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 or offline dataset, 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) , and more recently hybrid regimes that combine a large offline dataset with limited online interaction (Panaganti et al., 2024). 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 framework with general function approximation and by deriving finite-sample convergence guarantees. Our main contributions are summarized as follows.

(1) **First sample-efficient algorithm for online DR-RL with general function approximation.** We develop *Robust Fitted Learning with TV-Divergence Uncertainty Set (RFL-TV)*, the first algorithm for *purely online* DR-RL with general function approximation under TV-divergence uncertainty sets. *RFL-TV* integrates optimism for exploration into a fitted-learning scheme via a novel functional reformulation of the robust Bellman operator. Instead of standard state–action-wise bonuses as in tabular UCB methods, we use this reformulation to construct a *global* uncertainty quantifier over the function class, which aggregates estimation error more effectively and guides exploration. This yields a computationally efficient algorithm suitable for large-scale problems and, to the best of our knowledge, the first polynomial-time, polynomial-sample algorithm for purely online DR-RL beyond tabular and offline/hybrid settings.

(2) **Robust coverability as the fundamental complexity measure.** We introduce the *robust coverability coefficient* C_{rcov}, defined as the worst-case ratio between adversarial and nominal visitation measures across policies and time steps. Although a similar term is also studied in (Panaganti et al., 2024) for hybrid setting, our studies reveals its necessity in pure online setting with function approximation, which captures the intrinsic "information deficit" of learning a worst-case policy from nominal data. We show that (i) natural fail-state assumptions imply $C_{\mathrm{rcov}} < \infty$, and (ii)

C_{rcov} fully characterizes the sample complexity of online DR-RL, in direct analogy to—but strictly weaker than—classical coverability and concentrability conditions.

(3) **Dual robust fitted learning and global confidence sets.** We construct a dual robust Bellman residual based on a functional optimizer g and use it to build *global confidence sets* over value functions. Unlike tabular UCB methods and non-robust GOLF, which use per–state–action bonuses, *RFL-TV* maintains a single least-squares objective on the dual residual that simultaneously (i) approximates the worst-case Bellman operator and (ii) serves as a global uncertainty quantifier for exploration. This dual-driven robust fitted learning mechanism is specific to the DR-RL setting and contrasts with offline DR-RL methods, where the dual is analyzed under a fixed data distribution and not used to drive exploration.

(4) **Sharp regret and sample-complexity guarantees (general and linear settings).** We show that RFL-TV finds an ε-optimal robust policy with sample complexity $\tilde{\mathcal{O}}\Big(H^5\big(\min\{H,\sigma^{-1}\}\big)^2 C_{\text{rcov}}\varepsilon^{-2}\Big)$, up to logarithmic factors, plus an additive term linear in the dual approximation error. This bound is the first polynomial-order guarantee for robust online learning with general function approximation. It is independent of $|\mathcal{S}|$ and $|\mathcal{A}|$, demonstrating scalability to large or continuous spaces, with the cost of higher state-space dependence on H and σ^{-1} compared to existing online DR-RL results. Moreover, in d-dimensional *linear* TV-RMDPs our analysis specializes to a regret bound $\tilde{\mathcal{O}}\big(H^2\min\{H,\sigma^{-1}\}\sqrt{C_{\text{rcov}}^2 d^2 K}\big)$, which is near-optimal compared to the minimax lower bound (Liu et al., 2024), highlighting the sharpness of our theory.

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 a comprehensive offline dataset (Blanchet et al., 2023; Shi & Chi, 2024; Zhang et al., 2023; Liu & Xu, 2024a; Wang et al., 2024c;a). Recently, limited number of 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 with function approximation largely focuses on linear function classes. However, these classes are generally not closed under the robust Bellman operator, so approximation guarantees cannot be ensured. To circumvent this, most works impose strong structural assumptions on the underlying robust MDP—such as a small discount factor (Xu & Mannor, 2010; Tamar et al., 2014; Zhou et al., 2024) or a linear robust MDP model (Ma et al., 2022; Liu & Xu, 2024b;a; Liu et al., 2024; Wang et al., 2024a)—assumptions that are difficult to verify in practice. In contrast, we work with a broader, general function class to avoid these restrictions. General function approximation for DR-RL has so far been studied mainly in (Panaganti et al., 2022; 2024), which use a functional optimization approach but focus on offline or hybrid data settings with global coverage and thus avoid the exploration challenges of our fully online setting; moreover, (Panaganti et al., 2024) studies regularized robust MDPs, which differ from the DR-RL formulation considered here.

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 $\mathcal{U}_h^\sigma(s, a) \triangleq \big\{P \in \Delta(\mathcal{S}) : D(P, P_h^\star(\cdot|s, a)) \leq \sigma\big\}$, 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:

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

where for $p, q \in \Delta(\mathcal{S})$, $D_{\mathrm{TV}}\big(p, q\big) = \frac{1}{2} \sum_{s' \in \mathcal{S}} |p(s') - q(s')|$ 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 $P^\star$

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

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 is to minimize the *cumulative robust regret* over K episodes, defined as

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

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

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

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

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

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

When $C_{\mathrm{rcov}} = \infty$, 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), an online learning algorithm is efficient only if the coverability measure $C_{\mathrm{rcov}} < \infty$, 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. In this sense, our robust coverability condition plays a role analogous to the offline/hybrid coverage notions in (Panaganti et al., 2022; 2024), but is tailored to a different regime: their coverage constants compare robust occupancies to a fixed offline behavior distribution μ, whereas C_{rcov} compares robust occupancies to the nominal online occupancies induced by the learner's policy, specifically for online settings.

Empirical robust Bellman operator and functional optimization. Recall from eq. 5 that the robust value function is characterized as the fixed point of the robust Bellman operator. Hence, computing an optimal robust policy amounts to computing this fixed point. Directly evaluating the operator, however, is intractable: the mapping $\mathbb{E}_{\mathcal{U}_h^\sigma(s,a)}[\cdot]$ requires, for each (s,a), an optimization over an S–dimensional TV-uncertainty set, which quickly becomes prohibitive.

To address these issues, we construct an efficient empirical solution and adapt the approach in (Panaganti et al., 2022) to avoid pointwise scalar optimization by rewriting the problem as a *single* optimization over functions. Namely, we consider the probability space $(\mathcal{S} \times \mathcal{A}, \Sigma(\mathcal{S} \times \mathcal{A}), \mu)$, let $\mathcal{L}^1(\mu)$ denote the space of absolutely integrable dual functions, and consider the dual loss

$$\mathrm{Dual}_{loss}(g;f) = \mathbb{E}_{(s,a)\sim\mu}\big[\mathbb{E}_{s'\sim P^\star_{s,a}}[(g(s,a) - \max_{a'} f(s',a'))_+] - (1-\sigma)g(s,a)\big], \tag{8}$$

and its optimization is equivalent to the robust Bellman operator (Panaganti et al., 2022):

$$\inf_{g \in \mathcal{L}^1(\mu)} \mathrm{Dual}_{loss}(g;f) = \mathbb{E}_{(s,a)\sim\mu}\left[\mathbb{E}_{\mathcal{U}_h^\sigma(s,a)}[f]\right]. \tag{9}$$

We now construct an empirical dual loss $\widehat{\mathrm{Dual}}_{loss}(g;f)$ through a dataset, and obtain an approximate dual minimizer by solving $\inf_{g\in\mathcal{L}^1(\mu)} \widehat{\mathrm{Dual}}_{loss}(g;f)$. For efficiency, we instead optimize over another function class $\mathcal{G} = \{g : \mathcal{S} \times \mathcal{A} \to [0, 2H/\sigma]\}$ used to approximate the dual variables, which satisifies the following realizability assumption (deferred to Appendix C). We then approximate the robust Bellman operator for a given f and dataset $\mathcal{D}$ as $\underline{g}_f = \arg\min_{g\in\mathcal{G}} \widehat{\mathrm{Dual}}_{loss}(g;f)$, and then define the empirical robust Bellman operator

$$(\mathcal{T}_g^\sigma f)(s,a) \triangleq r(s,a) - \mathbb{E}_{s'\sim P^\star_{s,a}}[(g(s,a) - \max_{a'} f(s',a'))_+]) - (1-\sigma)g(s,a). \tag{10}$$

The next lemma quantifies how well $\mathcal{T}_{\underline{g}_f}^\sigma$ approximates $\mathcal{T}^\sigma$ in an $\mathcal{L}^1$ sense.

Lemma 1. *Let π be any policy, and let μ_h^π denote the visitation measure on $\mathcal{S} \times \mathcal{A}$ at step h induced by π under $P^\star$. Suppose $\mathcal{D}$ is a dataset collected by running π. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}} \left\| \mathcal{T}^\sigma f - \mathcal{T}_{\underline{g}_f}^\sigma f \right\|_{1,\mu^\pi} = \mathcal{O}\Big(H \min\{H, 1/\sigma\} \sqrt{2\log\big(8|\mathcal{G}||\mathcal{F}|/\delta\big) \big/ |\mathcal{D}|} + \xi_{\mathrm{dual}}\Big). \tag{11}$$

A similar result is derived for a fixed distribution (the offline dataset distribution) in (Panaganti et al., 2022; 2024), whereas we show it simultaneously hold for any policy and its induced distribution. Lemma 1 shows that our empirical functional optimization yields a uniformly accurate approximation to the robust Bellman operator under the $\mathcal{L}^1(\mu^\pi)$ norm. Crucially, the error is controlled *globally* with respect to the visitation measure μ^π, rather than pointwise in (s, a). This global control is what we leverage later to define our robust confidence sets and the global error term that drives the design and analysis of our main algorithm.

Remark 1 (Relation to φ-regularized RMDPs (Panaganti et al., 2024))**.** *Assumption 3 and Lemma 1 build on the dual functional machinery first developed by (Panaganti et al., 2022) and subsequently employed by (Panaganti et al., 2024) for φ-regularized RMDPs in a hybrid setting, where the policy value includes a Lagrangian penalty λ with $\lambda > 0$ and the guarantees scale with $(\lambda + H)$. Although the φ-regularized RMDPs recovers the standard RMDPs with $\lambda = 0$, our result cannot be obtained directly. This is due to that, the analysis in (Panaganti et al., 2024) is carried out explicitly for $\lambda > 0$ and we cannot set $\lambda = 0$ in their analysis to obtain ours.*

5 Robust Fitting Learning Algorithm

We then utilize our previous constructions and propose our Robust Fitted Learning (RFL) algorithm.

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

1: **Input:** Function class $\mathcal{F}$, Dual Function class $\mathcal{G}$, $\beta > 0$, $\sigma > 0$.
2: **Initialize:** $\mathcal{F}^{(0)} \leftarrow \mathcal{F}$, $\mathcal{D}_h^{(0)} \leftarrow \emptyset\ \forall h \in [H]$
3: **for** episode $k = 1, 2, \ldots, K$ **do**
4: Set $f^{(k)} \leftarrow \arg\max_{f \in \mathcal{F}^{(k-1)}} f(s_1, \pi_1^f(s_1))$ and $\pi^{(k)} \leftarrow \pi^{f^{(k)}}$
5: Execute $\pi^{(k)}$ and obtain a trajectory $(s_1^{(k)}, a_1^{(k)}, r_1^{(k)}), \ldots, (s_H^{(k)}, a_H^{(k)}, r_H^{(k)})$
6: Update dataset: $\mathcal{D}_h^{(k)} \leftarrow \mathcal{D}_h^{(k-1)} \cup \{(s_h^{(k)}, a_h^{(k)}, s_{h+1}^{(k)})\}\ \forall h \in [H]$
7: $\mathcal{F}_H^{(k)} \leftarrow \{0\}$
8: **for** $h = H - 1, ..., 1$ **do**
9: Update the confidence set, with notations defined in eq. 12:

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

10: **end for**
11: **end for**
12: **Output:** $\bar{\pi} = \mathrm{unif}(\pi^{(1:K)})$. `For PAC guarantee only.`

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 1 to construct the set. Namely, given a function f, we first solve the dual-variable approximation through the empirical functional optimization loss as

$$\underline{g}_f \triangleq \arg\min_{g \in \mathcal{G}} \sum_{(s,a,s') \in \mathcal{D}_h^{(k)}} \left(g(s, a) - \max_{a' \in \mathcal{A}} f(s', a') \right)_+ - (1 - \sigma) g(s, a). \tag{12}$$

We further capture the empirical robust Bellman error $L_h^{(k)}(f', f, g)$ via our functional optimization:

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

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.

Algorithmic novelties. While the overall template of "optimism + fitted value iteration" is reminiscent of GOLF (Xie et al., 2022), our setting differs fundamentally from both non-robust RL and offline DR-RL, and this is reflected in the design of RFL-TV. The datasets $\mathcal{D}_h^{(k)}$ are generated under *different* policies across episodes, so there is no single policy π with $\mathcal{D}_h^{(k)} \sim \mu^\pi$, and the fixed-distribution error quantification in Lemma 1 cannot be applied directly. Instead, we implement an optimistic, dual-driven fitted scheme in which a value–dual pair (f, g) is learned online: the dual network g both approximates the TV-robust Bellman operator and acts as a *global uncertainty quantifier* that defines optimistic confidence sets over $\mathcal{F}$ and guides exploration under this non-stationary data. The resulting regret decomposition and confidence bounds exploit the robust coverability coefficient C_{rcov} (Definition 3) to control the mismatch between the evolving on-dynamics distribution and the worst-case kernel P^ω. This dual-based, coverability-aware realization of "optimism + fitted value iteration" contrasts with GOLF's squared non-robust Bellman error (Xie et al., 2022) and with offline RFQI-style robust methods (Panaganti et al., 2022), where data come from a fixed distribution and the dual is not used to drive online exploration.

6 Theoretical Guarantees

We then develop the theoretical guarantees of our algorithm.

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

$$\text{Regret}(K) \leq \mathcal{O}\Big(\sqrt{C_{\text{rcov}}^2 H^4 (\min\{H, 1/\sigma\})^2 \log\big(KH|\mathcal{F}||\mathcal{G}|\delta^{-1}\big)K} + C_{\text{rcov}}\xi_{\text{dual}}\Big).$$

Proof-sketch. Our proof has two main ingredients: (i) a reduction of robust regret to a sum of *robust average Bellman errors* under a worst-case kernel, and (ii) uniform control of these Bellman errors via the dual-based empirical operator and robust coverability.

Step 1: From regret to robust average Bellman error. By Assumption 1 and the construction of the confidence sets, the optimistic estimates $f^{(k)}$ satisfy $f_h^{(k)} \geq Q_h^{\star,\sigma}$ pointwise for all k, h with high probability. Using the robust Bellman equation and the worst-case kernel P^ω in Definition 2, Lemma K.1 shows that the robust regret can be written as a sum of robust average Bellman errors:

$$\text{Regret}(K) \leq \sum_{k=1}^{K}\sum_{h=1}^{H} \varepsilon_{\text{TV}}^{\sigma}\big(f^{(k)}, \pi_f^{(k)}, h; P^\omega\big),$$

where $\varepsilon_{\text{TV}}^{\sigma}$ is some robust Bellman error and is defined in eq. 17. Notably, different from offline robust RL (Panaganti et al., 2022) and non-robust ones, our robust Bellman error is defined under the worst-case occupancy measure, which depends on the learner's greedy policy $\pi_f^{(k)}$ and the corresponding worst-case kernel P^ω. Both of them are changing during algorithms, and we need to tackle such distribution shift among episodes.

Step 2: Dual-based decomposition and bounds via coverability. Our strategy is to utilize the coverability to derive a uniform upper bound of the errors. For each (k, h), we add and subtract the

[2] 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).

dual-based empirical operator $\mathcal{T}_g^\sigma$ (eq. 10) and decompose (eq. 19 and eq. 20): $\mathrm{Regret}(K) \le I + II$, where I aggregates the Bellman residuals $f_h^{(k)} - \mathcal{T}_{g_{f^{(k)}}}^\sigma f_{h+1}^{(k)}$, and II aggregates the approximation error $\mathcal{T}_{g_{f^{(k)}}}^\sigma f_{h+1}^{(k)} - \mathcal{T}^\sigma f_{h+1}^{(k)}$. Since $f^{(k)}$ and $g^{(k)}$ minimize a global least-squares loss, Lemma K.2 shows that the empirical squared Bellman residuals contributing to I are bounded by $\mathcal{O}(\beta)$ on the observed data. A concentration type analysis then implies

$$I \le \mathcal{O}\big(HC_{\mathrm{rcov}} \min\{H, \sigma^{-1}\} + H\sqrt{C_{\mathrm{rcov}} \cdot \beta K \log K}\big).$$

For II, Lemma 1 provides a uniform bound on $\mathcal{T}^\sigma f - \mathcal{T}_{g_f}^\sigma f$ for all $f \in \mathcal{F}$ and policies π. Robust coverability (Definition 3) is then used to transfer this control from the nominal visitation μ_h^π to the worst-case visitation under P^ω, resulting in

$$\mathrm{II} = \mathcal{O}\Big(C_{\mathrm{rcov}} H^2 \min\{H, \sigma^{-1}\}\sqrt{2K\log\Big(8|\mathcal{G}||\mathcal{F}|KH/\delta\Big)} + C_{\mathrm{rcov}}\xi_{\mathrm{dual}}\Big).$$

Combining the bounds on I and II and setting β as in Theorem 1 hence implies the regret bound.

Our result is the first polynomial regret of robust online learning with general function approximation. It is free from the problem scales and hence enjoys better scalability. Our result is also comparable against the ones under offline/hybrid setting with general function approximation in (Panaganti et al., 2022; 2024), hence our algorithm can efficient learn RMDPs even without any pre-collected dataset.

Remark 2. *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.*

Technical novelties in the analysis. Our analysis departs substantially from both non-robust online RL (Xie et al., 2022) and offline/hybrid robust RL with function approximation (Panaganti et al., 2022; 2024). First, we learn a dual-based robust Bellman operator from *on-dynamics* data, but must certify performance under the *worst-case* kernel P^ω, so the dual optimization error is measured under nominal visitation while regret is defined under robust occupancy. To bridge this mismatch, we introduce the robust coverability coefficient C_{rcov}, which uniformly bounds density ratios between worst-case and nominal occupancies across episodes and time steps, and use it to propagate a *single* global dual error through the regret analysis. Second, unlike analyses that work with a fixed Bellman operator and only control approximation error in the value class $\mathcal{F}$, our confidence bounds must simultaneously handle errors in both $\mathcal{F}$ and the dual class $\mathcal{G}$ in the backup $\mathcal{T}_{g_f}^\sigma$, requiring a new dual optimization error lemma and a careful treatment under evolving on-policy distributions. Third, our regret decomposition explicitly ties these dual-based Bellman residuals to cumulative robust visitation, cleanly separating *algorithmic* quantities (exploration scale β, dual error ξ_{dual}) from the *structural* property C_{rcov} of the underlying RMDP.

By contrast, the offline robust RL analysis of Panaganti et al. (2022) assumes a *static* dataset drawn from a distribution satisfying a strong *global concentratability* condition that uniformly covers all policies and kernels in the uncertainty set, allowing all functional errors to be controlled under a single reference measure. In our fully online setting no such dataset exists: the data distributions μ^{π_k} evolve with learning, and the mismatch between nominal and worst-case kernels invalidates any global coverage assumption and forces an episode- and time-dependent analysis. Our robust coverability coefficient C_{rcov} is therefore a weaker, more local notion than the global concentratability used in the offline setting, yet our results show that it already suffices to obtain sample-efficient exploration guarantees in online distributionally robust RL.

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*

$$T = KH = \mathcal{O}\Big(H^5(\min\{H, \sigma^{-1}\})^2 C_{\mathrm{rcov}}^2 \log\Big(T\,|\mathcal{F}||\mathcal{G}|\delta^{-1}\Big)\varepsilon^{-2} + C_{\mathrm{rcov}}\xi_{\mathrm{dual}}\varepsilon^{-1}\Big)$$

Our bound is independent of S and A, indicating scalability to large state and action spaces. Moreover, as we shall discuss later, the dependences on other parameters, H, σ, ε, are also tight and near-optimal.

Remark 3. *We note that our results are obtained for a fixed uncertainty set level σ. However, when σ is varying, the required function class $|\mathcal{F}_\sigma|$ can also change and depend on σ. Nevertheless, note that $\sup_\sigma |\mathcal{F}_\sigma|$ is upper bounded by the tabular function class, which is independent from σ.*

We note that the minimax lower bound for general function approximation based RL is generally unattainable, due to the richness of function classes, even for non-robust setting. Thus to justify the tightness of our results, we compare our results with prior works, especially under two reductions: tabular and linear cases. A comprehensive comparison can be found in Table 1.

Remark 4 (Tabular)**.** *Our result can be reduced to the finite tabular case by taking $\mathcal{F}$ and $\mathcal{G}$ to be the full spaces of bounded functions $\mathcal{S} \times \mathcal{A} \to [0, H]$ and $\mathcal{S} \times \mathcal{A} \to [0, 2H/\sigma]$, so that the entropy terms satisfy $\log |\mathcal{F}|, \log |\mathcal{G}| = \tilde{\mathcal{O}}(SA)$ (Jin et al., 2021). Substituting these quantities into Theorem 1 yields a tabular regret bound of order $\tilde{\mathcal{O}}\big(\sqrt{C_{\mathrm{rcov}}^2 H^4 (\min\{H, 1/\sigma\})^2\, SA\, K}\big)$. Comparing with results with a similar coverage notion (He et al., 2025), which has a regret of $\tilde{\mathcal{O}}\big(\sqrt{C_{\mathrm{vr}} H^4 S^3 A\, K}\big)$, our algorithm has a better dependence on S and a worse dependence on H, which indicating the scalability of our method. Moreover, our algorithm has more applicability with general function approximation.*

Remark 5 (Linear TV-RMDPs)**.** *As another special case, we specialized our results to the d-rectangular linear RMDPs (Ma et al., 2022; Liu et al., 2024) in Appendix D. In particular, when the robust Bellman operator is linear in a d-dimensional feature map, we adapted the analysis of Theorem 1 and show that RFL-TV attains $\mathrm{Regret}(K) = \tilde{\mathcal{O}}\Big(\sqrt{C_{\mathrm{rcov}}^2\, H^4 (\min\{H, 1/\sigma\})^2\, d^2 K}\Big)$. Moreover, we show in Lemma K.4 that $C_{\mathrm{rcov}} \leq \mathcal{O}(d)$, hence the sample complexity is $\tilde{\mathcal{O}}\big(d^4 H^5 (\min\{H, 1/\sigma\})^2 \varepsilon^{-2}\big)$. Such a result is $\mathcal{O}(d^2 H^2)$-worse than the online learning in linear robust MDPs in (Liu et al., 2024), and $\mathcal{O}(d^2 H^3)$-worse than the minimax lower bound (Liu et al., 2024). However, our results hold for more general non-linear function classes.*

Setting	Online / Hybrid	Robustness	Sample complexity
General	Online, (Xie et al., 2022)	No	$\tilde{\mathcal{O}}\big(C_{\mathrm{cov}} H^3 \log(\lvert\mathcal{F}\rvert/\delta)\varepsilon^{-2}\big)$
	Online, **RFL-TV (ours), Thm. 1**	Yes	$\tilde{\mathcal{O}}\big(C_{\mathrm{rcov}}^2 H^5 (\min\{H, 1/\sigma\})^2 \log(\lvert\mathcal{F}\rvert\lvert\mathcal{G}\rvert/\delta)\varepsilon^{-2}\big)$
	Lower Bound	N/A	N/A
Tabular	Online, (Azar et al., 2017)	No	$\tilde{\mathcal{O}}\big(SAH^4\varepsilon^{-2}\big)$
	Online, (Lu et al., 2024)	Yes	$\tilde{\mathcal{O}}\big(\min\{H, \sigma^{-1}\} SAH^3 \varepsilon^{-2}\big)$
	Online, (He et al., 2025)	Yes	$\tilde{\mathcal{O}}\big(C_{\mathrm{vr}} S^3 AH^5 \varepsilon^{-2}\big)$
	Online, **RFL-TV (ours), Thm. 1**	Yes	$\tilde{\mathcal{O}}\big(C_{\mathrm{rcov}}^2 H^5 (\min\{H, 1/\sigma\})^2 SA\varepsilon^{-2}\big)$
	Lower Bound (Lu et al., 2024)	Yes	$\tilde{\Omega}\big(H^3 \min\{H, 1/\sigma\} SA\varepsilon^{-2}\big)$
Linear	Online, (He et al., 2023)	No	$\tilde{\mathcal{O}}\big(d^2 H^4 \varepsilon^{-2}\big)$
	Online, (Liu et al., 2024)	Yes	$\tilde{\mathcal{O}}\big(d^2 H^3 (\min\{H, 1/\sigma\})^2 \varepsilon^{-2}\big)$
	Hybrid, (Panaganti et al., 2024)	Yes	$\tilde{\mathcal{O}}\big(\max\{C^2(\pi^\star), 1\}\, d^3 H^3 (\lambda + H)^2 \varepsilon^{-2}\big)$
	Online, **RFL-TV (ours), Thm. 2**	Yes	$\tilde{\mathcal{O}}\big(C_{\mathrm{rcov}}^2 H^5 (\min\{H, 1/\sigma\})^2 d^2 \varepsilon^{-2}\big)$
	Lower Bound (Liu et al., 2024)	Yes	$\tilde{\Omega}\big(d^2 H^2 (\min\{H, 1/\sigma\})^2 \varepsilon^{-2}\big)$

Table 1: Comparison under general-function, tabular, and linear settings.

7 Conclusion

In this work, we introduced RFL-TV, a DR-RL algorithm with general function approximation under TV-uncertainty set for a purely online setting. 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_{rcov} 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. Reducing to both tabular and d-rectangular linear RMDP cases, our results are both tight and near-optimal against existing works and minimax lower bounds, implying the tightness and near-optimality of our results. Our hence algorithm stands for the first purely online, sample efficient algorithm for large scale DR-RL.

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.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 263–272. PMLR, July 2017. ISSN: 2640-3498.

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.

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 Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

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

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly Minimax Optimal Reinforcement Learning for Linear Markov Decision Processes. In *International Conference on Machine Learning*, pp. 12790–12822. PMLR, 2023.

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, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably Efficient Reinforcement Learning with Linear Function Approximation. In *Conference on learning theory*, pp. 2137–2143. PMLR, 2020.

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.

Kishan Panaganti, Adam Wierman, and Eric Mazumdar. Model-Free Robust ϕ-Divergence Reinforcement Learning Using Both Offline and Online Data. *arXiv preprint arXiv:2405.05468*, 2024.

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.

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, 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.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. *arXiv preprint arXiv:2012.08507*, 2020.

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 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.

B NUMERICAL EXPERIMENTS

B.1 EXPERIMENTAL SETUP ON CARTPOLE

Environment. We consider the standard `CartPole-v1` benchmark with a discrete action space. The state $s \in \mathbb{R}^4$ contains the cart position, cart velocity, pole angle, and pole angular velocity, and the action space is $\mathcal{A} = \{0, 1\}$, corresponding to applying a fixed horizontal force to the left or right. Episodes terminate either when the pole falls beyond the allowed angle or when the time limit is reached (maximum horizon $H = 500$). Rewards are the standard per-step rewards from the environment, and the agent aims to maximize the undiscounted return over each episode.

Robustness evaluation. We are interested in how the learned policies behave under several kinds of mismatch between training and test conditions. Policies are always *trained on the nominal* *environment* and are evaluated under the following perturbation families, applied only at evaluation time:

- **Action perturbation.** At each time step, with probability $\rho \in [0, 1]$ the environment ignores the agent's action and instead executes a uniformly random action in $\mathcal{A}$. We evaluate over a grid of perturbation levels and, for the plots in the main text, we focus on

$$\Gamma_{\text{act}} = \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\},$$

  where $\rho = 0$ corresponds to the nominal case (used internally for sanity checks but not always displayed in the figures).

- **Force-magnitude perturbation.** The horizontal push force applied in the dynamics is multiplied by a scalar factor η_{force}. We evaluate the learned policies on a finite set of scale values $\eta_{\text{force}} \in \Gamma_{\text{force}}$ that includes values

$$\Gamma_{\text{force}} = \{0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0\},$$

  where smaller values correspond to progressively weaker control inputs, and $\eta_{\text{force}} = 1.0$ is the nominal strength (used for training but not repeated in this sweep).

- **Pole-length perturbation.** The physical pole length is multiplied by a scalar factor η_{len}. At the configuration level, we specify the effective evaluation grid as

$$\Gamma_{\text{len}} = \{0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0\},$$

  covering shorter and longer poles relative to the nominal length.

In all settings, training is performed on the nominal environment ($\rho = 0, \eta_{\text{force}} = 1, \eta_{\text{len}} = 1$), while robustness is measured by evaluating the final policy on perturbed environments from the above families. Unless otherwise stated, each reported return corresponds to the average over 20 evaluation episodes and 3 independent random seeds $\{0, 1, 2\}$; we also plot 95% confidence intervals computed across seeds and episodes.

Practical RFL-TV agent. For CartPole we use a purely value-based implementation of RFL-TV with a discrete action space. The agent maintains two Q-networks Q_1, Q_2 (for Double Q-learning) and their target copies $\bar{Q}_1, \bar{Q}_2$, together with a dual network g and its target copy $\bar{g}$. All networks are multilayer perceptrons with ReLU activations:

- **Q-networks.** We maintain two Q-networks Q_1 and Q_2. Each network takes the state $s \in \mathbb{R}^4$ as input and outputs a vector in $\mathbb{R}^{|\mathcal{A}|}$, one value per discrete action ($|\mathcal{A}| = 2$ for CartPole). The architecture is a two-layer fully connected MLP with hidden sizes $(128, 128)$ and ReLU activations, followed by a linear output layer. The scalar value $Q_i(s, a)$ is obtained by indexing the corresponding component of this output vector.
- **Dual network.** The dual function $g(s, a)$ is parameterized by a network with the same backbone as the Q-networks: it takes $s \in \mathbb{R}^4$ as input, passes it through two fully connected ReLU layers with $(128, 128)$ units, and produces a vector in $\mathbb{R}^{|\mathcal{A}|}$, one value per action. The output is passed through a sigmoid and scaled so that $g(s, a) \in [0, 10]$ for all (s, a), enforcing non-negativity and preventing numerical blow-up in the dual updates.

Training protocol and robustness hyper-parameters. RFL-TV is trained off-policy on `CartPole-v1` using a replay buffer and an ε-greedy exploration strategy. Unless otherwise specified, we fix the discount factor to $\gamma = 0.99$ and use soft target updates with rate $\tau = 0.005$ for all target networks. Transitions are stored in a replay buffer of size 2×10^5, from which we sample mini-batches of size 256 and perform one gradient update per environment step. The Q-networks and dual network are optimized with Adam at a learning rate of 3×10^{-4}. Exploration uses ε-greedy action selection, where the exploration rate is initialized at $\varepsilon_{\text{start}} = 1.0$ and decayed linearly to $\varepsilon_{\text{end}} = 0.05$ over the first 200 episodes, and then held fixed at 0.05 for the remainder of training. Each configuration is trained for $K = 500$ episodes, and we report performance statistics over three random seeds $\{0, 1, 2\}$. The robust RFL-TV backup is parameterized by a TV-radius $\sigma \in [0, 1]$ and a slack parameter $\beta \geq 0$ that controls how strictly the dual constraint is enforced. On CartPole, we sweep $\sigma \in \{0.0, 0.2, 0.3, 0.4, 0.5, 0.6\}$ and treat the slack parameter β as a scalar hyperparameter

Table 2: Training hyper-parameters for RFL-TV on `CartPole-v1`.

Parameter	Symbol	Value
Discount factor	γ	0.99
Target update rate	τ	0.005
Replay buffer size	$\lvert\mathcal{D}\rvert$	2×10^5 transitions
Mini-batch size	B	256
Q-network learning rate	lr_Q	3×10^{-4}
Dual-network learning rate	lr_g	3×10^{-4}
Exploration start	$\varepsilon_{\text{start}}$	1.0
Exploration end	ε_{end}	0.05
Epsilon decay horizon	T_ε	200 episodes
Gradient updates per step	–	1
Training episodes	K	500
Evaluation episodes per configuration	–	20
Random seeds	–	$\{0, 1, 2\}$
TV-robustness radii	σ	$\{0.0, 0.2, 0.3, 0.4, 0.5, 0.6\}$
Slack parameter	β	$\{0.0, 0.5, 1.0\}$

Table 3: Network architectures for RFL-TV on CartPole.

Network	Hidden layers
Q-network Q_1, Q_2	$(128, 128)$ (ReLU)
Dual network g (default)	$(128, 128)$ (ReLU)
Dual network g (capacity sweep)	$(64, 64)$/ $(128, 128)$/ $(256, 256)$ (ReLU)

controlling how strictly we enforce the dual Bellman constraint. After normalizing rewards and values so that the dual residual has typical scale $\mathcal{O}(1)$, we sweep $\beta \in \{0.0, 0.5, 1.0\}$, spanning hard ($\beta = 0$) to moderately relaxed ($\beta = 1$) constraints, and report the best-performing setting. In our experiments, the best choice is $\beta = 0.0$ under action perturbations and $\beta = 0.5$ under force-magnitude and pole-length perturbations. The numerical values of all optimization hyper-parameters and network architectures are summarized in Tables 2 and 3.

Practical RFL-TV update (CartPole, discrete). For completeness, Algorithm 2 summarizes the training loop for the discrete practical RFL-TV agent used in the CartPole experiments. The pseudocode follows our implementation: we use Double Q-learning with a dual network that approximates the robust inner optimization under total variation, and we incorporate the slack parameter β by clipping the dual residual inside a quadratic penalty.

B.2 RFL-TV VS. FUNCTIONAL APPROXIMATION BENCHMARKS: GAINS UNDER SHIFT

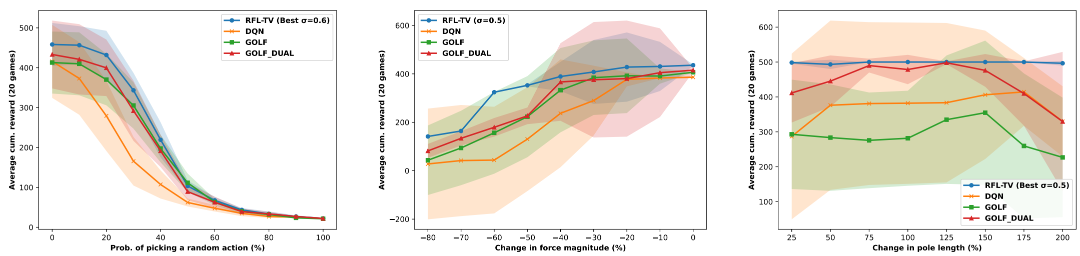
Figure 1 compares RFL-TV to three function-approximation baselines: DQN, the value-function method GOLF (Xie et al., 2022), and a dual-augmented variant GOLF-DUAL, which shares the same dual architecture as RFL-TV but is run with $\sigma = 0$. All three baselines are trained without explicit distributional robustness and thus correspond to the non-robust ($\sigma = 0$) setting. For RFL-TV, we fix the uncertainty radius to the value that achieves the best nominal CartPole performance on our σ grid, selecting $\sigma = 0.6$ for action perturbations, $\sigma = 0.5$ for force-magnitude perturbations, and $\sigma = 0.5$ for pole-length perturbations.

Across all three perturbation families, RFL-TV (with its best-performing $\sigma > 0$) consistently dominates the non-robust functional approximation baselines. Under *action perturbations*, for moderate noise levels $\rho \in [0.2, 0.5]$, RFL-TV achieves roughly 30–60% higher average return than DQN and about 15–30% higher than the best non-robust value-based baseline, with performance at $\rho \approx 0.3$ nearly twice that of DQN. For *force-magnitude shifts* of 40–80% from nominal, RFL-TV maintains average returns of roughly 150–400, while DQN stays below about 260 and GOLF/GOLF-DUAL lie mostly in the 60–380 range, corresponding to roughly ≈ 1.5–$3\times$ higher

Algorithm 2: Practical RFL-TV for CartPole

```
Inputs: TV radius σ, slack β, discount γ, target rate τ, batch size B, episodes K, horizon H,
    exploration schedule (ε_start, ε_end, K_decay).
Initialize replay buffer D ← ∅.
Initialize Q-networks Q_1, Q_2 and dual network g; set target networks Q̄_i ← Q_i for i = 1, 2 (and
    optionally ḡ ← g).
for k = 1, ..., K do
    Set ε_k by linearly decaying from ε_start to ε_end over K_decay episodes, then clamping.
    Reset environment and observe s_0.
    for t = 0, ..., H − 1 do
        With prob. ε_k sample a_t uniformly; otherwise
            a_t = arg max_a min{Q_1(s_t, a), Q_2(s_t, a)}.
        Execute a_t, observe (r_t, s_{t+1}, d_t), and store (s_t, a_t, r_t, s_{t+1}, d_t) in D.
        if |D| ≥ B then
            Sample minibatch {(s, a, r, s', d)}_{j=1}^B from D.
            *** Target value (Double Q) ***
            Compute Q̄_i(s', ·), i = 1, 2, and update
                v_next(s') = max_{a'} min{Q̄_1(s', a'), Q̄_2(s', a')}.
            *** Dual update with slack β ***
            Evaluate g(s, a) and define
                dual_term(s, a) = (g(s, a) − v_next(s'))_+ − (1 − σ) g(s, a).
            Compute residual
                r_dual(s, a) = |dual_term(s, a)| − β,   r̃_dual(s, a) = max{r_dual(s, a), 0},
            and minimize
                L_g = E[r̃_dual(s, a)^2]
            w.r.t. the parameters of g (one gradient step).
            *** Q-update using updated g ***
            Recompute
                dual_term^new(s, a) = (g(s, a) − v_next(s'))_+ − (1 − σ) g(s, a),
            and form targets
                y = r + (1 − d) γ (v_next(s') + dual_term^new(s, a)).
            Compute Q_i(s, a), i = 1, 2, and minimize
                L_Q = E[(Q_1(s, a) − y)^2 + (Q_2(s, a) − y)^2]
            w.r.t. the parameters of Q_1, Q_2 (one gradient step).
            Soft-update targets: Q̄_i ← (1 − τ) Q̄_i + τ Q_i,   i = 1, 2,
            and optionally: ḡ ← (1 − τ) ḡ + τ g.
        end if
        if d_t = 1 then
            break
        end if
    end for
end for
Return greedy policy: π(s) = arg max_a min{Q_1(s, a), Q_2(s, a)}.
```

returns than DQN at severe shifts ($\geq 60\%$) and typically a 5–15% gain over the GOLF baselines around the 40–50% shift region. For *pole-length changes* between 25% and 200% of nominal, RFL-TV stays near 500 reward throughout, while the best non-robust baseline ranges between ≈ 330 and 480, yielding about 5–50% higher return depending on the shift. Overall, for a fixed function class, turning on robustness in the Bellman update (via $\sigma > 0$ and the dual term) yields substantially better robustness to both action noise and dynamics misspecification than any of the non-robust functional approximation baselines. These trends also highlight that robustness is inherently σ-dependent: for a fixed training robustness level, performance eventually degrades as the test-time perturbation grows, so maintaining high returns under stronger shifts typically requires training with a larger σ and, in practice, possibly a more expressive function class.



(a) Action Perturbation (b) Force-magnitude Perturbation (c) Pole-length Perturbation

Figure 1: RFL-TV vs. Functional Approximation Algorithms

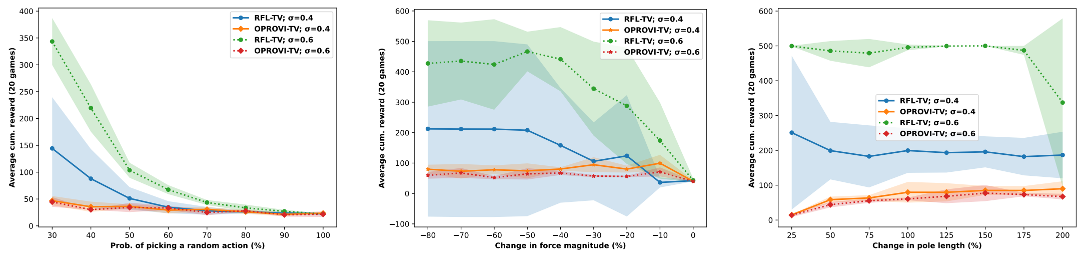
B.3 RFL-TV vs. Online Tabular TV-RMDP

Figure 2 evaluates how closely our practical RFL-TV implementation matches an ideal TV-robust planner by comparing it to OPROVI-TV (Lu et al., 2024), a tabular algorithm that exactly solves the TV-robust Bellman equations for a given radius σ. Although OPROVI-TV is restricted to small state spaces such as CartPole, it serves as a strong oracle-style baseline for TV-robust planning. In contrast, our practical RFL-TV implementation operates with neural function classes and sample-based updates, so its per-iteration computational cost depends on the network sizes, batch size, and action-space cardinality A, but *not* on the number of states S, making it applicable to large-scale problems where typically $S \gg A$. Across action perturbations and dynamics perturbations (force magnitude and pole length), RFL-TV with $\sigma \in \{0.4, 0.6\}$ consistently matches, and often exceeds the returns of OPROVI-TV at the same σ.

For action perturbations (random-action probability $\rho \in [0.3, 0.7]$), RFL-TV with $\sigma = 0.6$ achieves between roughly 100% and 400% higher average return than OPROVI-TV, while $\sigma = 0.4$ yields gains on the order of 30%–200% depending on the noise level; the two methods converge to similar near-random performance only as ρ approaches 1. Under force-magnitude perturbations, RFL-TV with $\sigma = 0.6$ improves over OPROVI-TV by about 100%–300% at large changes (40%–80% deviation from nominal), and $\sigma = 0.4$ still offers roughly 30%–150% gains. For pole-length perturbations, RFL-TV with $\sigma = 0.6$ maintains returns that are typically 150%–300% higher than the tabular baseline over most of the tested range, whereas $\sigma = 0.4$ yields about 30%–150% improvements. Overall, these trends indicate that a simple two-layer ReLU MLP (with 128–256 hidden units for both Q and dual networks) can closely track—and often outperform—the robust value structure computed by an exact tabular TV-RMDP solver, while enjoying computational complexity that scales with network size and A rather than S, which is particularly advantageous in regimes where $S \gg A$.

B.4 Balancing Robustness Radius and Dual-Network Capacity

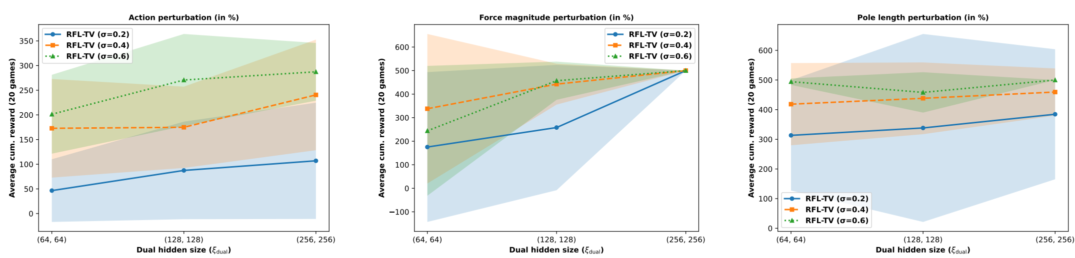
Figure 3 examines how the TV robustness radius σ and the dual-network width ξ_{dual} jointly shape the performance of RFL-TV. For each perturbation family (action noise, force–magnitude scaling, and pole–length scaling), we vary ξ_{dual} over two-layer MLPs with hidden sizes $(64, 64)$, $(128, 128)$, and $(256, 256)$ and evaluate RFL-TV for $\sigma \in \{0.2, 0.4, 0.6\}$ at a representative perturbation level. Note that enlarging the dual hidden size can only decrease the approximation gap ξ_{dual} to the ideal dual



(a) Action Perturbation (b) Force-magnitude Perturbation (c) Pole-length Perturbation

Figure 2: RFL-TV vs. OPROVI-TV (Tabular).

optimizer; in other words, we can view the dual width as a structural knob that monotonically reduces the realizability constant ξ_{dual}. Across all three families, increasing the dual capacity markedly improves robustness: moving from $(64, 64)$ to $(256, 256)$ yields roughly 40%–120% higher average return under action perturbations, about 50%–180% gains for force–magnitude shifts, and roughly 100%–250% gains for pole–length perturbations. At any fixed ξ_{dual}, larger robustness radii clearly help: compared to $\sigma = 0.2$, using $\sigma = 0.6$ improves returns by about 60%–160% under action noise, 30%–80% under force–magnitude changes, and 50%–150% under pole–length changes, with $\sigma = 0.4$ typically lying in between. This behaviour is natural: when σ is too small, the uncertainty set remains close to the nominal dynamics and the dual term contributes less, so the policy tends to overfit to the unperturbed environment and degrades sharply under shift. Larger radii ($\sigma \approx 0.4$–0.6), together with a sufficiently expressive dual network, force the optimizer to hedge against adversarial transitions, leading to policies that are more conservative around failure modes yet still high-reward under the moderately perturbed environments we evaluate on. In practice, these results suggest a simple tuning recipe: increase ξ_{dual} until the robust return curve flattens, and select σ in a moderate range where performance gains saturate (here around 0.4–0.6), thereby jointly controlling approximation quality and the strength of robustness.



(a) Action Perturbation (b) Force-magnitude Perturbation (c) Pole-length Perturbation

Figure 3: RFL-TV: uncertainty level σ vs. Uniform dual-approximation error ξ_{dual}.

C PROOF OF THE MAIN RESULTS

Assumption 3 ((Panaganti et al., 2022; 2024))**.** *For all $f \in \mathcal{F}$ and any policy π, there exists a uniform constant ξ_{dual} such that*

$$\inf_{g\in\mathcal{G}} \text{Dual}_{loss}(g; f) - \inf_{g\in\mathcal{L}^1(\mu^\pi)} \text{Dual}_{loss}(g; f) \;\leq\; \xi_{\text{dual}},$$

where μ^π is the visitation distribution induced by π under $P^\star$.

This assumption is not restrictive. Specially, note that $\mathcal{L}^1$ can be approximated by deep/wide neural networks (Goodfellow et al., 2016), which ensures Assumption 3 with such neural network classes.

We denote the robust Bellman operator as

$$[\mathcal{T}^\sigma f](s,a) = r(s,a) - \inf_{\eta\in[0,2H/\sigma]} \left\{ \mathbb{E}_{s'\in P_h^\star(s,a)} \left[\left(\eta - \max_{a'} f(s',a')\right)_+ \right] - \left(1-\sigma\right)\eta \right\}. \quad (13)$$

And we define the empirical duality loss as:

$$\widehat{\mathrm{Dual}}_{loss}(g; f) = \sum_{(s,a,s')\sim\mathcal{D}} \big((g(s,a) - \max_{a'} f(s',a'))_+ - (1-\sigma)g(s,a\big), \tag{14}$$

C.1 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

  $$C_h^{\mathrm{cv}} := \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sup_{\pi\in\Pi} d_h^{\pi,P^\omega}(s,a). \tag{15}$$

  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

  $$C_{\mathrm{rcov}} \;=\; \max_{h\in[H]} C_h^{\mathrm{cv}}. \tag{16}$$

- **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

  $$\varepsilon_{TV}^\sigma(f,\pi^f,h;Q) := \mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^f,Q}}\Big[f_h(s_h,a_h) - [\mathcal{T}_h^\sigma f_{h+1}](s_h,a_h)\Big]. \tag{17}$$

  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

  $$\begin{aligned}\mathrm{Regret}(K) &\le \sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\Big[f_h^{(k)}(s_h,a_h) - [\mathcal{T}_h^\sigma f_{h+1}^{(k)}](s_h,a_h)\Big],\\ &= \sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\Bigg[f_h^{(k)}(s_h,a_h) - \Big[\mathcal{T}^\sigma_{\underline{g}_{f_{h+1}^{(k)}},h} f_{h+1}^{(k)}\Big](s_h,a_h)\\ &\qquad + \Big[\mathcal{T}^\sigma_{\underline{g}_{f_{h+1}^{(k)}},h} f_{h+1}^{(k)}\Big](s_h,a_h) - [\mathcal{T}_h^\sigma f_{h+1}^{(k)}](s_h,a_h)\Bigg],\\ &= \mathrm{I} + \mathrm{II},\end{aligned} \tag{18}$$

  where we denote

  $$\mathrm{I} := \sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\Bigg[f_h^{(k)}(s_h,a_h) - \Big[\mathcal{T}^\sigma_{\underline{g}_{f_{h+1}^{(k)}},h} f_{h+1}^{(k)}\Big](s_h,a_h)\Bigg]. \tag{19}$$

  $$\mathrm{II} := \sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\Bigg[\Big[\mathcal{T}^\sigma_{\underline{g}_{f_{h+1}^{(k)}},h} f_{h+1}^{(k)}\Big](s_h,a_h) - [\mathcal{T}_h^\sigma f_{h+1}^{(k)}](s_h,a_h)\Bigg]. \tag{20}$$

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

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

Then, II can be written as

$$\text{II} := \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\left[\Delta_{k,h}(s_h,a_h)\right]. \tag{21}$$

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

$$\mathbb{E}_{d_h^{(k),P^\omega}}[\Delta_{k,h}] \;\le\; \left\|\frac{d_h^{(k),P^\omega}}{\mu_h^\pi}\right\|_\infty \|\Delta_{k,h}\|_{1,\mu_h^\pi}, \tag{22}$$

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

$$\left\|\frac{d_h^{(k),P^\omega}}{\mu_h^\pi}\right\|_\infty \;\le\; C_{\text{rcov}}. \tag{23}$$

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

$$\|\Delta_{k,h}\|_{1,\mu_h^\pi} = \mathcal{O}\left(\frac{H}{\sigma}\sqrt{\frac{2\log\big(8|\mathcal{G}||\mathcal{F}|KH/\delta\big)}{|\mathcal{D}_h^{(k)}|}} + \frac{\xi_{\text{dual}}}{KH}\right). \tag{24}$$

Step 3: Combine bounds. Hence, by combining eq. 22, eq. 23 and eq. 24, we get

$$\mathbb{E}_{d_h^{(k),P^\omega}}[\Delta_{k,h}] = \mathcal{O}\left(C_{\text{rcov}}\frac{H}{\sigma}\sqrt{\frac{2\log\big(8|\mathcal{G}||\mathcal{F}|KH/\delta\big)}{|\mathcal{D}_h^{(k)}|}} + C_{\text{rcov}}\frac{\xi_{\text{dual}}}{KH}\right). \tag{25}$$

Step 4: Summing over $k,h \in [K]\times[H]$. Summing the bound in eq. 25 over $k\in[K]$ and $h\in[H]$ yields the desired result:

$$\text{II} = \mathcal{O}\left(C_{\text{rcov}}\frac{H}{\sigma}\sqrt{2\log\Big(\tfrac{8|\mathcal{G}||\mathcal{F}|KH}{\delta}\Big)}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{|\mathcal{D}_h^{(k)}|}} + C_{\text{rcov}}\xi_{\text{dual}}\right). \tag{26}$$

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. 26 can be bounded by the following intergral, as

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{|\mathcal{D}_h^{(k)}|}} = \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{k}} \le H\left(1+\int_1^K\frac{dx}{\sqrt{x}}\right) = 2H\sqrt{K} - H \le 2H\sqrt{K}. \tag{27}$$

Applying eq. 27 in eq. 26, we get the final bound as

$$\text{II} = \mathcal{O}\left(C_{\text{rcov}}\,\frac{H^2}{\sigma}\,\sqrt{2K\log\Big(\tfrac{8|\mathcal{G}||\mathcal{F}|KH}{\delta}\Big)} \;+\; C_{\text{rcov}}\xi_{\text{dual}}\right). \tag{28}$$

- **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

$$\delta_h^{(k)}(\cdot,\cdot) := f_h^{(k)}(\cdot,\cdot) - \left[\mathcal{T}^\sigma_{\underline{g}_{f_{h+1}^{(k)}},h} f_{h+1}^{(k)}\right](\cdot,\cdot). \tag{29}$$

Then, I can be written as

$$\mathrm{I} := \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim d_h^{(k),P^\omega}}\left[\delta_h^{(k)}(s_h,a_h)\right]. \tag{30}$$

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

$$\widetilde{d}_h^{(k)}(s,a) \;:=\; \sum_{i=1}^{k-1} d_h^{(i),P^\omega}(s,a). \tag{31}$$

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 $k\in[K]$:

$$Q^{\star,\sigma}\in\mathcal{F}^{(k)} \quad \text{and} \quad \sum_{(s,a)} \tilde{d}_h^{(k)}(s,a)\left(\delta_h^{(k)}(s,a)\right)^2 \;\le\; \mathcal{O}(\beta). \tag{32}$$

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

$$\tau_h(s,a) \;=\; \min\Big\{\, t \;\big|\; \tilde{d}_h^{(t)}(s,a) \;\ge\; C_{\mathrm{rcov}}\cdot\mu_h^\pi(s,a)\Big\}, \tag{33}$$

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":

$$I = \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d_h^{(k),P^\omega}}\Big[\delta_h^{(k)}(s,a)\,\mathbb{I}\{k<\tau_h(s,a)\}\Big]}_{\text{conservative burn-in phase}} \tag{34}$$

$$+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d_h^{(k),P^\omega}}\Big[\delta_h^{(k)}(s,a)\,\mathbb{I}\{k\ge\tau_h(s,a)\}\Big]}_{\text{stable phase}}. \tag{35}$$

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

$$\mathbb{E}_{(s,a)\sim d_h^{(k),P^\omega}}\Big[\delta_h^{(k)}(s,a)\,\mathbb{I}\{k<\tau_h(s,a)\}\Big] \leq c_3\frac{H}{\sigma}\sum_{s,a} d_h^{(k),P^\omega}(s,a)\mathbb{I}\{k<\tau_h(s,a)\}. \tag{36}$$

Plugging eq. 36 in the conservative burn-in phase term of eq. 34, we get

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d_h^{(k),P^\omega}}\Big[\delta_h^{(k)}(s,a)\,\mathbb{I}\{k<\tau_h(s,a)\}\Big] \\
&\overset{(a)}{\leq} c_3\frac{H}{\sigma}\sum_{k=1}^{K}\sum_{h=1}^{H} d_h^{(k),P^\omega}(s,a)\mathbb{I}\{k<\tau_h(s,a)\} \\
&= c_3\frac{H}{\sigma}\sum_{h=1}^{H}\sum_{s,a}\sum_{k<\tau_h(s,a)} d_h^{(k),P^\omega}(s,a) \\
&\overset{(b)}{=} c_3\frac{H}{\sigma}\sum_{h=1}^{H}\sum_{s,a}\widetilde{d}_h^{\tau_h(s,a)}(s,a) \\
&= c_3\frac{H}{\sigma}\sum_{h=1}^{H}\sum_{s,a}\Big\{\widetilde{d}_h^{\tau_h(s,a)-1}(s,a) + d_h^{\tau_h(s,a)-1,P^\omega}(s,a)\Big\} \\
&\overset{(c)}{\leq} c_3\frac{H}{\sigma}\sum_{h=1}^{H}\sum_{s,a}\Big\{2C_{\mathrm{rcov}}\,\mu_h^\pi(s,a)\Big\} \\
&\overset{(d)}{=} c_3\frac{H^2}{\sigma}C_{\mathrm{rcov}}.
\end{aligned} \tag{37}$$

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. 31; ineq. (c) is due to eq. 33 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:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d_h^{(k),P^\omega}}\big[\delta_h^{(k)}(s,a)\,\mathbb{I}\{k\geq\tau_h(s,a)\}\big] \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a} d_h^{(k),P^\omega}(s,a)\left(\frac{\widetilde{d}_h^{(k)}(s,a)}{\widetilde{d}_h^{(k)}(s,a)}\right)^{1/2}\delta_h^{(k)}(s,a)\,\mathbb{I}\{k\geq\tau_h(s,a)\} \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a} d_h^{(k),P^\omega}(s,a)\left(\frac{\widetilde{d}_h^{(k)}(s,a)}{\widetilde{d}_h^{(k)}(s,a)}\right)^{1/2}\delta_h^{(k)}(s,a)\,\mathbb{I}\{k\geq\tau_h(s,a)\} \\
&\leq \sum_{h=1}^{H}\underbrace{\left(\sum_{k=1}^{K}\sum_{s,a}\frac{\big(\mathbb{I}\{t\geq\tau_h(x,a)\},d_h^{(k),P^\omega}(s,a)\big)^2}{\widetilde{d}_h^{(k)}(s,a)}\right)^{1/2}}_{\text{(A): extrapolation error}}\underbrace{\left(\sum_{k=1}^{K}\sum_{s,a}\tilde{d}_h^{(k)}(s,a)\,\big(\delta_h^{(k)}(s,a)\big)^2\right)^{1/2}}_{\text{(B): in-sample squared Bellman error}},
\end{aligned} \tag{38}$$

where the last inequality is Cauchy–Schwarz.

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

$$\text{(B)} \;\leq\; \mathcal{O}(\sqrt{\beta K}). \tag{39}$$

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. 33, for all $t \geq \tau_h(s, a)$ we have $\widetilde{d}_h^{(k)}(s, a) \geq C_{\text{rcov}} \mu_h^\pi(s, a)$, which implies $\widetilde{d}_h^{(k)}(s, a) \geq \frac{1}{2}\left(\widetilde{d}_h^{(k)}(s, a) + C_{\text{rcov}} \mu_h^\pi(s, a)\right)$. Thus,

$$
\begin{aligned}
(A) &= \sqrt{\sum_{k=1}^{K} \sum_{s,a} \frac{\left(\mathbb{I}\{k \geq \tau_h(s,a)\} d_h^{(k),P^\omega}(s,a)\right)^2}{\widetilde{d}_h^{(k)}(s,a)}} \\
&\leq \sqrt{2 \sum_{k=1}^{K} \sum_{s,a} \frac{d_h^{(k),P^\omega}(s,a) \cdot d_h^{(k),P^\omega}(s,a)}{\widetilde{d}_h^{(k)}(s,a) + C_{\text{rcov}} \cdot \mu_h^\pi(s,a)}} \\
&\leq \sqrt{2 \sum_{k=1}^{K} \sum_{s,a} \max_{\ell \in [K]} d_h^{(l),P^\omega}(s,a) \frac{d_h^{(k),P^\omega}(s,a)}{\widetilde{d}_h^{(k)}(s,a) + C_{\text{rcov}} \cdot \mu_h^\pi(s,a)}} \\
&\leq \sqrt{2 \left(\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \sum_{k=1}^{K} \frac{d_h^{(k),P^\omega}(s,a)}{\widetilde{d}_h^{(k)}(s,a) + C_{\text{rcov}} \cdot \mu_h^\star(s,a)} \right) \left(\sum_{s,a} \max_{l \in [K]} d_h^{(l),P^\omega}(s,a) \right)} \\
&\leq \mathcal{O}\left(\sqrt{C_{\text{rcov}} \log K}\right),
\end{aligned} \tag{40}
$$

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

To conclude, substitute eq. 39 and eq. 40 into eq. 38 to obtain

$$
\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{(s,a) \sim d_h^{(k),P^\omega}} \left[\delta_h^{(k)}(s,a)\, \mathbb{I}\{k \geq \tau_h(s,a)\}\right] \;\leq\; \mathcal{O}\left(H \sqrt{C_{\text{rcov}} \cdot \beta K \log K}\right). \tag{41}
$$

By applying eq. 37 and eq. 41 in eq. 34, we get

$$
I \leq \mathcal{O}\left(\frac{H^2}{\sigma} C_{\text{rcov}} + H \sqrt{C_{\text{rcov}} \cdot \beta K \log K}\right). \tag{42}
$$

Therefore, by applying eq. 42 and eq. 28 in eq. 18, we get

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

This concludes the proof of Theorem 1. □

D Specialization to Linear TV-RMDP

We now show that our regret bound for general functional approximation specializes to a near–dimension-optimal bound when the robust value function admits a linear representation, in the spirit of the d-rectangular linear RMDP framework of (Ma et al., 2022) and (Liu et al., 2024).

Assumption 4 (d-Rectangular Linear TV-RMDP)**.** *There exists a known feature map $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ for each $h \in [H]$ with $\sum_{i=1}^{d} \phi_{h,i}(s, a) = 1$ and $\phi_{h,i}(s, a) \geq 0$ for any $(i, s, a) \in [d] \times \mathcal{S} \times \mathcal{A}$ such that:*

(i) (Linear nominal model.) The reward and nominal kernel are linear:

$$
r_h(s,a) = \boldsymbol{\phi}_h(s,a)^\top \boldsymbol{\theta}_h, \qquad P_h^\star(\cdot \mid s,a) = \boldsymbol{\phi}_h(s,a)^\top \boldsymbol{\nu}_h^\star(\cdot),
$$

for some unknown probability measures $\{\boldsymbol{\nu}_h^\star\}_{h=1}^H$ over $\mathcal{S}$ and known vectors $\{\boldsymbol{\theta}_h\}_{h=1}^H$ with $\|\boldsymbol{\theta}_h\|_2 \leq \sqrt{d}$.

(ii) (d-rectangular TV uncertainty set.) *For each step h and feature index $i \in [d]$ we can parameterize our uncertainty set $\mathcal{P}$ by $\{\boldsymbol{\nu}_h^\star\}_{h=1}^H$, and thereby, can be defined as $\mathcal{P} = \mathcal{U}^\sigma(P^\star) = \bigotimes_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \mathcal{U}_h^\sigma(s,a;\boldsymbol{\nu}_h^\star)$, where $\mathcal{U}_h^\sigma(s,a;\boldsymbol{\nu}_h^\star)$ is defined as*

$$\mathcal{U}_h^\sigma(s,a;\boldsymbol{\nu}_h^\star) \triangleq \left\{ \sum_{i=1}^d \phi_{h,i}(s,a)\nu_{h,i}(\cdot) : \nu_{h,i} \in \Delta(\mathcal{S}) \text{ and } D_{TV}(\nu_{h,i}, \nu_{h,i}^\star(\cdot|s,a)) \le \sigma \right\}.$$

This is the TV analogue of the d-rectangular linear RMDP of (Liu et al., 2024, Sec. 3.2), specialized to TV divergence.

Linear function classes induced by the d-Rectangular linear TV-RMDP. Under the linear TV-RMDP structure in Assumption 4, we specialize our general functional class $\mathcal{F}$ and dual functional class $\mathcal{G}$ used by RFL-TV as linear function classes with a common feature map $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, and are denoted as follows:

$$\mathcal{F}^{lin} := \{\mathcal{F}_h^{lin}\}_{h=1}^H, \text{ where } \mathcal{F}_h^{lin} := \left\{ f_h : f_h(s,a) = \boldsymbol{\phi}_h(s,a)^\top \boldsymbol{w}_h,\ \boldsymbol{w}_h \in \mathbb{R}^d \right\}, \tag{43}$$

$$\mathcal{G}^{lin} := \{\mathcal{G}_h^{lin}\}_{h=1}^H, \text{ where } \mathcal{G}_h^{lin} := \left\{ g_h : g_h(s,a) = \boldsymbol{\phi}_h(s,a)^\top \boldsymbol{u}_h,\ \boldsymbol{u}_h \in \mathbb{R}^d \right\}. \tag{44}$$

The class $\mathcal{F}^{lin}$ is used to approximate robust Q–functions, while $\mathcal{G}^{lin}$ parameterizes the dual variables appearing in the TV–robust Bellman operator (via the functional dual loss in Eq. 8)[See Sec. 4 for the definition of the dual loss and its empirical counterpart].

Lemma 2 (Linear realizability and completeness)**.** *Suppose the linear RMDP satisfies Assumption 4. Then:*

(i) **Linear realizability of $Q^{\pi,\sigma}$ and $Q^{\star,\sigma}$.** *For any Markov policy π and any $\sigma \ge 0$, there exist vectors $\boldsymbol{w}_1^{\pi,\sigma}, \dots, \boldsymbol{w}_H^{\pi,\sigma} \in \mathbb{R}^d$ such that for all $h \in [H]$,*

$$Q_h^{\pi,\sigma}(s,a) = \boldsymbol{\phi}_h(s,a)^\top \boldsymbol{w}_h^{\pi,\sigma}, \qquad \forall (s,a) \in \mathcal{S}\times\mathcal{A}, \tag{45}$$

and, in particular, for the robust-optimal policy $\pi^\star$ there exist $\boldsymbol{w}_1^{\star,\sigma}, \dots, \boldsymbol{w}_H^{\star,\sigma}$ with

$$Q_h^{\star,\sigma}(s,a) = \boldsymbol{\phi}_h(s,a)^\top \boldsymbol{w}_h^{\star,\sigma}, \qquad \forall (s,a),\ h \in [H]. \tag{46}$$

Hence $Q^{\pi,\sigma}, Q^{\star,\sigma} \in \mathcal{F}^{lin}$.

(ii) **Closure under the robust Bellman operator.** *Let $f \in \mathcal{F}^{lin}$ with component functions $f_h(s,a) = \boldsymbol{\phi}_h(s,a)^\top \boldsymbol{w}_h$. Then, for each $h \in [H]$ there exists $\boldsymbol{w}_h' \in \mathbb{R}^d$ such that the robust Bellman backup satisfies*

$$\begin{aligned}[\mathcal{T}_h^\sigma f_{h+1}](s,a) &= r_h(s,a) + \mathbb{E}_{P\in\mathcal{U}_{\sigma,h}(s,a)}\left[V_{h+1}(s')\right] \\ &= \boldsymbol{\phi}_h(s,a)^\top \boldsymbol{w}_h', \qquad \forall (s,a),\end{aligned} \tag{47}$$

so that $\mathcal{T}_h^\sigma f_{h+1} \subseteq \mathcal{F}_h^{lin}$ for all h.

(iii) **Linear dual representation.** *For any $f \in \mathcal{F}^{lin}$, the dual minimizer $g_f^\star$ that attains the pointwise TV dual can be chosen in $\mathcal{G}^{lin}$, i.e., there exist $\boldsymbol{u}_1^f, \dots, \boldsymbol{u}_H^f \in \mathbb{R}^d$ such that*

$$g_{f,h}^\star(s,a) = \boldsymbol{\phi}_h(s,a)^\top \boldsymbol{u}_h^f, \qquad \forall (s,a),\ h \in [H]. \tag{48}$$

Consequently, the dual realizability error ξ_{dual} defined in Assumption 3 is zero when we take $\mathcal{G}^{lin} \equiv \mathcal{L}^1(\mu^\pi)$.

Proof. (i) Linear realizability of $Q^{\pi,\sigma}$ and $Q^{\star,\sigma}$. The linear robust MDP literature (e.g., (Ma et al., 2022, Prop. 3.2 and Lem. 4.1) and (Liu et al., 2024, Sec. 3.2)) implies that both the robust Bellman operator and the robust value functions preserve linearity in $\boldsymbol{\phi}_h$, yielding 45–47, the nominal kernel and all kernels in the TV uncertainty set are linear mixtures of the base measures $\{\boldsymbol{\nu}_h\}_{h=1}^H$, and the reward is linear in $\boldsymbol{\phi}_h$.

(ii) Closure under $\mathcal{T}_h^\sigma$. Let $f \in \mathcal{F}^{lin}$ with $f_{h+1}(s,a) = \boldsymbol{\phi}_{h+1}(s,a)^\top \boldsymbol{w}_{h+1}$. Define the value $V_{h+1}(s) = \max_{a\in\mathcal{A}} f_{h+1}(s,a)$. By the d-rectangular structure, any $P \in \mathcal{U}_h^\sigma(s,a)$ can be written as $P(\cdot \mid s,a) = \sum_{i=1}^d \phi_{h,i}(s,a)\,\nu_{h,i}(\cdot)$ with $\nu_{h,i} \in \mathcal{U}_h^\sigma(s,a;\boldsymbol{\nu}_h^\star)$. Thus

$$\inf_{P_h\in\mathcal{U}_h^\sigma(s,a)} \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\left[V_{h+1}^{\pi,\sigma}(s')\right] = \inf_{\nu_{h,1},\dots,\nu_{h,d}} \sum_{i=1}^d \phi_{h,i}(s,a)\,\mathbb{E}_{s'\sim\mu_{h,i}}[V_{h+1}(s')] \tag{49}$$

$$= \sum_{i=1}^d \phi_{h,i}(s,a) \inf_{\nu_{h,i}\in\mathcal{U}_h^\sigma(s,a;\boldsymbol{\nu}_h^\star)} \mathbb{E}_{s'\sim\nu_{h,i}}[V_{h+1}(s')] \tag{50}$$

$$= \sum_{i=1}^d \phi_{h,i}(s,a)\,\zeta_{h,i}(\boldsymbol{w}_{h+1}), \tag{51}$$

where each scalar $\zeta_{h,i}(\boldsymbol{w}_{h+1})$ depends only on V_{h+1} (and hence on $\boldsymbol{w}_{h+1}$) and the local TV ball at index i. We therefore obtain

$$[\mathcal{T}_h^\sigma f_{h+1}](s,a) = \boldsymbol{\phi}_h(s,a)^\top\boldsymbol{\theta}_h + \boldsymbol{\phi}_h(s,a)^\top\boldsymbol{\zeta}_h(\boldsymbol{w}_{h+1}) = \boldsymbol{\phi}_h(s,a)^\top\boldsymbol{w}_h', \tag{52}$$

with $\boldsymbol{w}_h' := \boldsymbol{\theta}_h + \boldsymbol{\zeta}_h(\boldsymbol{w}_{h+1})$. This yields 47 and shows that $\mathcal{T}_h^\sigma\{_{h+1} \subseteq \mathcal{F}_h^{lin}$.

(iii) Linear dual representation. Fix any $f \in \mathcal{F}^{lin}$ and (s,a,h). The TV dual expression for the robust Bellman operator (Eq. 13) writes the inner worst-case expectation as a one-dimensional convex optimization problem in a scalar dual variable η. Under the linear RMDP structure, $P_h^\star(\cdot \mid s,a) = \sum_{i=1}^d \phi_{h,i}(s,a)\nu_{h,i}^\star$, so the dual term is a weighted sum over coordinates i, and the optimal dual variable can be decomposed into coordinate-wise scalar duals $\eta_{h,i}^\star$ associated with each base measure $\nu_{h,i}^\star$ (see, e.g., the TV dual derivation in (Liu et al., 2024, Sec. 3.2). This yields an optimal dual function of the form

$$g_{f,h}^\star(s,a) = \sum_{i=1}^d \phi_{h,i}(s,a)\,\eta_{h,i}^\star = \boldsymbol{\phi}_h(s,a)^\top\boldsymbol{u}_h^f \tag{53}$$

for some $\boldsymbol{u}_h^f \in \mathbb{R}^d$. Collecting these across h we obtain $g_f^\star \in \mathcal{G}^{lin}$ as in 48. In particular, the infimum in the dual representation is attained within $\mathcal{G}^{lin}$, so the dual realizability error ξ_{dual} defined in Assumption 3 is zero when we set $\mathcal{G}^{lin} \equiv \mathcal{L}^1(\mu^\pi)$. □

Assumption 5 (Finite linear covering). *For $\varepsilon_0 = 1/K$, the union class $\mathcal{H} = \mathcal{F}^{lin} \cup \mathcal{G}^{lin}$ admits a finite ε_0-cover in $\|\cdot\|_\infty$ such that*

$$\log N_{\mathcal{H}}(\varepsilon_0) \;\le\; c_o dH \,\log\left(c_o K\right) \tag{54}$$

for some absolute constant $c_o > 0$.

This bound follows from standard metric-entropy results for linear predictors on a bounded domain (see, e.g., (Shalev-Shwartz & Ben-David, 2014, Thm. 14.5)). In our setting, the feature vectors satisfy the simplex constraints $\sum_i \phi_{h,i}(s,a) = 1$ and $\phi_{h,i}(s,a) \ge 0$ for all (s,a,h), which immediately implies $|\phi_h(s,a)|2 \le 1$. Together with the fact that the parameter vectors of $\mathcal{F}^{lin}$ and $\mathcal{G}^{lin}$ are restricted to a bounded ball, this ensures that every function in the union class $\mathcal{H} = \mathcal{F}^{lin} \cup \mathcal{G}^{lin}$ behaves as a linear predictor in an ambient space of dimension $d^{lin} = dH$, yielding a covering-number bound of the form $\log N_{\mathcal{H}}(\varepsilon_0) \le c_o dH \log\big(c_o/\varepsilon_0\big)$ for some absolute constant c_o.

Theorem 2 (Regret of RFL-TV in linear TV-RMDP). *For any $\delta \in (0,1]$, we set*

$$\beta = \mathcal{O}\Big(\min\{H, 1/\sigma\}\big(d^2 H \log(dKH/\delta)\big)\Big)$$

in RFL-TV. Then under Assumption 1–5, there exists an absolute constant $C > 0$ such that with probability at least $1-\delta$, it holds that

$$\textit{Regret}(K) = \mathcal{O}\Big(\sqrt{C_{\text{rcov}}^2 H^4 (\min\{H, 1/\sigma\})^2\, d^2\, K}\; \log\Big(\tfrac{dHK}{\delta}\Big)\Big).$$

Comparison with non-robust linear MDP. In the standard (non-robust) linear MDP setting, UCRL–VTR$^+$ and its refinements (Zhou et al., 2020; Jin et al., 2020; He et al., 2023) attain the minimax regret rate $\widetilde{\mathcal{O}}\big(\sqrt{d^2H^3K}\big)$. For linear TV-RMDP (with $\sigma = 0$), RFL-TV instead guarantees $\widetilde{\mathcal{O}}\big(\sqrt{C_{\rm rcov}^2 d^2 H^6 K}\big)$. Thus, our bound matches the optimal dependence on the feature dimension d and the number of episodes K, but incurs an additional polynomial factor in the horizon H and a multiplicative dependence on the robust coverage coefficient $C_{\rm rcov}$, reflecting the extra difficulty of controlling robust Bellman errors under partial coverage. Accordingly, we do not claim minimax optimality in $(H, C_{\rm rcov})$.

Comparison with linear RMDPs. On the robust side, several recent works study DR-RL with linear structure, but under settings that are fundamentally different from our online setting. (Ma et al., 2022) and (Wang et al., 2024a) analyse *offline* DR-RL with linear function approximation and obtain value-estimation rates of order $\widetilde{\mathcal{O}}(\sqrt{d/N})$ or $\widetilde{\mathcal{O}}(\sqrt{d^3/N})$ depending on coverage, where N is the number of trajectories. In the *online* setting, (Liu & Xu, 2024b) and (Liu et al., 2024) study d-rectangular linear RMDPs where the agent interacts online with a nominal (source) environment but the performance criterion is the worst-case value over a perturbed (target) environment, and attains regret rate $\widetilde{\mathcal{O}}\Big(\sqrt{d^2H^2(\min\{H, 1/\sigma\})^2K}\Big)$ together with an information-theoretic lower bound that is optimal in (d, K, σ) up to a $\sqrt{H}$ factor (Liu et al., 2024). (Panaganti et al., 2024) consider a different hybrid setting for φ-divergence RMDPs with general function approximation, and derive performance guarantees that scale with an appropriate complexity measure of the value-function class, leveraging both an offline dataset and online interaction with a nominal model.

By contrast, in the d-rectangular linear setting, Theorem 2 shows that RFL-TV achieves the regret bound of order $\widetilde{\mathcal{O}}\Big(\sqrt{C_{\rm rcov}^2\, H^4 \big(\min\{H, 1/\sigma\}\big)^2\, d^2 K}\Big)$. In *moderate coverage* regimes, where the data distribution provides good on-dynamics coverage and $C_{\rm rcov} = \mathcal{O}(1)$, this simplifies to $\widetilde{\mathcal{O}}\Big(H^2 \min\{H, 1/\sigma\}\, \sqrt{d^2K}\Big)$, which is optimal in its dependence on (d, K, σ) and matches the linear RMDP minimax lower bounds of Liu et al. (2024) up to an additional $\mathcal{O}(\sqrt{H^3})$ factor in the horizon. In *hard coverage* regimes, where the on-dynamics data poorly covers the robustly relevant state–action space, our coverability analysis in Lemma K.4 allows $C_{\rm rcov}$ to scale as $C_{\rm rcov} = \mathcal{O}(d)$, and the regret bound deteriorates to $\widetilde{\mathcal{O}}\Big(H^2 \min\{H, 1/\sigma\}\, \sqrt{d^4K}\Big)$. Such a result is $\mathcal{O}(d^2H^2)$-worse than the online learning in linear robust MDPs in (Liu et al., 2024), and $\mathcal{O}(\sqrt{d^2H^3})$-worse than the minimax lower bound (Liu et al., 2024). This yields a dimension dependence consistent with existing minimax lower bounds while explicitly quantifying the price of poor coverage via $C_{\rm rcov}$. Closing the remaining $\sqrt{H^3}$ gap in the horizon dependence and establishing matching lower bounds for online DR-RL with general function approximation (recovering the linear RMDP lower bounds as a special case) remain important directions for future work.

Remark 6. *(Panaganti et al., 2024) studies a hybrid φ-regularized RMDP that combines an offline dataset with online interactions. Under approximate value and dual realizability, a bilinear model of dimension d, and an offline concentratability coefficient $C(\pi^\star)$, they obtain a suboptimality bound of order $\mathcal{O}\Big(\max\{C(\pi^\star), 1\}\, (\lambda + H)\sqrt{d^3H^2K}\Big)$. By contrast, we specialize our general theorem to a d-rectangular linear TV-RMDP and show that RFL-TV achieves robust regret $\widetilde{\mathcal{O}}\Big(\sqrt{C_{\rm rcov}^2 H^4 (\min\{H, 1/\sigma\})^2\, d^2 K}\Big)$, which is better than the one in (Panaganti et al., 2024). This comparison implies our algorithm achieves a tighter sample complexity, even without any prior collected offline dataset.*

We also highlight that, conceptually, the two setups address different questions and are not directly comparable. (Panaganti et al., 2024) analyze a regularized *robust objective in a hybrid offline–online regime, where the parameter λ controls a trade-off induced by a φ-regularizer and the guarantees depend on offline coverage through $C(\pi^\star)$. In contrast, we study a* constrained *TV-RMDP in a purely online (off-dynamics) setting, where robustness is enforced via an explicit divergence ball of radius σ around the nominal model and performance is measured by cumulative regret with respect to the unconstrained TV-robust value. Our general theorem Theorem 1 applies to arbitrary parametric function classes.*

D.1 Proof of Theorem 2

Proof. We recall the proof of Theorem 1 and show how it specializes under the linear TV-RMDP structure in Assumption 4 together with the linear function classes $\mathcal{F}^{lin}, \mathcal{G}^{lin}$ defined in equation 43–equation 44.

Step 1: Starting point from the general regret proof. By the definition of robust regret in 6 and following the prrof lines 17-20, the robust regret can be decomposed as

$$\text{Regret}(K) \;\le\; I + II, \tag{55}$$

where I is the *Bellman-error term 19* term and II is the *Dual-approximation term 20* term.

By 42 and 28, we can bound I and II, respectively, as

$$\text{I} \le \mathcal{O}\Big(C_{\text{rcov}}\frac{H^2}{\sigma} + H\sqrt{C_{\text{rcov}}\,\beta K \log K}\Big) \tag{56}$$

$$\text{II} \le \mathcal{O}\Big(C_{\text{rcov}}\frac{H^2}{\sigma}\sqrt{K\log\Big(\frac{8|\mathcal{F}||\mathcal{G}|KH}{\delta}\Big)} + C_{\text{rcov}}\,\xi_{\text{dual}}\Big), \tag{57}$$

The factor $\log(|\mathcal{F}||\mathcal{G}|)$ here comes from the union bounds over the function classes in the concentration arguments, and ξ_{dual} is the dual realizability bias in Assumption 3.

Our goal is to rerun these two bounds when we instantiate $\mathcal{F} = \mathcal{F}^{lin}$ and $\mathcal{G} = \mathcal{G}^{lin}$ under the linear TV-RMDP structure.

Step 2: Consequences of the linear TV-RMDP structure. For better clarity, we work under the exact dual realizability condition, and we set $\xi_{\text{dual}} = 0$ for simplicity of proof [3].

Under Assumption 4, the linear classes $\mathcal{F}^{lin}, \mathcal{G}^{lin}$ together with Lemma 2 guarantee that all structural assumptions used in Theorem 1 remain valid when we instantiate the analysis with the linear TV-RMDP; the only resulting changes are as follows:

- The complexity term $\log(|\mathcal{F}||\mathcal{G}|)$ is replaced by a covering-number bound for the union class $\mathcal{H} \triangleq \mathcal{F}^{lin} \cup \mathcal{G}^{lin}$. By Assumption 5, for $\varepsilon_0 = 1/K$, the union class $\mathcal{H} = \mathcal{F}^{lin} \cup \mathcal{G}^{lin}$ admits an ε_0-cover in $\|\cdot\|_\infty$ with $\log N_{\mathcal{H}}(\varepsilon_0) \le c_0\, dH\, \log(c_0 K)$, for some absolute constant $c_0 > 0$ (Shalev-Shwartz & Ben-David, 2014).
- The dual bias term $C_{\text{rcov}}\xi_{\text{dual}}$ drops out.

The robust coverability constant C_{rcov} is unchanged, as it only depends on the failure-state assumption and the dynamics, not on the parametric structure of $\mathcal{F}$ and $\mathcal{G}$. Moreover, each stage h behaves as a d-dimensional linear class, and the full horizon class $\mathcal{H}$ has ambient dimension dH, yielding equation 54.

Step 4: Bounding II in the linear case. The derivation of the general bound equation 57 for II (Lemma K.3) uses ERM generalization bound Lemma T.1 and a union bound over all episodes, time steps, and function pairs $(f, g) \in \mathcal{F} \times \mathcal{G}$. In the linear case, we instead apply the same argument to a finite ε_0-net of $\mathcal{H}$.

More precisely, fix $\varepsilon_0 = 1/K$ and let $\mathcal{H}_0 \subset \mathcal{H}$ be a minimal ε_0-net under $\|\cdot\|_\infty$, such that $|\mathcal{H}_0| = N_{\mathcal{H}}(\varepsilon_0)$. We then repeat the concentration analysis of Lemma T.1, but take the union bound over the finite set $(k, h, \varphi) \in [K] \times [H] \times \mathcal{H}_0$ instead of $(k, h, f, g) \in [K] \times [H] \times \mathcal{F} \times \mathcal{G}$. The approximation error between any $f \in \mathcal{H}$ and its nearest neighbor $f' \in \mathcal{H}_0$ is at most ε_0 in $\|\cdot\|_\infty$ and

[3] By Lemma 2(iii), when we instantiate RFL-TV with the linear dual class $\mathcal{G}^{lin}$, the dual minimizer of the TV robust Bellman operator is exactly realizable, so the dual approximation error in Assumption 3 vanishes and we have $\xi_{\text{dual}} = 0$. For clarity, we therefore focus on this exact-realizability case in the sequel. If one instead works with a dual class that only approximately realizes the optimal dual (so $\xi_{\text{dual}} > 0$), the same proof strategy goes through with an additional additive term of order $C_{\text{rcov}}\,\xi_{\text{dual}}$ propagating from the bound on II (cf. 28) into the final regret bound; no other part of the argument needs to be modified, and the dependence on (K, d, H, σ) remains unchanged.

hence contributes only an $o(1)$ term in K to the final regret bound, which we absorb into the big–$\mathcal{O}$ notation.

Therefore, following the same steps of the proof of Lemma K.3 and setting $\xi_{\text{dual}} = 0$, we conclude that in the linear case equation 57 becomes

$$\text{II} \;\le\; \mathcal{O}\Big(C_{\text{rcov}}\frac{H^2}{\sigma}\sqrt{K\log\Big(\frac{8N_{\mathcal{H}}(\varepsilon_0)KH}{\delta}\Big)}\Big). \tag{58}$$

By covering-number bound equation 54, we obtain

$$\log\Big(\frac{8N_{\mathcal{H}}(\varepsilon_0)KH}{\delta}\Big) \;\le\; c_1 dH\log(c_1K) + \log\Big(\tfrac{c_2KH}{\delta}\Big) \;=:\; L_K, \tag{59}$$

for some absolute constants $c_1, c_2 > 0$. Hence, by applying 59 in 58, we get

$$\text{II} \;\le\; \mathcal{O}\Big(C_{\text{rcov}}\frac{H^2}{\sigma}\sqrt{K\,L_K}\Big), \qquad L_K = dH\log(c_1K) + \log\Big(\tfrac{c_2KH}{\delta}\Big). \tag{60}$$

Step 5: Bounding I in the linear case and choice of β. We now revisit the bound equation 56 on I. In the linear case, we again cover the union class $\mathcal{H}_0 \subset \mathcal{H}$ by a finite ε_0-net under $\|\cdot\|_\infty$, such that $|\mathcal{H}_0| = N_{\mathcal{H}}(\varepsilon_0)$, where we set $\varepsilon_0 = 1/(KH)$. Therefore, following the same steps of the proof of Lemma K.2 and using the same Freedman–cover argument as in Lemma T.4, but union-bounding over the finite set $N_{\mathcal{H}}(\varepsilon_0)$ instead of $\mathcal{F}\cup\mathcal{G}$, we obtain the same form of result with the complexity term $\log(|\mathcal{F}||\mathcal{G}|)$ replaced by $\log N_{\mathcal{H}}(\varepsilon_0)$. In particular, for $\varepsilon_0 = 1/(KH)$, Assumption 5 implies that $\log N_{\mathcal{H}}(\varepsilon_0)$ has the same order as in equation 54, so choosing

$$\beta = \mathcal{O}\Big((\min\{H, 1/\sigma\})^2\big(d^2H\log(c_1K) + \log(c_2KH/\delta)\big)\Big) = \mathcal{O}\big((\min\{H,1/\sigma\})^2 L_K\big) \tag{61}$$

is sufficient to reproduce the general bound equation 56, with the same constants as in Theorem 1. Substituting equation 61 into equation 56 yields

$$\begin{aligned}
\text{I} &\le \mathcal{O}\Big(C_{\text{rcov}}\frac{H^2}{\sigma} + H\sqrt{C_{\text{rcov}}\,\beta K\log K}\Big)\\
&= \mathcal{O}\Big(C_{\text{rcov}}\frac{H^2}{\sigma} + H\sqrt{C_{\text{rcov}}\,(\min\{H,1/\sigma\})^2 L_K\,K\log K}\Big)\\
&= \mathcal{O}\Big(C_{\text{rcov}}\frac{H^2}{\sigma} + H\sqrt{C_{\text{rcov}}(\min\{H,1/\sigma\})^2\,K\,L_K\,\log K}\Big).
\end{aligned} \tag{62}$$

Step 6: Combining the bounds of I and II. Combining equation 55, equation 60, and equation 62, we obtain

$$\text{Regret}(K) \le \mathcal{O}\Big(C_{\text{rcov}}\frac{H^2}{\sigma} + H\sqrt{C_{\text{rcov}}(\min\{H,1/\sigma\})^2\,K\,L_K\,\log K} + C_{\text{rcov}}\frac{H^2}{\sigma}\sqrt{K\,L_K}\Big). \tag{63}$$

The additive term $C_{\text{rcov}}H^2/\sigma$ is lower order than the $\sqrt{K}$ terms and can be absorbed into the leading big–$\mathcal{O}$ term. Also, since $K \ge 2$, $dH \ge 1$ and $\delta \in (0,1]$, we have $\frac{dHK}{\delta} \ge K$ and $\frac{dHK}{\delta} \ge \frac{1}{\delta}$, so $\log K \le \log(dHK/\delta)$ and $\log(1/\delta) \le \log(dHK/\delta)$, Using these facts and absorbing constants into $c_3 > 0$, we obtain

$$L_K \;\le\; c_3\Big(d^2H + \log\tfrac{1}{\delta}\Big)\log\Big(\tfrac{dHK}{\delta}\Big).$$

Moreover, $\log K \le \log(dHK/\delta)$, hence

$$\sqrt{L_K\log K} \;\le\; c_4\sqrt{d^2H + \log\tfrac{1}{\delta}}\;\log\Big(\tfrac{dHK}{\delta}\Big) = \mathcal{O}\Big(\sqrt{d^2H}\;\log\tfrac{dHK}{\delta}\Big),$$

where we used that $\log(dHK/\delta)$ dominates $\sqrt{\log(1/\delta)}$ and absolute constants c_4. Plugging this into equation 63, we deduce that

$$\text{Regret}(K) \;\le\; \mathcal{O}\Big(\sqrt{C_{\text{rcov}}^2H^4(\min\{H,1/\sigma\})^2\,d^2\,K}\;\log\Big(\tfrac{dHK}{\delta}\Big)\Big),$$

which proves the claim. □

D.2 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* $V^{\pi^f,Q} := \mathbb{E}_{a_{1:H}\sim\pi^f, s_{h+1}\sim Q_h}\left[\sum_{h=1}^H r_h(s_h, a_h)\right]$. *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. 17. Then, we can bound the regret as given in eq. 6 as,*

$$\text{Regret}(K) \le \sum_{k=1}^K \sum_{h=1}^H \varepsilon^\sigma_{TV}(f^{(k)}, \pi^{f^{(k)}}, h; P^\omega). \tag{64}$$

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 $\mathcal{T}^\sigma_h f$ in eq. 13, we get

$$[\mathcal{T}^\sigma_h f_{h+1}](s,a) = r_h(s,a) + \inf_{P\in\mathcal{U}^\sigma_h(s,a)} \mathbb{E}_P\left[\psi^f_{h+1}\right] \;\le\; r_h(s,a) + \mathbb{E}_{s'\sim Q_h(\cdot|s,a)}[\psi^f_{h+1}(s')]. \tag{65}$$

Thus, from eq. 65 we get

$$f_h(s,a) - [\mathcal{T}^\sigma_h f_{h+1}](s,a) \;\ge\; f_h(s,a) - r_h(s,a) - \mathbb{E}_{s'\sim Q_h}[\,\psi^f_{h+1}(s')\,]. \tag{66}$$

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

$$\sum_{h=1}^H \varepsilon^\sigma_{\mathrm{TV}}(f, \pi^f, h; Q) \;\ge\; \sum_{h=1}^H \mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^f,Q}}\left[f_h(s_h,a_h) - r_h(s_h,a_h) - \mathbb{E}_{Q_h}[\psi^f_{h+1}]\right]. \tag{67}$$

The right-hand side of eq. 67 follows the same proof-lines as in (Jiang et al., 2017, Lemma 1), yielding

$$\sum_{h=1}^H \varepsilon^\sigma_{\mathrm{TV}}(f, \pi^f, h; Q) \;\ge\; V^f - V^{\pi^f,Q}. \tag{68}$$

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

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

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 $V_h^{\star,\sigma}(s) \le V_h^{f^{(k)}}(s)$. Therefore, we can write

$$\begin{aligned}
\text{Regret}(K) &= \sum_{k=1}^K V_1^{\star,\sigma}(s_1) - V_1^{\pi^{(k)},\sigma}(s_1) \\
&\le \sum_{k=1}^K V_1^{f^{(k)}}(s_1) - V_1^{\pi^{(k)},P^\omega}(s_1) \\
&\le \sum_{k=1}^K \sum_{h=1}^H \varepsilon^\sigma_{\mathrm{TV}}(f^{(k)}, \pi^{f^{(k)}}, h; P^\omega) \qquad \text{[By eq. 68]}.
\end{aligned}$$

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

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 $f^{(k)} = \{f_1^{(k)}, \ldots, f^{(k)}{}_H\} \in \mathcal{F}$. 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 $\|f\|_\infty, \|\psi^f\|_\infty \leq H$ (this is the boundedness assumption used throughout). We consider the filtration induced as

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

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

We obtain $\underline{g}_{f_h} \in [0, 2H/\sigma]$ as a measurable minimizer of eq. 14 that satisfies Assumption 3. For the trajectory of episode k, we define

$$Z_h^{(k)}(f, \underline{g}_f) := \left(\underline{g}_{f^{(k)}_{h+1}}(s_h^k, a_h^k) - \psi^{f^{(k)}}_{h+1}(s_h^k, a_h^k)\right)_+ - (1-\sigma)\underline{g}_{f^{(k)}_{h+1}}(s_h^k, a_h^k), \tag{69}$$

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

$$\mathbb{E}\left[Z_h^{(k)}(\underline{g}_f, f)\middle|\mathcal{H}_h^{(k)}\right] = \left[\mathcal{T}^\sigma_{\underline{g}_{f^{(k)}_{h+1}},h} f^{(k)}_{h+1}\right](s_h^k, a_h^k) - r_h^{(k)}(s_h^k, a_h^k). \tag{70}$$

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

$$\begin{aligned} X_h^{(k)}(f, \underline{g}_f) &:= \left(f_h^{(k)}(s_h^k, a_h^k) - r_h^{(k)}(s_h^k, a_h^k) - Z_h^{(k)}(f^{(k)}, \underline{g}_{f^{(k)}})\right)^2 \\ &\quad - \left(\left[\mathcal{T}^\sigma_{\underline{g}_{f^{(k)}_{h+1}},h} f^{(k)}_{h+1}\right](s_h^k, a_h^k) - r_h^{(k)}(s_h^k, a_h^k) + Z_h^{(k)}(f^{(k)}, \underline{g}_{f^{(k)}})\right)^2, \end{aligned} \tag{71}$$

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

$$\begin{aligned} \mathbb{E}\left[X_h^{(k)}(f, \underline{g}_f)\middle|\mathcal{H}_h^{(k)}\right] &= \left(\delta_h^{(k)}(s_h^k, a_h^k)\right)^2 \\ \mathrm{Var}\left[X_h^{(k)}(f, \underline{g}_f)\middle|\mathcal{H}_h^{(k)}\right] &\leq c_2\Big(H\min\{H, 1/\sigma\}\Big)^2 \mathbb{E}\left[X_h^{(k)}(f, \underline{g}_f)\middle|\mathcal{H}_h^{(k)}\right], \end{aligned} \tag{72}$$

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

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

$$\left|\sum_{k=1}^K \left(X_h^{(k)}(f, \underline{g}_f) - \mathbb{E}\left[X_h^{(k)}(f, \underline{g}_f)\right]\right)\middle|\mathcal{H}_h^{(k)}\right| \leq \mathcal{O}\left(\sqrt{\log(1/\delta)\sum_{k=1}^K \mathbb{E}\left[X_h^{(k)}(f, \underline{g}_f)\middle|\mathcal{H}_h^{(k)}\right]} + \log(1/\delta)\right). \tag{73}$$

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

$$\sum_{t<k} \mathbb{E}\Big[\Big(\delta_h^{(t)}(s_h, a_h)\Big)^2 \Big| \mathcal{H}_h^{(t)}\Big] \leq \mathcal{O}(\beta), \tag{74}$$

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

Therefore, eq. 74 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

$$\begin{aligned}
\sum_{(s,a)} \widetilde{d}_h^{(k)}(s,a)\, \delta_h^{(t)}(s,a)^2 &\overset{(a)}{=} \sum_{t<k} \sum_{(s,a)} d_h^{(t),P^\omega}(s,a)\, \delta_h^{(t)}(s,a)^2 \\
&= \sum_{t<k} \mathbb{E}_{(s,a)\sim d_h^{(t),P^\omega}} \left[\delta_h^{(t)}(s,a)^2\right] \\
&\overset{(b)}{\leq} \mathcal{O}(\beta),
\end{aligned} \tag{75}$$

where (a) is by the definition of $\widetilde{d}_h^{(k)}(s,a)$ given by equation 31, and (b) is using equation 74.

☞ *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

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

As in eq. 72, $\mathbb{E}\Big[W_h^{(t)}(f, \underline{g}_f) \mid \mathcal{H}_h^{(t)}\Big] = \Big(f_h^{(t)}(s_h^t, a_h^t) - Q_h^{\star,\sigma}(s_h^t, a_h^t)\Big)^2$ 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 $|W_h^{(t)}(f, \underline{g}_f)| \leq c_1\Big(H \min\{H, 1/\sigma\}\Big)^2$ and $\mathrm{Var}\Big[W_h^{(t)}(f, \underline{g}_f) \mid \mathcal{H}_h^{(t)}\Big] \leq c_2\Big(H \min\{H, 1/\sigma\}\Big)^2 E\Big[W_h^{(t)}(f, \underline{g}_f) \mid \mathcal{H}_h^{(t)}\Big]$. 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 $\underline{g}_f$ denote any dual parameter obtained from the empirical optimization in eq. 14 for a given state–action value function f, and let $\mathcal{T}_g^\sigma$ be as defined in eq. 10. Then, under Definition 2, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}} \left\|\mathcal{T}^\sigma f - \mathcal{T}_{\underline{g}_f}^\sigma f\right\|_{1,\mu^\pi} = \mathcal{O}\left(H \min\{H, 1/\sigma\} \sqrt{\frac{2\log(8|\mathcal{G}||\mathcal{F}|/\delta)}{|\mathcal{D}|}} + \xi_{\text{dual}}\right). \tag{76}$$

Proof. Fix an arbitrary $f \in \mathcal{F}$ and recall that $\underline{g}_f$ as defined in eq. 14, where $\widehat{\text{Dual}}_{loss}$ is given in eq. 14. For notational convenience, define the dual objective

$$\Phi_f(x) \ := \ \mathbb{E}_{(s,a)\sim\mu^\pi, s'\sim P^\star_{s,a}}\Big[h_f(x)\Big], \text{ where } h_f(x) := \big(x - \max_{a'} f(s', a')\big)_+ - (1-\sigma)x.$$

Using the dual representation in eq. 13, the difference between the true robust Bellman operator and its empirical counterpart can be written as

$$\left\|\mathcal{T}^\sigma f - \mathcal{T}^\sigma_{\underline{g}_f} f\right\|_{1,\mu^\pi} = \Phi_f\big(\underline{g}_f\big) - \mathbb{E}_{(s,a)\sim\mu^\pi}\left[\inf_{\eta\in[0,2H/\sigma]} h_f(\eta)\right]. \tag{77}$$

Next, we use the functional reformulation, which (by the interchange rule for integral functionals (Rockafellar & Wets, 1998, Theorem 14.60)) (as given in Lemma T.2) states that

$$\mathbb{E}_{(s,a)\sim\mu^\pi}\left[\inf_{\eta\in[0,2H/\sigma]} h_f(\eta)\right] = \inf_{g\in\mathcal{L}^1(\mu^\pi)} \Phi_f(g).$$

Substituting this into eq. 77 gives

$$\begin{aligned}\left\|\mathcal{T}^\sigma f - \mathcal{T}^\sigma_{\underline{g}_f} f\right\|_{1,\mu^\pi} &= \Phi_f(\underline{g}_f) - \inf_{g\in\mathcal{L}^1(\mu^\pi)} \Phi_f(g) \\ &= \big[\Phi_f(\underline{g}_f) - \inf_{g\in\mathcal{G}} \Phi_f(g)\big] + \big[\inf_{g\in\mathcal{G}} \Phi_f(g) - \inf_{g\in\mathcal{L}^1(\mu^\pi)} \Phi_f(g)\big].\end{aligned}$$

The second bracket is controlled by the approximate dual realizability assumption (Assumption 3), which gives

$$\inf_{g\in\mathcal{G}} \Phi_f(g) - \inf_{g\in\mathcal{L}^1(\mu^\pi)} \Phi_f(g) \ \le \ \xi_{\text{dual}}.$$

Hence,

$$\left\|\mathcal{T}^\sigma f - \mathcal{T}^\sigma_{\underline{g}_f} f\right\|_{1,\mu^\pi} \ \le \ \Phi_f(\underline{g}_f) - \inf_{g\in\mathcal{G}} \Phi_f(g) + \xi_{\text{dual}}. \tag{78}$$

We now bound the optimization error term $\Phi_f(\underline{g}_f) - \inf_{g\in\mathcal{G}} \Phi_f(g)$. Consider the loss function as

$$\ell_f(g, (s,a,s')) := \big(g(s,a) - \max_{a'} f(s',a')\big)_+ - (1-\sigma)g(s,a),$$

so that $\Phi_f(g) = \mathbb{E}_{(s,a,s')}\big[\ell_f(g,(s,a,s'))\big]$ and $\widehat{\text{Dual}}_{loss}(g; f)$ in eq. 14 is the empirical average of ℓ_f over $\mathcal{D}$. Since $f \in \mathcal{F}$ and $g \in \mathcal{G}$ take values in $[0, H]$ and $[0, 2H/\sigma]$, respectively, we have $|\ell_f(g,(s,a,s'))| \le 5H/\sigma$, and $\ell_f(\cdot,(s,a,s'))$ is $(2-\sigma)$-Lipschitz in g.

By applying the empirical risk minimization generalization bound ((Panaganti et al., 2022, Lemma 3)) together with the Lipschitz-based bound in eq. 81 of Lemma T.1, we obtain that, with probability at least $1-\delta$,

$$\Phi_f(\underline{g}_f) - \inf_{g\in\mathcal{G}} \Phi_f(g) \ \le \ \frac{4H(2-\sigma)}{\sigma}\sqrt{\frac{2\log|\mathcal{G}|}{|\mathcal{D}|}} + \frac{25H}{\sigma}\sqrt{\frac{2\log(8/\delta)}{|\mathcal{D}|}}. \tag{79}$$

Combining equation 78 and equation 79, and then taking a union bound over $f \in \mathcal{F}$, we conclude that, with probability at least $1-\delta$,

$$\begin{aligned}\sup_{f\in\mathcal{F}} \left\|\mathcal{T}^\sigma f - \mathcal{T}^\sigma_{\underline{g}_f} f\right\|_{1,\mu^\pi} &\le 25(3-\sigma)\frac{H}{\sigma}\sqrt{\frac{2\log\big(8|\mathcal{G}||\mathcal{F}|/\delta\big)}{|\mathcal{D}|}} + \xi_{\text{dual}} \\ &\le C\,\frac{H}{\sigma}\sqrt{\frac{2\log\big(8|\mathcal{G}||\mathcal{F}|/\delta\big)}{|\mathcal{D}|}} + \xi_{\text{dual}},\end{aligned}$$

for some absolute constant $C > 0$, which proves the claimed big-$\mathcal{O}$ bound. □

Lemma K.4. *For any policy π and transition kernel P, define the step-h state and state–action visitation measures as*

$$\rho_h^{\pi,P}(s) := \Pr(s_h = s \mid \pi, P), \qquad d_h^{\pi,P}(s,a) := \Pr(s_h = s, a_h = a \mid \pi, P) = \rho_h^{\pi,P}(s)\,\pi(a \mid s).$$

For each step $h \in [H]$, define the robust coefficient as

$$C_h^{\mathrm{cv,rob}} := \sup_{P\in\mathcal{U}} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sup_\pi d_h^{\pi,P}(s,a),$$

and its state-only counterpart

$$\widetilde{C}_h^{\mathrm{cv,rob}} := \sup_{P\in\mathcal{U}} \sum_{s\in\mathcal{S}} \sup_\pi \rho_h^{\pi,P}(s).$$

Under the d-rectangular linear TV-RMDP assumption in Assumption 4 and the definition of C_{rcov} in Definition 3, the robust coverability coefficient satisfies

$$C_{\mathrm{rcov}} \le \max_h \widetilde{C}_h^{\mathrm{cv,rob}} \le \mathcal{O}(Ad).$$

Proof. The first inequality is straightforward, as $P^w \in \mathcal{U}$.

Fix any kernel $P \in \mathcal{U}$. By Lemma T.6,

$$\sum_{(s,a)} \sup_\pi d_h^{\pi,P}(s,a) \le \sum_{(s,a)} \sup_\pi \rho_h^{\pi,P}(s) = A \sum_s \sup_\pi \rho_h^{\pi,P}(s).$$

Applying Lemma T.7 pointwise in s yields

$$\sum_s \sup_\pi \rho_h^{\pi,P}(s) \le \sum_s \max_{i\in[d]} \nu_{h-1,i}(s),$$

for the corresponding signed measures $\{\nu_{h-1,i}\}$. It hence completes the proof by applying Lemma T.8. □

D.3 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 (Panaganti et al., 2022), Lemma 3)**.** *Let P be a distribution on $\mathcal{X}$ and let $\mathcal{H}$ be a hypothesis class of real-valued functions on $\mathcal{X}$. Assume the loss $loss : \mathcal{H}\times\mathcal{X} \to \mathbb{R}$ satisfies*

$$|loss(h,x)| \le c_0, \quad \forall\, h \in \mathcal{H},\ x \in \mathcal{X}, \quad \text{for some constant } c_0 > 0.$$

Given an i.i.d. sample $\mathcal{D} = \{X_i\}_{i=1}^N$ from P, define the empirical risk minimizer $\widetilde{h} \in \arg\min_{h\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^N loss(h, X_i)$. For any $\delta \in (0,1)$ and any population risk minimizer $h^\star \in \arg\min_{h\in\mathcal{H}} \mathbb{E}_{X\sim P}[loss(h,X)]$, the following holds with probability at least $1-\delta$:

$$\mathbb{E}_{X\sim P}[loss(\widetilde{h},X)] - \mathbb{E}_{X\sim P}[loss(h^\star,X)] \le 2R(loss\circ\mathcal{H}\circ\mathcal{D}) + 5c_0\sqrt{\frac{2\log(8/\delta)}{N}}, \tag{80}$$

where $R(loss\circ\mathcal{H}\circ\mathcal{D})$ is the empirical Rademacher complexity of the loss-composed class $loss\circ\mathcal{H}$, defined by

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

with $\{\sigma_i\}_{i=1}^N$ independent of $\{X_i\}_{i=1}^N$ and i.i.d. according to a Rademacher random variable σ (i.e., $\mathbb{P}(\sigma=1) = \mathbb{P}(\sigma=-1) = 0.5$). Moreover, if $\mathcal{H}$ is finite, $|\mathcal{H}| < \infty$, and there exist constants $c_1, c_2 > 0$ such that

$$|h(x)| \le c_0 \quad \forall\, h \in \mathcal{H},\ x \in \mathcal{X}, \qquad \text{and} \qquad loss(h,x) \text{ is } c_1\text{-Lipschitz in } h,$$

then with probability at least $1-\delta$ we further have

$$\mathbb{E}_{X\sim P}[loss(\widetilde{h},X)] - \mathbb{E}_{X\sim P}[loss(h^\star,X)] \le 2c_1c_2\sqrt{\frac{2\log(|\mathcal{H}|)}{N}} + 5c_0\sqrt{\frac{2\log(8/\delta)}{N}}. \tag{81}$$

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 7. *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*

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

and hence $C_{\mathrm{rcov}} \leq SA$.

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

$$\Pr\left(\sum_{t=1}^T M_t \geq \sqrt{2S_T \ln(1/\delta)} + \tfrac{b}{3} \ln(1/\delta) \right) \leq \delta.$$

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 $d^{(t)}(z)/\mu(z) \leq C$ for all $(z,t) \in \mathcal{Z} \times [K]$. Then for all $z \in \mathcal{Z}$,*

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

Lemma T.6. *For any kernel P, state s, action a, and step h,*

$$\sup_\pi d_h^{\pi,P}(s,a) \leq \sup_\pi \rho_h^{\pi,P}(s).$$

Proof. For any π, $d_h^{\pi,P}(s,a) = \rho_h^{\pi,P}(s)\,\pi(a \mid s) \leq \rho_h^{\pi,P}(s)$, hence $\sup_\pi d_h^{\pi,P}(s,a) \leq \sup_\pi \rho_h^{\pi,P}(s)$. □

Lemma T.7. *Fix step h and a kernel P with bases $\{\nu_{h-1,i}\}_{i=1}^d$. For any policy π,*

$$\rho_h^{\pi,P}(\cdot) \;=\; \sum_{i=1}^d z_{h-1,i}^\pi\, \nu_{h-1,i}(\cdot) \quad \textit{for some } z_{h-1}^\pi \in \Delta_d. \tag{82}$$

Consequently, for any state s,

$$\sup_\pi \rho_h^{\pi,P}(s) \;\leq\; \max_{i \in [d]} \nu_{h-1,i}(s). \tag{83}$$

Proof. It holds that

$$\begin{aligned}\rho_h^{\pi,P}(\cdot) &= \sum_{s',a} \rho_{h-1}^{\pi,P}(s')\,\pi(a \mid s')\,P_{h-1}(\cdot \mid s', a) \\ &= \sum_{s',a} \rho_{h-1}^{\pi,P}(s')\,\pi(a \mid s') \sum_{i=1}^{d} \phi_i(s', a)\,\nu_{h-1,i}(\cdot) \\ &= \sum_{i=1}^{d} \underbrace{\Big(\sum_{s',a} \rho_{h-1}^{\pi,P}(s')\,\pi(a \mid s')\,\phi_i(s', a) \Big)}_{=: z_{h-1,i}^{\pi}} \nu_{h-1,i}(\cdot).\end{aligned}$$

Because $\phi_i(\cdot) \geq 0$ and $\sum_{i=1}^d \phi_i(\cdot) = 1$, we have $z_{h-1}^{\pi} \geq 0$ and $\sum_{i=1}^d z_{h-1,i}^{\pi} = 1$, hence $z_{h-1}^{\pi} \in \Delta_d$ and equation 82 holds. For any state s,

$$\rho_h^{\pi,P}(s) = \sum_{i=1}^{d} z_{h-1,i}^{\pi}\,\nu_{h-1,i}(s) \leq \max_i \mu_{h-1,i}(s),$$

and taking $\sup_\pi$ gives equation 83. □

Lemma T.8. *Let $\{\nu_i\}_{i=1}^d$ be some probability measure on $\mathcal{S}$ with in $\mathcal{P}_i$. Then*

$$\sum_{s \in \mathcal{S}} \max_{i \in [d]} \nu_i(s) \;\leq\; d. \tag{84}$$

Proof. Note that

$$\sum_{s \in \mathcal{S}} \max_{i \in [d]} \nu_i(s) \leq \sum_{i=1}^{d} \sum_{s} \nu_i(s) = d,$$

where the last equality is from the fact that ν_i is some probability measure. □

E 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.