\delta$, we finally get

$$\sup_{f\in\mathcal{F}} \|\mathcal{T}^\sigma f - \mathcal{T}^\sigma_{\widehat{g}_f} f\|_{1,\mu^\pi} \leq 25(3-\sigma)\frac{H}{\sigma}\sqrt{\frac{2\log(8|\mathcal{G}||\mathcal{F}|/\delta)}{|\mathcal{D}|}} + \xi_{\text{dual}}$$
$$\leq C\frac{H}{\sigma}\sqrt{\frac{2\log(8|\mathcal{G}||\mathcal{F}|/\delta)}{|\mathcal{D}|}} + \xi_{\text{dual}}$$

which concludes the proof. □

C.2 Technical Lemmas

We now state a result for the generalization bounds on empirical risk minimization (ERM) problems. This result is adapted from (Shalev-Shwartz & Ben-David, 2014, Theorem 26.5, Lemma 26.8, Lemma 26.9).

Lemma T.1 (ERM generalization bound)**.** *Let P be the data generating distribution on the space $\mathcal{X}$ and let $\mathcal{H}$ be a given hypothesis class of functions. Assume that for all $x\in\mathcal{X}$ and $h\in\mathcal{H}$ we have that $|l(h,x)|\leq c_1$ for some positive constant $c_1>0$. Given a dataset $\mathcal{D}=\{X_i\}_{i=1}^N$, generated independently from P, denote $\hat{h}$ as the ERM solution, i.e.*

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^N l(h,X_i).$$

For any fixed $\delta\in(0,1)$ and $h^*\in\arg\min_{h\in\mathcal{H}}\mathbb{E}_{X\sim P}[l(h,X)]$, we have

$$\mathbb{E}_{X\sim P}[l(\hat{h},X)] - \mathbb{E}_{X\sim P}[l(h^*,X)] \leq 2R(l\circ\mathcal{H}\circ\mathcal{D}) + 5c_1\sqrt{\frac{2\log(8/\delta)}{N}}, \tag{59}$$

with probability at least $1-\delta$, where $R(\cdot)$ is the Rademacher complexity of $l\circ\mathcal{H}$ given by

$$R(l\circ\mathcal{H}\circ\mathcal{D}) = \frac{1}{N}\mathbb{E}_{\{\sigma_i\}_{i=1}^N}\left(\sup_{g\in l\circ\mathcal{H}}\sum_{i=1}^N \sigma_i g(X_i)\right),$$

in which σ_i's are independent from X_i's and are independently and identically distributed according to the Rademacher random variable σ, i.e. $\mathbb{P}(\sigma=1)=0.5=\mathbb{P}(\sigma=-1)$.

Furthermore, if $\mathcal{H}$ is a finite hypothesis class, i.e. $|\mathcal{H}|<\infty$, with $|h\circ x|\leq c_2$ for all $h\in\mathcal{H}$ and $x\in\mathcal{X}$, and $l(h,x)$ is c_3-Lipschitz in h, then we have

$$\mathbb{E}_{X\sim P}[l(\hat{h},X)] - \mathbb{E}_{X\sim P}[l(h^*,X)] \leq 2c_2c_3\sqrt{\frac{2\log(|\mathcal{H}|)}{N}} + 5c_1\sqrt{\frac{2\log(8/\delta)}{N}}, \tag{60}$$

with probability at least $1-\delta$.

We now mention two important concepts from variational analysis (Rockafellar & Wets, 1998) literature that is useful to relate minimization of integrals and the integrals of pointwise minimization under special class of functions.

Definition 5 (Decomposable spaces and Normal integrands (Rockafellar & Wets, 1998)(Definition 14.59, Example 14.29))**.** A space $\mathcal{X}$ of measurable functions is a decomposable space relative to an underlying measure space $(\Omega,\mathcal{A},\mu)$, if for every function $x_0\in\mathcal{X}$, every set $A\in\mathcal{A}$ with $\mu(A)<\infty$, and any bounded measurable function $x_1:A\to\mathbb{R}$, the function

$$x(\omega) = x_0(\omega)\mathbf{1}(\omega\notin A) + x_1(\omega)\mathbf{1}(\omega\in A)$$

belongs to $\mathcal{X}$. A function $f:\Omega\times\mathbb{R}\to\mathbb{R}$ (finite-valued) is a normal integrand, if and only if $f(\omega,x)$ is $\mathcal{A}$-measurable in ω for each x and is continuous in x for each ω.

Remark 2. *A few examples of decomposable spaces are $\mathcal{L}^p(\mathcal{S}\times\mathcal{A},\Sigma(\mathcal{S}\times\mathcal{A}),\mu)$ for any $p\geq 1$ and $\mathcal{M}(\mathcal{S}\times\mathcal{A},\Sigma(\mathcal{S}\times\mathcal{A}))$, the space of all $\Sigma(\mathcal{S}\times\mathcal{A})$-measurable functions.*

Lemma T.2 ((Rockafellar & Wets, 1998), Theorem 14.60)**.** *Let $\mathcal{X}$ be a space of measurable functions from Ω to $\mathbb{R}$ that is decomposable relative to a σ-finite measure μ on the σ-algebra $\mathcal{A}$. Let $f : \Omega \times \mathbb{R} \to \mathbb{R}$ (finite-valued) be a normal integrand. Then, we have*

$$\inf_{x\in\mathcal{X}} \int_{\omega\in\Omega} f(\omega, x(\omega))\mu(d\omega) = \int_{\omega\in\Omega} \left(\inf_{x\in\mathcal{X}} f(\omega, x)\right) \mu(d\omega).$$

Moreover, as long as the above infimum is not $-\infty$, we have that

$$x' \in \arg\min_{x\in\mathcal{X}} \int_{\omega\in\Omega} f(\omega, x(\omega))\mu(d\omega),$$

if and only if $x'(\omega) \in \arg\min_{x\in\mathbb{R}} f(\omega, x)\mu$ almost surely.

Lemma T.3 (Equivalence of robust coverability and cumulative visitation (Xie et al., 2022), Lemma 3)**.** *Recall the definition of C_{rcov} as given in Definition 3 and the* cumulative visitation *for every layer $h \in [H]$ as given in Definition 4. Then*

$$C_{\mathrm{rcov}} \;=\; \max_{h\in[H]} C_h^{cv}.$$

Lemma T.4 (Freedman's inequality (e.g., (Agarwal et al., 2014)))**.** *Let $\{M_t\}_{t\le T}$ be a real-valued martingale difference sequence w.r.t. filtration $\{\mathcal{G}_t\}$ with $|M_t| \le b$ a.s. and let $S_T = \sum_{t=1}^{T} \mathbb{E}[M_t^2 \mid \mathcal{G}_{t-1}]$. Then for any $\delta \in (0, 1)$,*

$$\Pr\Big(\sum_{t=1}^{T} M_t \ge \sqrt{2S_T \ln(1/\delta)} + \tfrac{b}{3}\ln(1/\delta)\Big) \le \delta.$$

Lemma T.5 (Per-state-action elliptic potential lemma (Lattimore & Szepesvári, 2020))**.** *Let $d^{(1)}, d^{(2)}, \ldots, d^{(K)}$ be an arbitrary sequence of distributions over a set $\mathcal{Z}$ (e.g., $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$), and let $\mu \in \Delta(\mathcal{Z})$ be a distribution such that $d^{(t)}(z)/\mu(z) \le C$ for all $(z, t) \in \mathcal{Z} \times [K]$. Then for all $z \in \mathcal{Z}$,*

$$\sum_{k=1}^{K} \frac{d^{(k)}(z)}{\sum_{i<t} d^{(k)}(z) + C \cdot \mu(z)} \;\le\; \mathcal{O}(\log K)\,.$$