# Distributionally Robust Off-Dynamics Reinforcement Learning: Provable Efficiency with Linear Function Approximation

**Zhishuai Liu**
Duke University
zhishuai.liu@duke.edu

**Pan Xu**
Duke University
pan.xu@duke.edu

## Abstract

We study off-dynamics Reinforcement Learning (RL), where the policy is trained on a source domain and deployed to a distinct target domain. We aim to solve this problem via online distributionally robust Markov decision processes (DRMDPs), where the learning algorithm actively interacts with the source domain while seeking the optimal performance under the worst possible dynamics that is within an uncertainty set of the source domain's transition kernel. We provide the first study on online DRMDPs with function approximation for off-dynamics RL. We find that DRMDPs' dual formulation can induce nonlinearity, even when the nominal transition kernel is linear, leading to error propagation. By designing a $d$-rectangular uncertainty set using the total variation distance, we remove this additional nonlinearity and bypass the error propagation. We then introduce DR-LSVI-UCB, the first provably efficient online DRMDP algorithm for off-dynamics RL with function approximation, and establish a polynomial suboptimality bound that is independent of the state and action space sizes. Our work makes the first step towards a deeper understanding of the provable efficiency of online DRMDPs with linear function approximation. Finally, we substantiate the performance and robustness of DR-LSVI-UCB through different numerical experiments.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. 

## 1 INTRODUCTION

The Markov decision process (MDP) is a prevalent model in dynamic decision-making and reinforcement learning (Puterman, 2014; Sutton and Barto, 2018). A central challenge in employing MDPs in various applications lies in the lack of knowledge of model parameters, notably the transition kernels. Existing studies mostly hinge on the assumption that the environment in which a policy is trained is identical to that in which it is deployed. However, in practical scenarios where this assumption is violated, standard RL methods are prone to severe failures (Farebrother et al., 2018; Packer et al., 2018; Zhao et al., 2020), a phenomenon known as the *sim-to-real gap*. Infectious disease control (Laber et al., 2018; Liu et al., 2023a) exemplifies such a case wherein an agent trains policies on simulators extensively utilized in environmental studies. Nonetheless, these simulators cannot fully capture the environmental evolution complexity, and environmental changes may also occur over time, further contributing to the sim-to-real gap. Another instance is found in robotics learning, where slight variations between training and testing environments, such as terrain or target parameters, may lead to task failure (Maitin-Shepard et al., 2010; Tobin et al., 2017; Peng et al., 2018).

Learning under the sim-to-real gap can be conceptualized as an off-dynamics RL problem (Koos et al., 2012; Wulfmeier et al., 2017; Eysenbach et al., 2020; Jiang et al., 2021), where an agent trains a policy in an accessible source domain, such as a simulator or the present environment, then deploys the learned policy in a distinct target domain, which could be the real environment the agent encounters during operation or a future changing environment. The dynamics shift between environments necessitates a robust strategy for policy learning in the source domain, ensuring that the policy can work effectively in different yet structurally similar target domains.

Distributionally robust Markov decision process (DR-

MDPs) (Satia and Lave Jr, 1973; Nilim and El Ghaoui, 2005; Iyengar, 2005) address the sim-to-real gap challenge by modeling the uncertainty of transition kernels. It aims to learn a robust policy that performs well under the worst-case transition kernel within the uncertainty set defined based on the source environment (Xu and Mannor, 2006; Wiesemann et al., 2013; Zhang et al., 2021; Yang et al., 2022; Panaganti et al., 2022; Shi and Chi, 2022; Yang et al., 2023b; Shen et al., 2024). Existing DRMDP research can be categorized based on the assumption on the source domain: (i) planning problems where the exact model is assumed known, (ii) learning under a generative model, and (iii) learning from offline datasets utilizing specific data coverage assumptions. However, in practice, formulating and solving a planning problem is often infeasible due to imperfect knowledge or complexity of the source domain. Similarly, an accurate generative model representing the source domain is usually unavailable. Additionally, most data coverage assumptions require the datasets have sufficient coverage of distributions induced by the optimal policy under any transition kernel in the uncertainty set. Since the optimal policy is usually unknown and there are infinite number of transition kernels in the uncertainty set, practical verification of data coverage assumptions is intractable. Thus, when incremental collection of data through active interactions with the source domain is feasible, online algorithms without relying on additional oracles or data coverage assumptions about the optimal policy will be preferred. We refer to this as the online DRMDP problem.

Another significant challenge in RL is the ubiquitous presence of applications with arbitrarily large state and action spaces, which require suitable function approximations to alleviate the curse of dimensionality. Although approaches based on linear function approximation have exhibited theoretical and empirical success in numerous settings under standard MDP (Bhandari et al., 2018; Modi et al., 2020; Jin et al., 2020; He et al., 2023, 2021; Yang and Wang, 2020), DRMDP encounters additional difficulties when combined with linear function approximations since the dual formulation in worst-case analyses may induce extra nonlinearity, even when the source domain transition kernel is linear (Tamar et al., 2014; Pinto et al., 2017; Derman et al., 2018; Mankowitz et al., 2019; Derman et al., 2020; Zhang et al., 2021; Badrinath and Kalathil, 2021). Consequently, the theoretical understanding of online DRMDPs with function approximation remains elusive, even when the approximation is linear. This leads to the open question:

*When is it possible to design a provably efficient algorithm for online DRMDPs with linear function approximation?*

In this work, we provide the first analysis of online DRMDP with linear function approximation where an agent actively interacts with the source domain to learn a robust policy.

**Our main contributions** are summarized as follows.

- We first investigate the differences in applying linear function approximation in DRMDPs with uncertainty sets defined on different probability divergence metrics. We show that the strong duality for Chi-square or Kullback-Leibler (KL) based DRMDPs induces additional nonlinearity which can cause severe error amplification and regret accumulation (see Remark 4.4 for more details). We then identify a feasible setting that assumes a $d$-rectangular linear DRMDP and a total variation (TV) based uncertainty set, which permits linear representations on the robust Q-functions, and bypasses the error amplification and regret accumulation.
- We introduce a model-free online algorithm, viz., DR-LSVI-UCB, based on the LSVI-UCB algorithm in the non-robust setting (Jin et al., 2020). The design of the DR-LSVI-UCB incorporates a robust Upper Confidence Bonus (UCB) quantity and a truncated estimation of the robust state-action value function at the MDP's fail state, both of which are explicitly devised for the online DRMDP setting (refer to Remark 4.5 for more details).
- We prove an average suboptimality bound for DR-LSVI-UCB in the order of $\tilde{O}(\sqrt{H^4d^4/K})$, where $H$ is the horizon length, $d$ the feature dimension, and $K$ the number of episodes. Our result matches the average regret[1] bound of its non-robust counterpart LSVI-UCB (Jin et al., 2020) regarding $H$ and $K$, but is worse regarding feature dimension by a factor of $\sqrt{d}$. To the best of our knowledge, this is the first non-asymptotic suboptimality bound for online DRMDPs with linear function approximation, which guarantees efficient robust learning in off-dynamics RL. Interestingly, when reduced to the tabular setting where $d = SA$ with $S$ and $A$ being the state and action space sizes, the average suboptimality gap of DR-LSVI-UCB exactly matches the average regret bound of LSVI-UCB, indicating tabular DRMDPs with a TV uncertainty set might not be more challenging than the standard tabular MDP.
- We perform numerical experiments to illustrate the efficacy of DR-LSVI-UCB on a simulated linear MDP environment and an emulated American put option environment (Tamar et al., 2014). Our

[1]Since in DRMDP, we trade off the performance in the source domain for the robustness in the target domain, we evaluate a robust algorithm by its suboptimality gap from the optimal robust policy, comparable to the average regret in standard MDP, i.e., the cumulative regret divided by $K$.

results demonstrate that the policies derived by DR-LSVI-UCB are robust against dynamics shifts, further substantiating our theoretical findings.

## 2 RELATED WORK

**Episodic Linear MDP** Our study focuses on the episodic linear MDP setting. Specifically, we assume the nominal transition probability in our DRMDP admits the linear MDP structure. There has been a recent surge in research on episodic linear MDPs (Yang and Wang, 2020; Jin et al., 2020; Modi et al., 2020; Zanette et al., 2020; Wang et al., 2020a; He et al., 2021; Wagenmaker et al., 2022; Ishfaq et al., 2023; He et al., 2023). The most relevant study to ours is the seminal work of Jin et al. (2020), which introduced a model-free online algorithm, LSVI-UCB, for standard RL. Through a 'Hoeffding-type' exploration bonus, LSVI-UCB can actively explore the nominal environment and achieves a $\tilde{O}(\sqrt{d^3H^4K})$ regret bound. However, the episodic linear MDP setting still remains understudied in the context of DRMDPs.

**DRMDPs** Numerous works have extensively studied the DRMDP framework under different settings. Xu and Mannor (2006); Wiesemann et al. (2013); Yu and Xu (2015); Mannor et al. (2016); Goyal and Grand-Clement (2023) studied the DRMDP assuming the exact environment is known, and establishing DRMDPs as classic planning problems. Zhou et al. (2021); Yang et al. (2022); Panaganti and Kalathil (2022); Xu et al. (2023); Shi et al. (2023); Yang et al. (2023a) studied the DRMDP assuming the access to a generative model. Panaganti et al. (2022); Shi and Chi (2022); Blanchet et al. (2023) studied the DRMDP in the offline RL setting assuming strong data coverage or concentratability conditions. Moreover, Dong et al. (2022) studied the online DRMDP under the episodic tabular MDP setting. They proposed a model-based algorithm ROPO, which achieves an average suboptimality bound of $\tilde{O}(\sqrt{H^4S^2A/K})$ under the $(s,a)$-rectangular assumption. However, their method cannot deal with settings where state space size $S$ and action space size $A$ are large or infinite in practical applications.

**DRMDPs with linear function approximation** Tamar et al. (2014) first proposed to use linear function approximation to solve DRMDPs with large state and action spaces, and provided an asymptotic convergence guarantee for their sampling-based approach. Badrinath and Kalathil (2021) proposed a model-free online algorithm based on linear projection, and provided the corresponding asymptotic convergence guarantee. Recently, Ma et al. (2022) pointed out that the nonlinearity of DRMDPs might make linear projection fall short, resulting in poor decision-making. Ma et al. (2022) then studied the novel $d$-rectangular linear DRMDP that naturally admits linear representations of the robust state-action value function. They studied the offline setting and proposed two value iteration based algorithms under the uniformly well-explored dataset assumption and the sufficient coverage of the optimal policy assumption, respectively. Blanchet et al. (2023) also studied the offline $d$-rectangular linear DRMDP based on the robust partial coverage assumption. However, the data coverage assumptions cannot be verified and guaranteed in practice as we discussed in Remark 5.4. Thus, an online algorithm, which automates the acquisition of the optimal robust policy through actively interacting with the source domain, for the episodic $d$-rectangular linear DRMDP is in need.

## 3 DISTRIBUTIONALLY ROBUST MDP WITH LINEAR FUNCTION APPROXIMATION

### 3.1 Preliminaries

A finite horizon Markov decision process can be denoted as MDP$(\mathcal{S}, \mathcal{A}, H, P, r)$. Here $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $H \in \mathbb{Z}_+$ is the horizon length, $P = \{P_h\}_{h=1}^H$ and $r = \{r_h\}_{h=1}^H$ are the set of transition kernels and reward functions, respectively. For each step $h \in [H]$, we denote $P_h(\cdot|s,a)$ as the transition probability measure over the next state if action $a$ is taken at state $s$, and $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the deterministic reward function, which for simplicity is assumed to be known.

A non-stationary Markov policy $\pi = \{\pi_h\}_{h=1}^H$ is a sequence of decision rules, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ is the policy at step $h$ and $\Delta(\mathcal{A})$ is the probability simplex defined over the action space $\mathcal{A}$. For any transition kernel $P$ and any policy $\pi$, we define the value function and the state-action value function (viz., the Q-function) at step $h$ as

$$V_h^{\pi,P}(s) := \mathbb{E}^P\big[\textstyle\sum_{t=h}^H r_t(s_t,a_t)\big|s_h = s, \pi\big],$$
$$Q_h^{\pi,P}(s,a) := \mathbb{E}^P\big[\textstyle\sum_{t=h}^H r_t(s_t,a_t)\big|s_h = s, a_h = a, \pi\big].$$

As the rewards are bounded in $[0,1]$, thus any value function and Q-function are bounded in $[0,H]$.

A finite horizon distributionally robust Markov decision process (DRMDP) is formally defined by a tuple DRMDP$(\mathcal{S}, \mathcal{A}, H, \mathcal{U}^\rho(P^0), r)$. Here, $P^0 = \{P_h^0\}_{h=1}^H$ is the set of nominal transition kernels, and $\mathcal{U}^\rho(P^0) = \bigotimes_{h\in[H]} \mathcal{U}_h^\rho(P_h^0)$ denotes an uncertainty set centered around the nominal transition kernel with an uncertainty level $\rho \geq 0$. $\mathcal{U}_h^\rho(P_h^0)$ is often defined as a ball centered around $P_h^0$ with radius $\rho$ based on different probability divergence measures (Iyengar, 2005; Yang et al., 2022; Xu et al., 2023).

In contrast with the standard MDP where only the nominal transition kernel $P^0$ is considered, in DRMDPs, we consider all transition kernels within the uncertainty set $\mathcal{U}^\rho(P^0)$. Then for $h \in [H]$ and any policy $\pi$, we define the robust value function $V_h^{\pi,\rho} : \mathcal{S} \to \mathbb{R}$ as the value function under the worst possible transition kernel within the uncertainty set:

$$V_h^{\pi,\rho}(s) = \inf_{P \in \mathcal{U}^\rho(P^0)} V_h^{\pi,P}(s), \quad \forall (h,s) \in [H] \times \mathcal{S}.$$

Accordingly, we define the robust state-action value function as $Q_h^{\pi,\rho}(s,a) = \inf_{P \in \mathcal{U}^\rho(P^0)} Q_h^{\pi,P}(s,a)$, for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$.

We then define the optimal robust value function and optimal robust state-action value function: $\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, $V_h^{\star,\rho}(s) = \sup_{\pi \in \Pi} V_h^{\pi,\rho}(s)$, $Q_h^{\star,\rho}(s,a) = \sup_{\pi \in \Pi} Q_h^{\pi,\rho}(s,a)$, where $\Pi$ is the set of all (possibly randomized and nonstationary) policies. Then the optimal robust policy $\pi^\star = \{\pi_h^\star\}_{h=1}^H$, defined as the policy that achieves the optimal robust value function, is given by $\pi^\star = \arg\sup_{\pi \in \Pi} V_h^{\pi,\rho}(s)$, for any $(h,s) \in [H] \times \mathcal{S}$. Our goal is to learn the optimal robust policy by actively interacting with the nominal environment within $K$ episodes. At the beginning of episode $k$, the agent receives an initial state $s_1^k$. Denote $\pi^k$ as the current policy of the agent. We use $V_1^{\star,\rho}(s_1^k) - V_1^{\pi^k,\rho}(s_1^k)$ to measure the suboptimality of policy $\pi^k$ at episode $k$. Hence, we are interested in the average suboptimality of an algorithm after $K$ episodes, i.e., $\mathrm{AveSubopt}(K)$, defined as follows

$$\mathrm{AveSubopt}(K) = \frac{1}{K} \sum_{k=1}^K \left[V_1^{\star,\rho}(s_1^k) - V_1^{\pi^k,\rho}(s_1^k)\right].$$

### 3.2 $d$-Rectangular Linear DRMDP

In this paper, we define the uncertainty set $\mathcal{U}_h^\rho(P_h^0)$ based on a linear structure of the nominal transition kernel $P_h^0$, called the linear MDP (Jin et al., 2020; Wei et al., 2021; Wagenmaker et al., 2022; He et al., 2023).

**Assumption 3.1.** (Linear MDP) Given a known state-action feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ satisfying $\sum_{i=1}^d \phi_i(s,a) = 1$, $\phi_i(s,a) \geq 0$, for any $(i,s,a) \in [d] \times \mathcal{S} \times \mathcal{A}$, we assume the reward function $\{r_h\}_{h=1}^H$ and nominal transition kernels $\{P_h^0\}_{h=1}^H$ have linear structures. Specifically, for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, $r_h(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h \rangle$, and $P_h^0(\cdot|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h^0(\cdot) \rangle$, where $\{\boldsymbol{\theta}_h\}_{h=1}^H$ are known vectors with bounded norm $\|\boldsymbol{\theta}_h\|_2 \leq \sqrt{d}$ and $\{\boldsymbol{\mu}_h\}_{h=1}^H$ are unknown probability measures over $\mathcal{S}$.

Assumption 3.1 is slightly stronger than the linear MDP studied in the standard RL literature. Following similar works in DRMDPs (Ma et al., 2022; Blanchet et al., 2023), we assume the coordinates of the feature mapping $\boldsymbol{\phi}(\cdot,\cdot)$ to be positive and add up to one, which could be achieved by normalization. Meanwhile, the factor measures $\{\boldsymbol{\mu}_h\}_{h=1}^H$ are required to be proper probability measures. Under these additional constraints, the nominal transition kernel $P_h^0(\cdot|s,a)$ can be seen as a mixture of factor distributions $\boldsymbol{\mu}_h(\cdot)$ with the aggregated feature $\phi(s,a)$ determining the weights.

To incorporate the linear structure of $P^0$ into the uncertainty set $\mathcal{U}_h^\rho(P_h^0)$, we adopt the notion of $d$-rectangular uncertainty set (Ma et al., 2022; Goyal and Grand-Clement, 2023). More specifically, we assume $\mathcal{U}^\rho(P^0)$ is parameterized by $\{\boldsymbol{\mu}_h^0\}_{h=1}^H$ and can be decomposed into $\mathcal{U}_h^\rho(P_h^0) = \bigotimes_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)$, where $\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0) = \{\sum_{i=1}^d \phi_i(s,a)\mu_{h,i}(\cdot) : \mu_{h,i}(\cdot) \in \mathcal{U}_{h,i}^\rho(\mu_{h,i}^0), \forall i \in [d]\}$, and $\mathcal{U}_{h,i}^\rho(\mu_{h,i}^0)$ is defined as

$$\mathcal{U}_{h,i}^\rho(\mu_{h,i}^0) = \left\{\mu : \mu \in \Delta(\mathcal{S}), D(\mu||\mu_{h,i}^0) \leq \rho\right\}. \tag{3.1}$$

Here $D(\cdot||\cdot)$ is a probability divergence metric that will be instantiated later. We remark that the factor uncertainty sets $\{\mathcal{U}_{h,i}^\rho(\mu_{h,i}^0)\}_{i \in [d]}$ are independent of the state-action pair $(s,a)$, and also independent with each other. As we will show in the proof of Proposition 4.3, these attributes are essential in deriving that, for all policies, the robust Q-functions are always linear in the feature mapping $\boldsymbol{\phi}(\cdot,\cdot)$.

### 3.3 Robust Bellman Equation and the Optimal Policy in DRMDPs

We show that the robust value function and the robust Q-function defined in DRMDPs satisfy the following robust Bellman equation. We denote $[\mathbb{P}_h V](s,a) = \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V(s')]$ for simplicity.

**Proposition 3.2.** (Robust Bellman equation) Under the $d$-rectangular linear DRMDP setting, for any nominal transition kernel $P^0 = \{P_h^0\}_{h=1}^H$ and any stationary policy $\pi = \{\pi_h\}_{h=1}^H$, the following robust Bellman equation holds: for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} Q_h^{\pi,\rho}(s,a) &= r_h(s,a) + \inf_{P_h(\cdot|s,a) \in \mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} [\mathbb{P}_h V_{h+1}^{\pi,\rho}](s,a) \\ V_h^{\pi,\rho}(s) &= \mathbb{E}_{a \sim \pi_h(\cdot|s)}\left[Q_h^{\pi,\rho}(s,a)\right]. \end{aligned} \tag{3.2}$$

Furthermore, it is well-known that the optimal (robust) value function can be achieved by a deterministic and stationary policy in standard MDPs (Sutton and Barto, 2018; Agarwal et al., 2019) and tabular DRMDPs with $(s,a)$-rectangular assumption (Iyengar, 2005; Nilim and El Ghaoui, 2005). Similarly, we show that the optimal robust value function and Q-function can be achieved by a deterministic and stationary policy $\pi^\star$ in the $d$-rectangular linear DRMDP.

**Proposition 3.3.** (Existence of the optimal policy) Assume the nominal transition kernel $P^0$ satisfies Assumption 3.1 and the uncertainty set $\mathcal{U}^\rho(P^0)$ is defined as in Section 3.2. Then there exists a deterministic and stationary policy $\pi^\star$ such that $V_h^{\pi^\star,\rho}(s) = V_h^{\star,\rho}(s)$ and $Q_h^{\pi^\star,\rho}(s,a) = Q_h^{\star,\rho}(s,a)$, for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$.

The results in Propositions 3.2 and 3.3 have been used in existing analyses of DRMDPs without proof (Ma

et al., 2022; Blanchet et al., 2023). For completeness, we provide their proofs in Appendix B. With these results, we can safely restrict the policy class $\Pi$ to the deterministic and stationary one. This leads to the robust Bellman optimality equation:

$$Q_h^{\star,\rho}(s,a) = r_h(s,a) + \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} [\mathbb{P}_h V_{h+1}^{\star,\rho}](s,a)$$
$$V_h^{\star,\rho}(s) = \max_{a\in\mathcal{A}} Q_h^\star(s,a). \tag{3.3}$$

(3.3) suggests that the optimal robust policy is greedy with respect to the optimal robust Q-function. Therefore, it suffices to estimate $Q_h^{*,\rho}$ to find $\pi^*$.

### 3.4 DRMDPs with TV divergence

In this work, we focus on the total variation (TV) distance as the probability divergence metric employed in defining the uncertainty set (3.1). Given any two probability distributions $P$ and $Q$, the TV divergence, denoted by $D_{TV}(P\|Q)$, can be expressed as

$$D_{TV}(P\|Q) = 1/2 \int_{\mathcal{S}} |P(s) - Q(s)| ds. \tag{3.4}$$

The optimization problem in (3.2) has the following dual formulation under the TV uncertainty set.

**Proposition 3.4.** (Strong duality for TV (Shi et al., 2023, Lemma 4)). Given any probability measure $\mu^0$ over $\mathcal{S}$, a fixed uncertainty level $\rho$, the uncertainty set $\mathcal{U}^\rho(\mu^0) = \{\mu : \mu \in \Delta(\mathcal{S}), D_{TV}(\mu||\mu^0) \le \rho\}$, and any function $V : \mathcal{S} \to [0,H]$, we obtain

$$\inf_{\mu\in\mathcal{U}^\rho(\mu^0)} \mathbb{E}_{s\sim\mu} V(s) = \max_{\alpha\in[V_{\min},V_{\max}]} \{\mathbb{E}_{s\sim\mu^0}[V(s)]_\alpha$$
$$-\rho(\alpha - \min_{s'}[V(s')]_\alpha)\}, \tag{3.5}$$

where $[V(s)]_\alpha = \min\{V(s),\alpha\}$, $V_{\min} = \min_s V(s)$ and $V_{\max} = \max_s V(s)$. Notably, the range of $\alpha$ can be relaxed to $[0,H]$ without impacting the optimization.

## 4 ROBUST LEAST SQUARE VALUE ITERATION WITH UCB EXPLORATION

### 4.1 Linear Representation of the Robust State-Action Value Function

Recall the strong duality in (3.5), we need to solve the minimization problem, $\min_{s'}[V(s')]_\alpha$, which is challenging when it is not convex with respect to $s'$ and computationally inefficient when $\mathcal{S}$ is large. To overcome this issue, we make the same fail-state assumption made in the function approximation setting (Panaganti et al., 2022) and show that it is compatible with the $d$-rectangular linear DRMDP.

**Assumption 4.1.** (Fail-state) The linear MDP has a 'fail state' $s_f$, such that for all $(h,a) \in [H]\times\mathcal{A}$, $r_h(s_f,a) = 0$, $P_h^0(s_f|s_f,a) = 1$.

The existence of fail states is natural in many real-world applications such as the collapse of a robot in robotics (Panaganti et al., 2022). As another example in the context of cancer treatments, patients could die, or the cancer may advance further, during the course of a finite-stage treatment process (Goldberg and Kosorok, 2012; Zhao et al., 2018; Liu et al., 2023b), both of which could be considered as fail states.

We show that Assumption 4.1 is compatible with the linear MDP structure. In particular, we show that we can extend the original $d$-rectangular linear DRMDP as follows. First, we define a new feature mapping $\tilde{\boldsymbol{\phi}} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{d+1}$ based on the original one:

$$\tilde{\boldsymbol{\phi}}(s_f,a) = [1,0,\cdots,0]^\top, \quad \forall a\in\mathcal{A},$$
$$\tilde{\boldsymbol{\phi}}(s,a) = \big[0,\boldsymbol{\phi}(s,a)^\top\big]^\top, \quad \forall (s,a)\in\mathcal{S}/\{s_f\}\times\mathcal{A}.$$

It is easy to verify that $\sum_{i=1}^{d+1}\tilde{\phi}_i(s,a) = 1, \forall(s,a)\in\mathcal{S}\times\mathcal{A}$, and $\tilde{\phi}_i(s,a) \ge 0$, $\forall i \in [d+1]$. Let $\tilde{\boldsymbol{\theta}}_h = [0,\boldsymbol{\theta}_h^\top]^\top$, and $\tilde{\boldsymbol{\mu}}_h^0(\cdot) = [\delta_{s_f}(\cdot),\boldsymbol{\mu}_h^0(\cdot)^\top]^\top$, where $\delta_{s_f}$ is the Dirac delta distribution with mass at $s_f$. We can show that the reward functions and transition kernels are still linear based on the new notations. Then we can define the same $d$-rectangular uncertainty set as in Section 3.2. For simplicity, we assume the fail-state assumption holds in the original linear MDP in this paper.

**Remark 4.2.** Under Assumption 4.1, Proposition 3.4 can be further simplified. For any function $V : \mathcal{S}\to[0,H]$ with $\min_{s\in\mathcal{S}} V(s) = V(s_f) = 0$, we have $\inf_{\mu\in\mathcal{U}^\rho(\mu^0)}\mathbb{E}_{s\sim\mu}V(s) = \max_{\alpha\in[0,H]}\{\mathbb{E}_{s\sim\mu^0}[V(s)]_\alpha - \rho\alpha\}$. Then with the fail state $s_f$, for any $(\pi,h,a)\in\Pi\times[H]\times\mathcal{A}$, we have $Q_h^{\pi,\rho}(s_f,a) = 0$, and $V_h^{\pi,\rho}(s_f) = 0$.

Now we show that the robust Q-function $Q_h^{\pi,\rho}(\cdot,\cdot)$ is linear in the feature mapping $\boldsymbol{\phi}(\cdot,\cdot)$ for any policy $\pi$.

**Proposition 4.3.** Under Assumptions 3.1 and 4.1, for any $(\pi,s,a,h)\in\Pi\times\mathcal{S}\times\mathcal{A}\times[H]$, the robust Q-function $Q_h^{\pi,\rho}(s,a)$ has a linear form as follows:

$$Q_h^{\pi,\rho}(s,a) = \langle\boldsymbol{\phi}(s,a),\boldsymbol{\theta}_h + \boldsymbol{\nu}_h^{\pi,\rho}\rangle\mathbb{1}\{s\neq s_f\},$$

where $\boldsymbol{\nu}_h^{\pi,\rho} = (\nu_{h,1}^{\pi,\rho},\ldots,\nu_{h,d}^{\pi,\rho})^\top$, $\nu_{h,i}^{\pi,\rho} = \max_{\alpha\in[0,H]}\{z_{h,i}^\pi(\alpha) - \rho\alpha\}$, and $z_{h,i}^\pi(\alpha) = \mathbb{E}^{\mu_{h,i}^0}[V_{h+1}^{\pi,\rho}(s')]_\alpha$.

Therefore, with the known feature mapping $\boldsymbol{\phi}(\cdot,\cdot)$, it suffices to estimate the weight vectors $\{\boldsymbol{\nu}_h^{\pi,\rho}\}_{h=1}^H$ to recover the robust Q-functions. Based on Proposition 4.3, we can iteratively perform backward induction to estimate the robust Q-functions. Specifically, given any estimated robust Q-function at step $h+1$, $Q_{h+1}^k(s,a)$, and estimated robust value function $V_{h+1}^k(s) = \max_{a\in\mathcal{A}} Q_{h+1}^k(s,a)$, the one step backward induction leads to the following linear term

$$\langle\boldsymbol{\phi}(s,a),\boldsymbol{\theta}_h + \boldsymbol{\nu}_h^{\rho,k}\rangle\mathbb{1}\{s\neq s_f\},$$

where $\nu_{h,i}^\rho := \max_{\alpha\in[0,H]}\{z_{h,i}(\alpha) - \rho\alpha\}$ and $z_{h,i}(\alpha) := \mathbb{E}^{\mu_{h,i}^0}[V_{h+1}^k(s')]_\alpha$, for any $i\in[d]$. According to the linear structure defined in Assumption 3.1 on the nominal transition kernel, $z_{h,i}(\alpha)$ is the parameter of the

following linear formulation,

$$\left[\mathbb{P}_h^0\left[V_{h+1}^k\right]_\alpha\right](s,a) = \left\langle \boldsymbol{\phi}(s,a), \boldsymbol{z}_h(\alpha)\right\rangle,$$

which is an expectation with respect to the nominal transition kernel $P_h^0$. Therefore, we can collect trajectories $\{(s_h^\tau, a_h^\tau, s_{h+1}^\tau)\}_{\tau=1}^{k-1}$ and estimate $z_{h,i}(\alpha)$ from samples. In particular, we will solve the following ridge regression problem with regularizer $\lambda > 0$,

$$\begin{aligned}\hat{\boldsymbol{z}}_h(\alpha) = \operatorname{argmin}_{\boldsymbol{z}\in\mathbb{R}^d} \textstyle\sum_{\tau=1}^{k-1} \left(\left[V_{h+1}^k(s_{h+1}^\tau)\right]_\alpha \right.\\ \left. -\boldsymbol{\phi}_h^{\tau\top}\boldsymbol{z}\right)^2 + \lambda\|\boldsymbol{z}\|_2^2,\end{aligned} \tag{4.1}$$

with the close-form solution being

$$\hat{\boldsymbol{z}}_h(\alpha) = \left(\Lambda_h^k\right)^{-1}\left[\textstyle\sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau [V_{h+1}^k(s_{h+1})]_\alpha\right], \tag{4.2}$$

where $\boldsymbol{\phi}_h^\tau$ is a shorthand notation for $\boldsymbol{\phi}(s_h^\tau, a_h^\tau)$, and $\Lambda_h^k = \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau (\boldsymbol{\phi}_h^\tau)^\top + \lambda \mathbf{I}$. We then approximate $\boldsymbol{\nu}_h^{\rho,k}$ by $\hat{\nu}_{h,i}^{\rho,k} = \max_{\alpha\in[0,H]}\{\hat{z}_{h,i}(\alpha) - \rho\alpha\}, i \in [d]$, and obtain the estimated robust Q-function at step $h$:

$$Q_h^k(s,a) = \left\langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h + \hat{\boldsymbol{\nu}}_h^{\rho,k}\right\rangle \mathbb{1}\{s \neq s_f\}. \tag{4.3}$$

**Remark 4.4.** Thanks to the linear representation of the robust Q-function in terms of $\boldsymbol{\phi}(\cdot,\cdot)$ (Proposition 4.3) and the linear dependence on the value function $V(s)$ in strong duality (Proposition 3.4), we can apply ridge regression with the estimated value function $V_{h+1}^k(s')$ as the target.

In comparison, the strong duality under KL uncertainty set (Shi and Chi, 2022) is $\inf_{\mu\in\mathcal{U}^\rho(\mu^0)} \mathbb{E}_{s\sim\mu} V(s) = \max_{\alpha\in[0,H/\rho]}\{-\alpha \log \mathbb{E}_{s\sim\mu^0}[e^{-V(s)/\alpha}] - \alpha\rho\}$. Since the expectation is nonlinear in the value function $V(s)$, we have to apply ridge regression with $\exp(-V(s)/\alpha)$ as the target, and take logarithm back to the approximator (see (8) - (10) of Ma et al. (2022) for details). This logarithm operation could amplify the approximation error by $\exp(H)$, which leads to the $O(\exp(H/\underline{\beta}))$ term in Theorem 4.1 of Ma et al. (2022). This amplified error could accumulate through the backward induction and ultimately lead to an $O(\exp(H^2))$ term in the regret bound of online DR-MDPs. Similar argument applies to the Chi-square divergence based uncertainty set, with the strong duality $\inf_{\mu\in\mathcal{U}^\rho(\mu^0)} \mathbb{E}_{s\sim\mu} V(s) = \max_{\alpha\in[0,H]}\{\mathbb{E}_{s\sim\mu^0}[V(s)]_\alpha - \sqrt{\rho \mathrm{Var}_{s\sim\mu^0}([V(s)]_\alpha)}\}$ (Shi et al., 2023). The nonlinearity could lead to $O(H)$ error amplification in the regression approximation and $O(H^H)$ error accumulation in the regret bound in online DRMDPs. This justifies our choice of TV distance in the definition of the $d$-rectangular uncertainty set.

### 4.2 UCB Exploration in DRMDP

In online DRMDPs, the ridge estimator in (4.3) is not sufficient for finding the optimal robust policy due to being greedy on past data that provides only partial information of the environment. Hence, we propose to incorporate a robust Upper Confidence Bonus (UCB) in the Q-function estimation to explore the source environment to avoid such myopic behavior.

We present our algorithm DR-LSVI-UCB in Algorithm 1. In each episode, DR-LSVI-UCB consists of two phases. In Phase 1 (Lines 2-14), it updates the robust Q-function estimation through backward induction. Specifically, the parameters used to form the robust Q-function estimation are updated by first solving ridge regressions according to (4.1) and then solving optimization problems derived from Proposition 3.4. Next, a robust UCB is added to the Q-function estimation, whose exact form will be discussed in Remark 4.5. Finally, we truncate the robust Q-function at the fail state, by setting $Q_h^{k,\rho}(s_f, a) = 0$ for any $a \in \mathcal{A}$. In Phase 2 (Lines 15-17), it executes the greedy policy associated with the estimated robust Q-function to explore the source domain, and collects a new trajectory.

**Remark 4.5.** In Line 9 of Algorithm 1, we denote $\alpha_i^\star = \operatorname{argmax}_{\alpha\in[0,H]}\{z_{h,i}^k(\alpha) - \rho\alpha\}$ for any $i \in [d]$. Then we compute $\nu_{h,i}^{\rho,k} = z_{h,i}^k(\alpha_i^\star) - \rho\alpha_i^\star$, where $z_{h,i}^k(\alpha_i^\star)$ is the $i$-th element of vector $\boldsymbol{z}_h^k(\alpha_i^\star)$. This immediately implies that we have to solve $d$ distinct ridge regressions in Line 8 to obtain different coordinates of $\boldsymbol{\nu}_h^{\rho,k}$. This further leads to our design of the robust UCB term in Line 11, $\beta \sum_{i=1}^d \phi_i(s,a)[\mathbf{1}_i^\top (\Lambda_h^k)^{-1}\mathbf{1}_i]^{1/2}$, which consists of $d$ different upper confidence bonuses. This design is motivated from the optimism principle used in standard MDPs (Azar et al., 2017; Jin et al., 2020), where a bonus term proportional to the approximation error is added to guide exploration. A distinctive feature of the robust UCB term in Algorithm 1 is that the approximation error arises from $d$ ridge regressions, due to the $d$-rectangular uncertainty set.

**Remark 4.6.** In practice, Algorithm 1 extends to broader scenarios, where uncertainty level $\rho$ varies across different uncertainty sets $\{\mathcal{U}_{h,i}^\rho(\mu_{h,i}^0)\}_{i,h=1}^{d,H}$. We denote $\boldsymbol{\rho} = \{\rho_{h,i}\}_{i,h=1}^{d,H}$, where $\rho_{h,i}$ is the uncertainty level for the $i$-th factor uncertainty set at step $h$. To generalize Algorithm 1, we simply replace $\rho$ in Line 9 with $\rho_{h,i}$. This updated algorithm handles varied uncertainty levels with $\boldsymbol{\rho}$ chosen to satisfy various objectives. Importantly, due to the bounded range of $\{\rho_{h,i}\}_{i,h=1}^{d,H}$ in $[0,1]$ and the independence of factor uncertainty sets, heterogeneity in uncertainty level does not impact our analysis. Therefore, the modified algorithm maintains the average suboptimality bound of the original algorithm, as depicted in Section 5.

## 5 MAIN THEORETICAL RESULTS

Now we present our main result for Algorithm 1.

**Theorem 5.1.** Under Assumptions 3.1 and 4.1, there

**Algorithm 1** DR-LSVI-UCB

**Require:** Parameters $\beta > 0$ and $\lambda > 0$

**for** episode $k = 1, \cdots, K$ **do**
  Receive the initial state $s_1^k$.
  **for** stage $h = H, \cdots, 1$ **do**
    $\Lambda_h^k \leftarrow \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)\boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top + \lambda \boldsymbol{I}$
    **if** $h = H$ **then**
      $\boldsymbol{\nu}_h^{\rho,k} \leftarrow 0$
    **else**
      Update $z_h^k(\alpha)$ according to (4.2).
      $\nu_{h,i}^{\rho,k} \leftarrow \max_{\alpha \in [0,H]}\{z_{h,i}^k(\alpha) - \rho\alpha\}, i \in [d]$
    **end if**
    $\Gamma_h^k(s,a) \leftarrow \beta \sum_{i=1}^d \phi_i(s,a)[\mathbf{1}_i^\top (\Lambda_h^k)^{-1}\mathbf{1}_i]^{1/2}$
    $Q_h^{k,\rho}(s,a) \leftarrow \min\{\boldsymbol{\phi}(s,a)^\top(\boldsymbol{\theta}_h + \boldsymbol{\nu}_h^{\rho,k}) + \Gamma_h^k(s,a), H-h+1\}_+ \mathbb{1}\{s \neq s_f\}$
    $\pi_h^k(s) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_h^{k,\rho}(s,a)$
  **end for**
  **for** stage $h = 1, \cdots, H$ **do**
    Take the action $a_h^k \leftarrow \pi_h^k(s_h^k)$, and receive the next state $s_{h+1}^k$.
  **end for**
**end for**

exists an absolute constant $c > 0$ such that, for any fixed $p \in (0,1)$, if we set $\lambda = 1$ and $\beta = c \cdot dH\sqrt{\iota}$ with $\iota = \log(3dKH/p)$ in Algorithm 1, then with probability at least $1-p$ the average suboptimality of DR-LSVI-UCB satisfies

$$\text{AveSubopt}(K) \le \sqrt{2H^3 \log(3/p)/K} + 2\beta/K \underbrace{\sum_{k=1}^K \sum_{h=1}^H \sum_{i=1}^d \phi_{h,i}^k \sqrt{\mathbf{1}_i^\top (\Lambda_h^k)^{-1}\mathbf{1}_i}}_{d\text{-rectangular estimation error}}, \quad (5.1)$$

where $\phi_{h,i}^k$ is the $i$-th element of $\boldsymbol{\phi}_h^k = \boldsymbol{\phi}(s_h^k, a_h^k)$ and $\mathbf{1}_i$ is the one hot vector with its $i$-th entry being 1.

The $d$-rectangular estimation error in (5.1) resembles the regression error $\sum_{k=1}^K \sum_{h=1}^H \sqrt{(\boldsymbol{\phi}_h^k)^\top (\Lambda_h^k)^{-1}\boldsymbol{\phi}_h^k}$ in the standard episodic linear MDP literature (Jin et al., 2020; He et al., 2021, 2023). However, it cannot be easily bounded by the elliptical potential lemma (Abbasi-Yadkori et al., 2011, Lemma 11), as its summands are not quadratic terms $\|\phi_h^k\|_{(\Lambda_h^k)^{-1}}$ but weighted sum of diagonal elements of $(\Lambda_h^k)^{-1}$, i.e., $\sum_{i=1}^d \phi_{h,i}^k [(\Lambda_h^k)^{-1}]_{ii}^{1/2}$. As shown in Remark 4.5, this term primarily originates from the necessity to solve $d$ distinct ridge regressions at each episode $k$ and step $h$, due to the structure of the $d$-rectangular uncertainty set. This represents a unique challenge in DRMDPs analysis with linear function approximation. Similar terms also appear in the proof of Theorem 4.1 in Ma et al. (2022) and Theorem 6.3 in Blanchet et al. (2023), which share our setting. However, their final results do not explicitly showcase this due to strong coverage assumptions on offline dataset, which may not hold in practice and are inapplicable to the off-dynamics learning setting in our paper, which requires active and incremental data collection via interaction with the source environment.

In the following, we will instantiate the average suboptimality bound in Theorem 5.1 on different examples. We start with the tabular MDP, where the number of states and actions are finite. We set dimension $d = |\mathcal{S}| \times |\mathcal{A}|$ and the feature mapping $\boldsymbol{\phi}(s,a) = \boldsymbol{e}_{(s,a)}$ as the canonical basis in $\mathbb{R}^d$. Then the $d$-rectangular assumption degenerates to the $(s,a)$-rectangular assumption (Goyal and Grand-Clement, 2023). It turns out that with this specific structure of feature mapping $\boldsymbol{\phi}(s,a) = \boldsymbol{e}_{(s,a)}$, we can bound the $d$-rectangular estimation error without further assumption.

**Corollary 5.2.** Under the setting of tabular MDP with $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$, there exists an absolute constant $c > 0$ such that, for any fixed $p \in (0,1)$, if we set $\lambda$ and $\beta$ in Algorithm 1 as in Theorem 5.1, then with probability at least $1-p$, the average suboptimality of DR-LSVI-UCB is at most $\tilde{\mathcal{O}}(\sqrt{H^4 S^3 A^3/K})$.

Note that $d = SA$ in the tabular setting. Our result in Corollary 5.2 aligns with the average regret bound $\tilde{O}(\sqrt{H^4 d^3/K})$ of LSVI-UCB in standard MDP, which can be derived by dividing the cumulative regret bound in Theorem 3.1 of Jin et al. (2020) by $K$. In addition, Dong et al. (2022) also studied the online DR-MDP problem under the $(s,a)$-rectangular assumption and proposed an algorithm with an average suboptimality bound of $\tilde{O}(\sqrt{H^4 S^2 A/K})$, improving our result by a factor of $\sqrt{SA}$. However, their algorithm is model-based and only designed for $(s,a)$-rectangular robust tabular MDPs, which is not extendable to the function approximation setting. In contrast, our DR-LSVI-UCB algorithm is model-free and amenable to function approximation. Moreover, DR-LSVI-UCB is designed for the more general $d$-rectangular linear DR-MDPs, covering a broader scope than solely the $(s,a)$-rectangular robust tabular MDPs.

Next, we consider the general $d$-rectangular linear DR-MDP setting. Under an assumption on the inherent structure of linear MDP, we have the following average suboptimality bound.

**Corollary 5.3.** For all $(\pi, h) \in \Pi \times [H]$, assume that

$$\mathbb{E}_\pi[\boldsymbol{\phi}(s_h, a_h)\boldsymbol{\phi}(s_h, a_h)^\top] \ge \alpha I, \quad (5.2)$$

where $\alpha > 0$. Then there exists an absolute constant $c > 0$ such that, for any fixed $p \in (0,1)$, if we set $\lambda$ and $\beta$ in Algorithm 1 as in Theorem 5.1, then with probability at least $1-p$ the average suboptimality of DR-LSVI-UCB is at most $\tilde{\mathcal{O}}(\sqrt{d^2 H^4/(\alpha^2 K)})$.

**Remark 5.4.** Note that $\alpha$ represents the lower bound of the smallest eigenvalue of $\mathbb{E}_\pi[\boldsymbol{\phi}(s_h, a_h)\boldsymbol{\phi}(s_h, a_h)^\top]$, which can be upper bounded by $1/d$ (Wang et al.,

2020b). When $\alpha = O(1/d)$, Corollary 5.3 suggests an average suboptimality bound of $\tilde{\mathcal{O}}(\sqrt{d^4H^4/K})$. Moreover, Blanchet et al. (2023) studied the offline setting of $d$-rectangular linear DRMDP with TV uncertainty set. Under the robust partial coverage assumption on the offline dataset, their model-based algorithm P$^2$MPO achieves $\tilde{O}(\sqrt{d^4H^4/c^\dagger K})$ suboptimality bound, where $c^\dagger$ is a problem dependent constant related to the robust partial coverage assumption. If we further assume $c^\dagger = O(1)$, then the suboptimality bound of P$^2$MPO is the same as DR-LSVI-UCB.

In contrast with P$^2$MPO, DR-LSVI-UCB does not require a precollected offline dataset satisfying the strong coverage assumption, which is unrealistic in practice. In particular, the robust partial coverage assumption requires that the offline dataset has sufficient coverage of distributions induced by the optimal robust policy and any transition kernel in the uncertainty set. Since the optimal robust policy is unknown, and there are infinite transition kernels in the uncertainty set, it's practically impossible to verify this robust partial coverage assumption. Instead, our algorithm employs an online incremental approach to explore data through active interactions with the source domain. Additionally, we can numerically compute the $d$-rectangular estimation error in (5.1), and then acquire a specific value of the high probability upper bound of the average suboptimality according to (5.1).

In addition, P$^2$MPO is computationally intractable. For example, even when the model space in their algorithm is specified for $d$-rectangular linear DRMDPs, their algorithm requires exact solution of a supremum problem, $\sup_{\nu\in\mathcal{V}}$, over the value function class to obtain a confidence region $\widehat{\mathcal{P}}_h$, and the solution of an infimum problem, $\inf_{P_h\in\widehat{\mathcal{P}}_h}$, over the confidence region $\widehat{\mathcal{P}}_h$ (see (3.1) and (6.1) in Blanchet et al. (2023) for details). These requirements make P$^2$MPO computationally intractable. In contrast, our proposed DR-LSVI-UCB algorithm is not only statistically efficient, but also computationally efficient.

**Remark 5.5.** When $\alpha = O(1/d)$, the average suboptimality bound of DR-LSVI-UCB, $\tilde{O}(\sqrt{d^4H^4/K})$, matches the average regret bound $\tilde{O}(\sqrt{d^3H^4/K})$ for LSVI-UCB in standard linear MDPs (Jin et al., 2020, Theorem 3.1) with respect to horizon length $H$ and number of episodes $K$. However, our result in the robust setting incurs an extra $\sqrt{d}$ term concerning the feature dimension. This factor emerges from the necessity for Algorithm 1 to solve $d$ distinct ridge regressions to estimate the parameter of the $d$-rectangular uncertainty set (refer to Lines 8, 9, 12 of Algorithm 1). An intriguing open question remains whether this additional $\sqrt{d}$ factor can be mitigated through algorithm design or a more refined analysis.

# 6 EXPERIMENTS

In this section, we compare DR-LSVI-UCB with its non-robust counterpart, LSVI-UCB (Jin et al., 2020), on two off-dynamics RL problems. All numerical experiments were conducted on a MacBook Pro with a 2.6 GHz 6-Core Intel CPU. The implementation of our DR-LSVI-UCB algorithm is available at https://github.com/panxulab/Distributionally-Robust-LSVI-UCB.

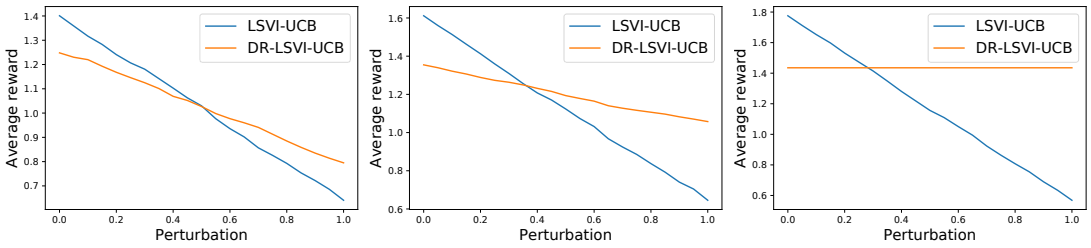
## 6.1 Simulated Off-Dynamics Linear MDPs

We first construct a linear MDP as the source domain, where the learning horizon $H = 3$, and the state space is $\mathcal{S} = \{x_1, \cdots, x_5\}$. At each step, the action $a$ is chosen from $\mathcal{A} = \{-1, 1\}^4 \subset \mathbb{R}^4$. The initial state is always $x_1$, which can transit to $x_2$, $x_4$ or $x_5$ with nonzero probabilities, where $x_4$ and $x_5$ are absorbing states. From $x_2$, the next state can be $x_3$, $x_4$ or $x_5$, and from $x_3$, it can only transit to $x_4$ or $x_5$. We design the transition probabilities and rewards such that they both depend on $\langle \xi, a\rangle$, which is bounded in $[-\|\xi\|_1, \|\xi\|_1]$ by the definition of $\mathcal{A}$, where $\xi \in \mathbb{R}^4$ is a hyperparameter of the MDP instance. We verify that this MDP satisfies Assumption 3.1 with $d = 4$. We then construct target domains by perturbing the transition probability at $x_1$ of the source domain such that the divergence is up to $q \in (0, 1)$ in TV distance. Due to the space limit, we defer more details on the construction and verification of the source domain as well as the perturbation of the target domain to Appendix A.1.

In our experiments, we consider different source MDP instances by setting $\|\xi\|_1 \in \{0.1, 0.2, 0.3\}$. To implement the uncertainty set in DR-LSVI-UCB, we use heterogeneous uncertain levels $\rho_{h,i}$ for $h \in [H]$ and $i \in [d]$ as we discussed in Remark 4.6. In particular, we set $\rho_{1,4} = 0.5$ and $\rho_{h,i} = 0$ for all other cases. We evaluate different policies based on their average rewards achieved in the target domain, which are illustrated in Figure 1. It can be seen that LSVI-UCB outperforms DR-LSVI-UCB when the dynamics shift is small, but significantly underperforms when the dynamics shift is moderate or substantial, which verifies the robustness of our DR-LSVI-UCB. We also conduct an ablation study on the effect of different values of $\rho_{1,4}$ on the performance of DR-LSVI-UCB, which is deferred to Appendix A.1 due to the space limit.

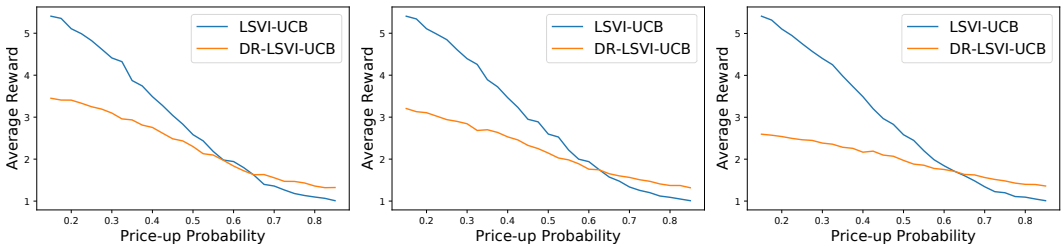
## 6.2 Simulated American Put Option

We then evaluate our algorithm in a simulated American put option problem (Tamar et al., 2014; Zhou et al., 2021; Ma et al., 2022). There is a price model in this problem, which is assumed to follow the Bernoulli




(a) $\|\xi\|_1 = 0.1$, $\rho_{1,4} = 0.5$ (b) $\|\xi\|_1 = 0.2$, $\rho_{1,4} = 0.5$ (c) $\|\xi\|_1 = 0.3$, $\rho_{1,4} = 0.5$

Figure 1: Simulation results under different source domains. The $x$-axis represents the perturbation level corresponding to different target environments. $\rho_{1,4}$ is the uncertainty level in our DR-LSVI-UCB algorithm.




(a) $\rho = 0.3$ (b) $\rho = 0.4$ (c) $\rho = 0.5$

Figure 2: Results for the simulated American put option problem. $\rho$ is the uncertainty level in DR-LSVI-UCB.

distribution

$$s_{h+1} = \begin{cases} 1.02 s_h, & \text{w.p. } p_u \\ 0.98 s_h, & \text{w.p. } 1 - p_u, \end{cases} \tag{6.1}$$

where $p_u \in (0, 1)$ is the probability that the price goes up in the next step. The initial price $s_0$ is generated uniformly from $[95, 105]$. At each step $h$, an agent can take one of the two actions: exercising the option ($a_e$) or not exercising the option ($a_{ne}$). If exercising the option, the agent receives a reward of $\max\{0, 100 - s_h\}$, and the next state would be the exit state. If not exercising the option, the agent receives 0 reward, and the next state $s_{h+1}$ is generated based on the Bernoulli distribution in (6.1). We limit the number of trading steps to $H$.

In order to employ linear function approximation, we construct a feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d+1}$ motivated by Ma et al. (2022). Specifically, we first construct the set of anchor states, $\{s_i\}_{i=1}^d$, where $s_1 = 80$, $s_{i+1} - s_i = \Delta$ and $\Delta = 60/d$. Then we define,

$$\phi(s_h, a) = \begin{cases} [\varphi_1(s_h), \cdots, \varphi_d(s_h), 0], & \text{if } a = a_e \\ [0, \cdots, 0, \max\{0, 100 - s_h\}], & \text{if } a = a_{ne}, \end{cases}$$

where $\varphi_i(s) = \max\{0, 1 - |s_h - s_i|/\Delta\}$, $i \in [d]$. In our simulation, we set the price-up probability of the source domain to $p_u = 0.5$, maximum trading steps $H$ to 10, and the feature dimension $d$ to 20. Moreover, we consider various target domains, each with a price-up probability falling within the range of $[0.15, 0.85]$. We conduct experiments on different uncertainty levels $\rho$ for DR-LSVI-UCB, and plot the average rewards for LSVI-UCB and DR-LSVI-UCB on target domains in Figure 2. It can be seen that the average rewards of robust policies are more stable over different target domains. In particular, DR-LSVI-UCB outperforms LSVI-UCB under worst-cases when the price-up probability of the target domain is much higher than that of the source domain.

## 7 CONCLUSION

We studied off-dynamics RL under the framework of online DRMDPs with linear function approximation. We proposed a model-free algorithm DR-LSVI-UCB, which learns the optimal robust policy through active interaction with the source domain. This is the first provably efficient DRMDP algorithm for off-dynamics RL with function approximation. We established the first non-asymptotic suboptimality bound for this setting, which is independent of state and action space sizes. We validated the performance and robustness of DR-LSVI-UCB on carefully designed instances. It remains an intriguing open question whether the theoretical bounds for online DRMDPs can match that of standard linear MDPs. It is also of great interest to derive lower bounds on $d$-rectangular linear DRMDPs to see its fundamental limits.

## Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. PX was supported in part by the National Science Foundation (DMS-2323112) and the Whitehead Scholars Program at the Duke University School of Medicine. The views and conclusions in this paper are those of the authors and should not be interpreted as representing any funding agencies.

### References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24. (pp. 7 and 27.)

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. (2019). Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32. (p. 4.)

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR. (p. 6.)

Badrinath, K. P. and Kalathil, D. (2021). Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pages 511–520. PMLR. (pp. 2 and 3.)

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR. (p. 2.)

Blanchet, J., Lu, M., Zhang, T., and Zhong, H. (2023). Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *arXiv preprint arXiv:2305.09659*. (pp. 3, 4, 5, 7, and 8.)

Derman, E., Mankowitz, D., Mann, T., and Mannor, S. (2020). A bayesian approach to robust reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 648–658. PMLR. (p. 2.)

Derman, E., Mankowitz, D. J., Mann, T. A., and Mannor, S. (2018). Soft-robust actor-critic policy-gradient. *arXiv preprint arXiv:1803.04848*. (p. 2.)

Dong, J., Li, J., Wang, B., and Zhang, J. (2022). Online policy optimization for robust mdp. *arXiv preprint arXiv:2209.13841*. (pp. 3 and 7.)

Eysenbach, B., Asawa, S., Chaudhari, S., Levine, S., and Salakhutdinov, R. (2020). Off-dynamics reinforcement learning: Training for transfer with domain classifiers. *arXiv preprint arXiv:2006.13916*. (p. 1.)

Farebrother, J., Machado, M. C., and Bowling, M. (2018). Generalization and regularization in dqn. *arXiv preprint arXiv:1810.00123*. (p. 1.)

Goldberg, Y. and Kosorok, M. R. (2012). Q-learning with censored data. *Annals of statistics*, 40(1):529. (p. 5.)

Goyal, V. and Grand-Clement, J. (2023). Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226. (pp. 3, 4, and 7.)

He, J., Zhao, H., Zhou, D., and Gu, Q. (2023). Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR. (pp. 2, 3, 4, and 7.)

He, J., Zhou, D., and Gu, Q. (2021). Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR. (pp. 2, 3, and 7.)

Ishfaq, H., Lan, Q., Xu, P., Mahmood, A. R., Precup, D., Anandkumar, A., and Azizzadenesheli, K. (2023). Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. *arXiv preprint arXiv:2305.18246*. (p. 3.)

Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280. (pp. 2, 3, and 4.)

Jiang, Y., Zhang, T., Ho, D., Bai, Y., Liu, C. K., Levine, S., and Tan, J. (2021). Simgan: Hybrid simulator identification for domain adaptation via adversarial reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2884–2890. IEEE. (p. 1.)

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR. (pp. 2, 3, 4, 6, 7, 8, 27, and 28.)

Koos, S., Mouret, J.-B., and Doncieux, S. (2012). The transferability approach: Crossing the reality gap in evolutionary robotics. *IEEE Transactions on Evolutionary Computation*, 17(1):122–145. (p. 1.)

Laber, E. B., Meyer, N. J., Reich, B. J., Pacifici, K., Collazo, J. A., and Drake, J. M. (2018). Optimal treatment allocations in space and time for on-line control of an emerging infectious disease. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 67(4):743–789. (p. 1.)

Liu, Z., Clifton, J., Laber, E. B., Drake, J., and Fang, E. X. (2023a). Deep spatial q-learning for infectious disease control. *Journal of Agricultural, Biological and Environmental Statistics*, pages 1–25. (p. 1.)

Liu, Z., Zhan, Z., Liu, J., Yi, D., Lin, C., and Yang, Y. (2023b). On estimation of optimal dynamic treatment regimes with multiple treatments for survival data-with application to colorectal cancer study. *arXiv preprint arXiv:2310.05049*. (p. 5.)

Ma, X., Liang, Z., Xia, L., Zhang, J., Blanchet, J., Liu, M., Zhao, Q., and Zhou, Z. (2022). Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*. (pp. 3, 4, 6, 7, 8, and 9.)

Maitin-Shepard, J., Cusumano-Towner, M., Lei, J., and Abbeel, P. (2010). Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In *2010 IEEE International Conference on Robotics and Automation*, pages 2308–2315. IEEE. (p. 1.)

Mankowitz, D. J., Levine, N., Jeong, R., Shi, Y., Kay, J., Abdolmaleki, A., Springenberg, J. T., Mann, T., Hester, T., and Riedmiller, M. (2019). Robust reinforcement learning for continuous control with model misspecification. *arXiv preprint arXiv:1906.07516*. (p. 2.)

Mannor, S., Mebel, O., and Xu, H. (2016). Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509. (p. 3.)

Modi, A., Jiang, N., Tewari, A., and Singh, S. (2020). Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR. (pp. 2 and 3.)

Nilim, A. and El Ghaoui, L. (2005). Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798. (pp. 2 and 4.)

Packer, C., Gao, K., Kos, J., Krähenbühl, P., Koltun, V., and Song, D. (2018). Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*. (p. 1.)

Panaganti, K. and Kalathil, D. (2022). Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR. (p. 3.)

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. (2022). Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35:32211–32224. (pp. 2, 3, and 5.)

Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. (2018). Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE. (p. 1.)

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. (2017). Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. PMLR. (p. 2.)

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons. (p. 1.)

Satia, J. K. and Lave Jr, R. E. (1973). Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740. (p. 2.)

Shen, Y., Xu, P., and Zavlanos, M. (2024). Wasserstein distributionally robust policy evaluation and learning for contextual bandits. *Transactions on Machine Learning Research*. Featured Certification. (p. 2.)

Shi, L. and Chi, Y. (2022). Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*. (pp. 2, 3, and 6.)

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. (2023). The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*. (pp. 3, 5, and 6.)

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press. (pp. 1 and 4.)

Tamar, A., Mannor, S., and Xu, H. (2014). Scaling up robust mdps using function approximation. In *International conference on machine learning*, pages 181–189. PMLR. (pp. 2, 3, and 8.)

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE. (p. 1.)

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434. (p. 22.)

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press. (p. 25.)

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. (2022). Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, pages 22430–22456. PMLR. (pp. 3 and 4.)

Wang, R., Du, S. S., Yang, L., and Salakhutdinov, R. R. (2020a). On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 33:17816–17826. (p. 3.)

Wang, R., Foster, D. P., and Kakade, S. M. (2020b). What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*. (p. 7.)

Wei, C.-Y., Jahromi, M. J., Luo, H., and Jain, R. (2021). Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR. (p. 4.)

Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183. (pp. 2 and 3.)

Wulfmeier, M., Posner, I., and Abbeel, P. (2017). Mutual alignment transfer learning. In *Conference on Robot Learning*, pages 281–290. PMLR. (p. 1.)

Xu, H. and Mannor, S. (2006). The robustness-performance tradeoff in markov decision processes. *Advances in Neural Information Processing Systems*, 19. (pp. 2 and 3.)

Xu, Z., Panaganti, K., and Kalathil, D. (2023). Improved sample complexity bounds for distributionally robust reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 9728–9754. PMLR. (p. 3.)

Yang, L. and Wang, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR. (pp. 2 and 3.)

Yang, W., Wang, H., Kozuno, T., Jordan, S. M., and Zhang, Z. (2023a). Avoiding model estimation in robust markov decision processes with a generative model. *arXiv preprint arXiv:2302.01248*. (p. 3.)

Yang, W., Zhang, L., and Zhang, Z. (2022). Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248. (pp. 2 and 3.)

Yang, Z., Guo, Y., Xu, P., Liu, A., and Anandkumar, A. (2023b). Distributionally robust policy gradient for offline contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 6443–6462. PMLR. (p. 2.)

Yu, P. and Xu, H. (2015). Distributionally robust counterpart in markov decision processes. *IEEE Transactions on Automatic Control*, 61(9):2538–2543. (p. 3.)

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. (2020). Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR. (p. 3.)

Zhang, H., Chen, H., Boning, D., and Hsieh, C.-J. (2021). Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*. (p. 2.)

Zhao, W., Queralta, J. P., and Westerlund, T. (2020). Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pages 737–744. IEEE. (p. 1.)

Zhao, Y.-Q., Zhu, R., Chen, G., and Zheng, Y. (2018). Constructing stabilized dynamic treatment regimes for censored data. *arXiv preprint arXiv:1808.01332*. (p. 5.)

Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR. (pp. 3 and 8.)

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   The code of our implementation is available at https://github.com/panxulab/Distributionally-Robust-LSVI-UCB.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

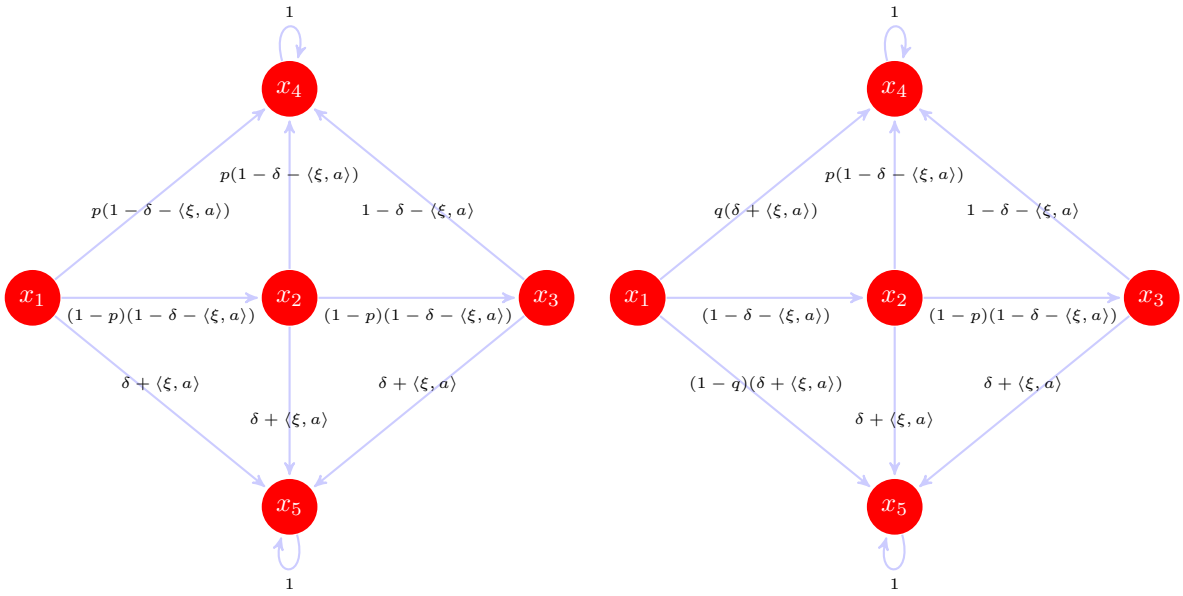
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# A EXPERIMENT SETUP AND ADDITIONAL RESULTS

In this section, we provide additional details and more experimental results for our numerical study in Section 6.

## A.1 Simulated Linear MDP

We first describe the details about the construction of the source and target linear MDPs in Section 6.1 and then provide the implementation of our method. We also present more ablation study on the robustness of our method with respect to the input parameter $\rho_{1,4}$ which stands for the uncertainty level.




(a) The source MDP environment.

(b) The target MDP environment.

Figure 3: The source and the target linear MDP environments. The value on each arrow represents the transition probability. For the source MDP, there are five states and three steps, with the initial state being $x_1$, the fail state being $x_4$, and $x_5$ being an absorbing state with reward 1. The target MDP on the right is obtained by perturbing the transition probability at the first step of the source MDP, with others remaining the same.

**Construction of the linear MDP** The source environment MDP is showed in Figure 3(a). We recall that the learning horizon is $H = 3$, the state space is $\mathcal{S} = \{x_i\}_{i=1}^5$, and the action space is $\mathcal{A} = \{-1, 1\}^4 \subset \mathbb{R}^4$. The initial state in each episode is always $x_1$. We construct the feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ with $d = 4$ as follows:

$$\begin{aligned}
\phi(x_1, a) &= (1 - \delta - \langle \xi, a\rangle, 0, 0, \delta + \langle \xi, a\rangle)^\top, \\
\phi(x_2, a) &= (0, 1 - \delta - \langle \xi, a\rangle, 0, \delta + \langle \xi, a\rangle)^\top, \\
\phi(x_3, a) &= (0, 0, 1 - \delta - \langle \xi, a\rangle, \delta + \langle \xi, a\rangle)^\top, \\
\phi(x_4, a) &= (0, 0, 1, 0)^\top, \\
\phi(x_5, a) &= (0, 0, 0, 1)^\top,
\end{aligned}$$

where the $\delta$ and $\xi$ are hyperparameters. We then define the reward parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_h\}_{h=1}^3$ as

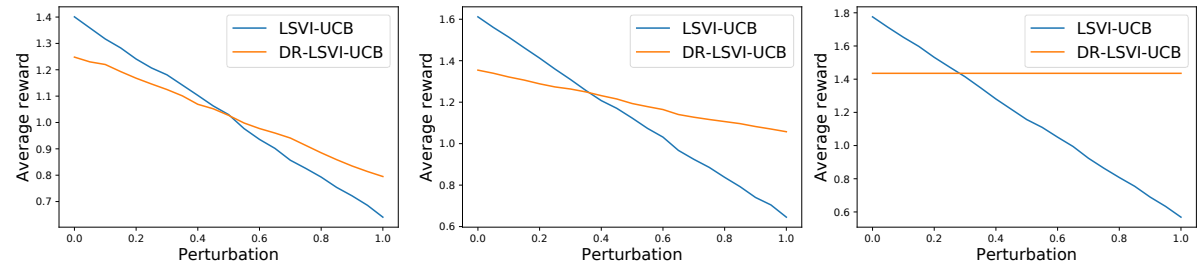
$$\boldsymbol{\theta}_1 = (0, 0, 0, 0)^\top, \ \boldsymbol{\theta}_2 = (0, 0, 0, 1)^\top \text{ and } \boldsymbol{\theta}_3 = (0, 0, 0, 1)^\top,$$

and the factor distributions $\{\boldsymbol{\mu}_h\}_{h=1}^2$ as

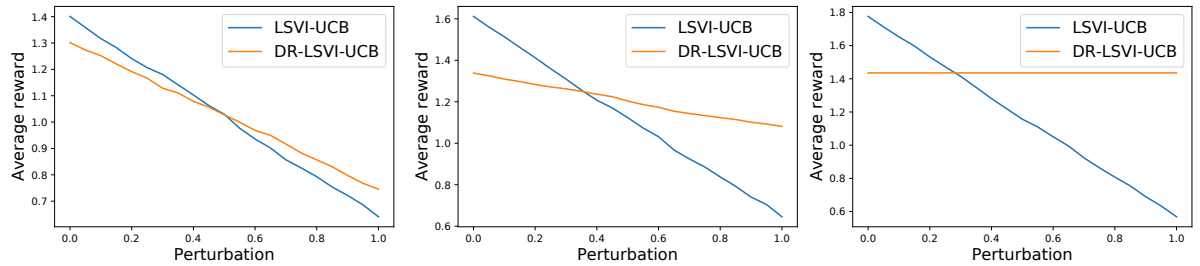
$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ((1-p)\delta_{x_2} + p\delta_{x_4}, (1-p)\delta_{x_3} + p\delta_{x_4}, \delta_{x_4}, \delta_{x_5})^\top, \tag{A.1}$$

where the $\delta_x$ is a Dirac measure which puts an atom on element $x$, and $p$ is a hyperparameter. With these notations, we define the linear reward functions as

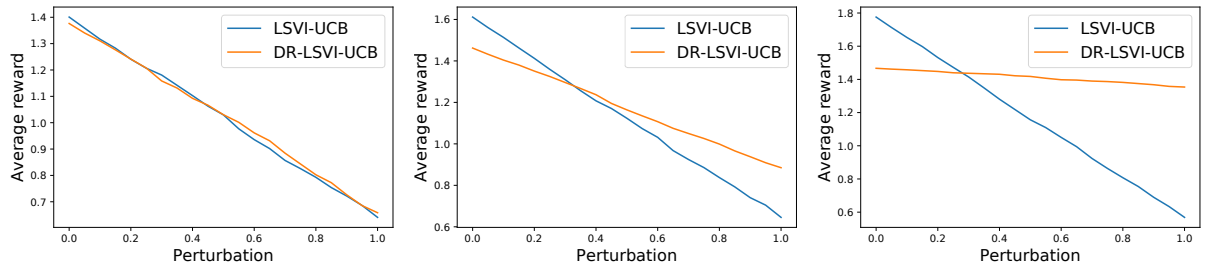
$$r_h(s, a) = \boldsymbol{\phi}(s, a)^\top \boldsymbol{\theta}_h, \ \forall (h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A},$$




(a) $\|\xi\|_1 = 0.1$, $\rho_{1,4} = 0.5$ (b) $\|\xi\|_1 = 0.2$, $\rho_{1,4} = 0.5$ (c) $\|\xi\|_1 = 0.3$, $\rho_{1,4} = 0.5$




(d) $\|\xi\|_1 = 0.1$, $\rho_{1,4} = 0.4$ (e) $\|\xi\|_1 = 0.2$, $\rho_{1,4} = 0.4$ (f) $\|\xi\|_1 = 0.3$, $\rho_{1,4} = 0.4$




(g) $\|\xi\|_1 = 0.1$, $\rho_{1,4} = 0.3$ (h) $\|\xi\|_1 = 0.2$, $\rho_{1,4} = 0.3$ (i) $\|\xi\|_1 = 0.3$, $\rho_{1,4} = 0.3$

Figure 4: Simulation results under different source domains. The $x$-axis represents the perturbation level corresponding to different target environments. $\rho_{1,4}$ is the input uncertainty level for our DR-LSVI-UCB algorithm.

and the linear transition kernels as

$$P_h(\cdot|s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_h(\cdot),\ \forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}.$$

Note that by construction, $x_4$ is a fail state in this MDP as (i) $P_h(x_4|x_4,a) = 1, \forall (h,a) \in [H] \times \mathcal{A}$, and (ii) $r_h(x_4,a) = 0,\ \forall (h,a) \in [H] \times \mathcal{A}$. Thus, it is easy to verify that the constructed source MDP satisfies Assumptions 3.1 and 4.1. In our simulation, we set $p = 0.001$, $\delta = 0.3$, $\xi = (1/\|\xi\|_1, 1/\|\xi\|_1, 1/\|\xi\|_1, 1/\|\xi\|_1)^\top$ and $\|\xi\|_1 = \{0.1, 0.2, 0.3\}$. Next, we construct several target domains, as showed in Figure 3(b), by perturbing the source domain. Specifically, we only perturb the factor distributions $\boldsymbol{\mu}_1$ in (A.1) for the fist step of the MDP, which is changed to

$$\boldsymbol{\mu}_1^{\text{perturbed}} = \left(\delta_{x_2}, \delta_{x_3}, \delta_{x_4}, (1-q)\delta_{x_5} + q\delta_{x_4}\right)^\top, \tag{A.2}$$

where $q$ is a factor that controls the perturbation level. In our simulation, we consider difference values of $q$ in the range $[0,1]$. Moreover, We train policies in the source domain through 100 epochs, and test those policies by computing the average reward in target domains through 100 epochs.

**Ablation study** We also conduct additional experiments to study the impact of $\rho_{1,4}$ on the robustness of our algorithm. In particular, we vary the value of $\rho_{1,4}$ in the range $\{0.3, 0.4, 0.5\}$ and set all other $\rho_{h,i} = 0$. Results of ablation study are showed in Figure 4.

To interpret the results, we first delve deeper into the source linear MDP in Figure 3(a). Note that $x_5$ is an absorbing state, and $r_h(x_5,a) = 1,\ \forall (h,a) \in [H] \times \mathcal{A}$. For any $(s,a,h) \in \{x_1, x_2, x_3, x_4\} \times \mathcal{A} \times [H]$, we have

$r_h(s,a) \le \delta + \|\xi\|_1 < 1$. Thus, the maximum reward is obtained from transitions starting from $x_5$, which can then be regarded as the goal state. Thus, in the source domain, the optimal strategy at the first step is to take action $(1,1,1,1)$, which leads to the largest transition probability, $\delta + \|\xi\|_1$, to $x_5$. However, in target domains, if action $(1,1,1,1)$ is taken at the first step, it results in a probability of $(1-q)(\delta + \|\xi\|_1)$ for transitioning to state $x_5$, and also a non-negligible probability of $q(\delta + \|\xi\|_1)$ for transitioning to the fail state $x_4$. Intuitively, when $q$ is large enough, action $(1,1,1,1)$ loses its advantage as it with high probability could cause a failure. Concretely, some calculation shows that when

$$q > \frac{4 - 2(\delta + \|\xi\|_1)(3 - \delta - \|\xi\|_1)}{(4 - 2(\delta + \|\xi\|_1))}, \tag{A.3}$$

the optimal action at the first step would be $(-1,-1,-1,-1)$, otherwise action $(1,1,1,1)$ would be the optimal action. Thus, the optimal policies learned in the source domain by the LSVI-UCB algorithm, which is non-robust, would fail in target domains where the perturbation level $q$ satisfies (A.3). This is consistent with our observation for all the settings in Figure 4, where we see a significant performance drop of LSVI-UCB when the perturbation level increases.

In contrast, the performance of DR-LSVI-UCB is more robust to the dynamics shift between the source and target domains, as exemplified in Figure 4(a). In scenarios where the MDP instance parameter $\xi$ remains the same, such as in Figures 4(a), 4(d) and 4(g), the performance of DR-LSVI-UCB gradually becomes more robust in the target domain as the uncertainty level, characterized by the parameter $\rho_{1,4}$, increases. This is because when $\rho_{1,4}$ is large enough, it become more likely that the uncertainty set considered by DR-LSVI-UCB will include the transition kernel of the target domain. This finding aligns with our theoretical analysis of the proposed DR-LSVI-UCB algorithm.

# B PROOF OF MAIN RESULTS

In this section, we provide the proofs of the robust Bellman equation, the existence of the optimal robust policy, and the linear representation of the robust Q-function.

## B.1 Proof of Proposition 3.2

We first prove the robust Bellman equation for $d$-rectangular linear DRMDPs. Specifically, we will prove the following stronger statement: there exists a set of transition kernels $\tilde{P}^\pi = \{\tilde{P}_h^\pi\}_{h=1}^H$ satisfying $\tilde{P}_h^\pi \in \mathcal{U}_h^\rho(P_h^0)$, such that

1. Robust Bellman equation holds,

$$V_h^{\pi,\rho}(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}\big[Q_h^{\pi,\rho}(s,a)\big], \tag{B.1a}$$

$$Q_h^{\pi,\rho}(s,a) = r_h(s,a) + \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\big[V_{h+1}^{\pi,\rho}(s')\big]. \tag{B.1b}$$

2. The following expressions for robust value function and robust Q-function hold,

$$V_h^{\pi,\rho}(s) = V_h^{\pi,\{\tilde{P}_i^\pi\}_{i=h}^H}(s), \tag{B.2a}$$

$$Q_h^{\pi,\rho}(s,a) = Q_h^{\pi,\{\tilde{P}_i^\pi\}_{i=h}^H}(s,a). \tag{B.2b}$$

*Proof.* We prove this proposition by induction. First, we start at the last stage $H$. The conclusion holds trivially because no transitions are involved. Suppose the conclusion holds for stage $h+1$, say there exist transition kernels $\{\tilde{P}_i^\pi\}_{i=h+1}^H$ such that

$$V_{h+1}^{\pi,\rho}(s) = V_{h+1}^{\pi,\{\tilde{P}_i^\pi\}_{i=h+1}^H}(s). \tag{B.3}$$

By the definition of $Q_h^{\pi,\rho}$, we have for any $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$Q_h^{\pi,\rho}(s,a) = \inf_{P\in\mathcal{U}^\rho(P^0)} \mathbb{E}^{\{P_i\}_{i=h}^H}\left[\sum_{i=h}^H r_i(s_i,a_i)\Big| s_h = s, a_h = a, \pi\right] \tag{B.4}$$

$$
\begin{aligned}
&= \inf_{P_i \in \mathcal{U}_i^\rho(P_i^0), h \le i \le H} \mathbb{E}^{\{P_i\}_{i=h}^H} \bigg[ \sum_{i=h}^H r_i(s_i, a_i) \Big| s_h = s, a_h = a, \pi \bigg] \\
&= r_h(s,a) + \inf_{P_i \in \mathcal{U}_i^\rho(P_i^0), h \le i \le H} \int_{\mathcal{S}} P_h(ds'|s,a) \mathbb{E}^{\{P_i\}_{i=h+1}^H} \bigg[ \sum_{i=h+1}^H r_i(s_i, a_i) \Big| s_{h+1} = s', \pi \bigg] \\
&\le r_h(s,a) + \inf_{P_h(\cdot|s,a) \in \mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \int_{\mathcal{S}} P_h(ds'|s,a) \mathbb{E}^{\{\tilde{P}_i\}_{i=h+1}^H} \bigg[ \sum_{i=h+1}^H r_i(s_i, a_i) \Big| s_{h+1} = s', \pi \bigg]. \qquad \text{(B.5)}
\end{aligned}
$$

For $d$-rectangular linear DRMDP, the uncertainty sets $\{\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ are closed, and the factor uncertainty sets $\{\mathcal{U}_{h,i}^\rho\}_{h,i=1}^{H,d}$ are decoupled from the state-action pair $(s,a)$. Thus, there exists a valid distribution $\tilde{P}_h^\pi$ such that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$
\tilde{P}_h^\pi(\cdot|s,a) = \mathop{\arg\inf}_{P_h(\cdot|s,a) \in \mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \int_{\mathcal{S}} P_h(ds'|s,a) \mathbb{E}^{\{\tilde{P}_i\}_{i=h+1}^H} \bigg[ \sum_{i=h+1}^H r_i(s_i, a_i) \Big| s_{h+1} = s', \pi \bigg]. \qquad \text{(B.6)}
$$

Then by (B.3) and the definition of $V_h^{\pi,\rho}$ and $V_h^{\pi,P}$, we have

$$
\begin{aligned}
Q_h^{\pi,\rho}(s,a) &\le r_h(s,a) + \inf_{P_h(\cdot|s,a) \in \mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \int_{\mathcal{S}} P_h(ds'|s,a) V_{h+1}^{\pi,\{\tilde{P}_i^\pi\}_{i=h+1}^H}(s') && \text{(B.7)} \\
&= r_h(s,a) + \inf_{P_h(\cdot|s,a) \in \mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \int_{\mathcal{S}} P_h(ds'|s,a) V_{h+1}^{\pi,\rho}(s') && \text{(B.8)} \\
&= r_h(s,a) + \inf_{P_h(\cdot|s,a) \in \mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \int_{\mathcal{S}} P_h(ds'|s,a) \inf_{P_i \in \mathcal{U}_i^\rho(P_i^0), h+1 \le i \le H} V_{h+1}^{\pi,\{P_i\}_{i=h+1}^H}(s') && \text{(B.9)} \\
&= r_h(s,a) + \inf_{P_i \in \mathcal{U}_i^\rho(P_i^0), h \le i \le H} \int_{\mathcal{S}} P_h(ds'|s,a) V_{h+1}^{\pi,\{P_i\}_{i=h+1}^H}(s') && \text{(B.10)} \\
&= r_h(s,a) + \inf_{P \in \mathcal{U}^\rho(P^0)} \int_{\mathcal{S}} P_h(ds'|s,a) V_{h+1}^{\pi,\{P_i\}_{i=h+1}^H}(s'),
\end{aligned}
$$

where (B.7) follows from (B.5) and the definition of $V_{h+1}^{\pi,P}$, (B.8) follows from (B.3), and (B.9) follows from the definition of $V_{h+1}^{\pi,\rho}$. Note that the RHS of (B.10) equals to $Q_h^{\pi,\rho}(s,a)$. Therefore, all the inequalities are actually equations. On the other hand, from (B.8) we have

$$
Q_h^{\pi,\rho}(s,a) = r_h(s,a) + \inf_{P_h(\cdot|s,a) \in \mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \int_{\mathcal{S}} P_h(ds'|s,a) V_{h+1}^{\pi,\rho}(s').
$$

This finishes the proof of Statement (B.1b) for step $h$.

On the other hand, by combining (B.6) and (B.5), we have

$$
Q_h^{\pi,\rho}(s,a) = \mathbb{E}^{\{\tilde{P}_i^\pi\}_{i=h}^H} \bigg[ \sum_{i=h}^H r_i(s_i, a_i) \Big| s_h = s, a_h = a, \pi \bigg] = Q_h^{\pi,\{\tilde{P}_i^\pi\}_{i=h}^H}(s,a), \qquad \text{(B.11)}
$$

which proves the existence of $\{\tilde{P}_i^\pi\}_{i=h}^H$ in Statement (B.2b).

Based on the existence of $\{\tilde{P}_i^\pi\}_{i=h}^H$, next we prove Statement (B.1a) and Statement (B.2a). By the definition of $V_h^{\pi,\rho}$, we have

$$
\begin{aligned}
V_h^{\pi,\rho}(s) &= \inf_{P_i \in \mathcal{U}_i^\rho(P_i^0), h \le i \le H} \mathbb{E}^{\{P_i\}_{i=h}^H} \bigg[ \sum_{i=h}^H r_i(s_i, a_i) \Big| s_h = s, \pi \bigg] \\
&= \inf_{P_i \in \mathcal{U}_i^\rho(P_i^0), h \le i \le H} \sum_{a\in\mathcal{A}} \pi_h(a|s) \mathbb{E}^{\{P_i\}_{i=h}^H} \bigg[ \sum_{i=h}^H r_i(s_i, a_i) \Big| s_h = s, a_h = a, \pi \bigg] \\
&\le \sum_{a\in\mathcal{A}} \pi_h(a|s) \mathbb{E}^{\{\tilde{P}_i^\pi\}_{i=h}^H} \bigg[ \sum_{i=h}^H r_i(s_i, a_i) \Big| s_h = s, a_h = a, \pi \bigg]. \qquad \text{(B.12)}
\end{aligned}
$$

By applying (B.11) to (B.12), we further have

$$V_h^{\pi,\rho}(s) \le \sum_{a\in\mathcal{A}} \pi_h(a|s) Q_h^{\pi,\rho}(s,a) \tag{B.13}$$

$$= \sum_{a\in\mathcal{A}} \pi_h(a|s) \inf_{P_i\in\mathcal{U}_i^\rho(P_i^0), h\le i\le H} \mathbb{E}^{\{P_i\}_{i=h}^H}\left[\sum_{i=h}^H r_i(s_i,a_i)\Big| s_h=s, a_h=a, \pi\right] \tag{B.14}$$

$$= \inf_{P_i\in\mathcal{U}_i^\rho(P_i^0), h\le i\le H} \sum_{a\in\mathcal{A}} \pi_h(a|s) \mathbb{E}^{\{P_i\}_{i=h}^H}\left[\sum_{i=h}^H r_i(s_i,a_i)\Big| s_h=s, a_h=a, \pi\right], \tag{B.15}$$

where (B.14) follows form the definition of $Q_h^{\pi,\rho}$. Now note that the RHS of (B.15) equals to $V_h^{\pi,\rho}(s)$. Therefore all the inequalities are actually equations. On the other hand, by (B.13) we have

$$V_h^{\pi,\rho}(s) = \sum_{a\in\mathcal{A}} \pi_h(a|s) Q_h^{\pi,\rho}(s,a). \tag{B.16}$$

This proves (B.1a) for stage $h$. By combining (B.16) with (B.11), we further have

$$V_h^{\pi,\rho}(s) = \mathbb{E}^{\{\tilde{P}_i^\pi\}_{i=h}^H}\left[\sum_{i=h}^H r_i(s_i,a_i)\Big| s_h=s, \pi\right].$$

This proves Statement (B.2a) the $V_h^{\pi,\rho}$ for stage $h$. Finally, by using an induction argument, we can finish the proof of the Statement (B.1) and (B.2). Thus, we finish the proof of Proposition 3.2. □

### B.2 Proof of Proposition 3.3

We then prove the existence of the optimal robust policy for the $d$-rectangular linear DRMDP.

*Proof.* We first define a policy $\tilde{\pi} = \{\tilde{\pi}_h\}_{h=1}^H$ such that for all $h\in[H]$,

$$\tilde{\pi}_h(s) = \operatorname*{argmax}_{a\in\mathcal{A}} \left\{ r_h(s,a) + \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \mathbb{E}_{s'\sim P_h(\cdot|s,a)} V_{h+1}^{\star,\rho}(s) \right\}. \tag{B.17}$$

Next we show that $\tilde{\pi}$ is optimal, i.e., for all $(h,s)\in[H]\times\mathcal{S}$,

$$V_h^{\tilde{\pi},\rho}(s) = V_h^{\star,\rho}(s).$$

We prove this by induction. For the last stage $H$, the conclusion holds trivially:

$$V_H^{\star,\rho}(s) = \sup_{\pi\in\Pi} V_H^{\pi,\rho}(s) = \sup_{\pi\in\Pi} \mathbb{E}\big[r_H(s_H,a_H)|s_H=s,\pi\big] = \max_{a\in\mathcal{A}} r_H(s,a) = V_H^{\tilde{\pi},\rho}(s).$$

Now suppose that the conclusion hold for stage $h+1$, i.e., for all $s\in\mathcal{S}$

$$V_{h+1}^{\tilde{\pi},\rho}(s) = V_{h+1}^{\star,\rho}(s).$$

By Proposition 3.2, we have

$$V_h^{\tilde{\pi},\rho}(s) = \mathbb{E}_{a\sim\tilde{\pi}_h(\cdot|s)}\left[Q_h^{\tilde{\pi},\rho}(s,a)\right]$$

$$= \mathbb{E}_{a\sim\tilde{\pi}_h(\cdot|s)}\left[r_h(s,a) + \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\left[V_{h+1}^{\tilde{\pi},\rho}(s')\right]\right]$$

$$= \mathbb{E}_{a\sim\tilde{\pi}_h(\cdot|s)}\left[r_h(s,a) + \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\left[V_{h+1}^{\star,\rho}(s)\right]\right] \tag{B.18}$$

$$= \max_{a\in\mathcal{A}}\left[r_h(s,a) + \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\left[V_{h+1}^{\star,\rho}(s)\right]\right], \tag{B.19}$$

where (B.18) follows from the induction assumption and (B.19) follows from the definition of $\tilde{\pi}_h$ in (B.17).

On the other hand, by the definition of $V_h^{\star,\rho}(s)$, for any $s \in \mathcal{S}$, we have

$$
\begin{aligned}
V_h^{\star,\rho}(s) &= \sup_{\pi\in\Pi} V_h^{\pi,\rho}(s) \\
&= \sup_{\pi\in\Pi} \mathbb{E}_{a\sim\pi_h(\cdot|s)}\big[Q_h^{\pi,\rho}(s,a)\big] && \text{(B.20)} \\
&= \sup_{\pi\in\Pi} \mathbb{E}_{a\sim\pi_h(\cdot|s)}\Big[r_h(s,a) + \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\big[V_{h+1}^{\pi,\rho}(s')\big]\Big] && \text{(B.21)} \\
&\le \sup_{\pi\in\Pi} \mathbb{E}_{a\sim\pi_h(\cdot|s)}\Big[r_h(s,a) + \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\big[V_{h+1}^{\star,\rho}(s')\big]\Big] && \text{(B.22)} \\
&= \max_{a\in\mathcal{A}} \mathbb{E}\Big[r_h(s,a) + \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\big[V_{h+1}^{\star,\rho}(s)\big]\Big],
\end{aligned}
$$

where (B.20) and (B.21) follow from Proposition 3.2, (B.22) is due to the fact that $V_{h+1}^{\star,\rho}(s') \ge V_{h+1}^{\pi,\rho}(s')$, $\forall s' \in \mathcal{S}$. Then by (B.19), we have $V_h^{\star,\rho}(s) \le V_h^{\pi,\rho}(s), \forall s \in \mathcal{S}$. Trivially, we also have $V_h^{\star,\rho}(s) \ge V_h^{\pi,\rho}(s)$ holds for all $s \in \mathcal{S}$. Consequently, we obtain $V_h^{\star,\rho}(s) = V_h^{\pi,\rho}(s)$, $\forall s \in \mathcal{S}$. By using an induction argument, we finish the proof. □

### B.3 Proof of Proposition 4.3

Next, we prove that for any policy $\pi$, the robust Q-function $Q_h^{\pi,\rho}(\cdot,\cdot)$ is always linear with respect to the feature mapping $\boldsymbol{\phi}(\cdot,\cdot)$. Before presenting the proof, we first recall and define some notions. First recall the fail state that is denoted as $s_f$. The feature mapping $\tilde{\boldsymbol{\phi}} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{d+1}$ is defined as

$$
\begin{aligned}
&\tilde{\boldsymbol{\phi}}(s_f,a) = [1,0,\cdots,0]^\top, \quad a \in \mathcal{A}, \\
&\tilde{\boldsymbol{\phi}}(s,a) = [0,\boldsymbol{\phi}(s,a)^\top]^\top, \quad \forall (s,a) \in \mathcal{S}/\{s_f\}\times\mathcal{A}.
\end{aligned}
$$

Accordingly, we define

$$
\tilde{\boldsymbol{\theta}}_h = \big[0,\boldsymbol{\theta}_h^\top\big]^\top,\ \tilde{\boldsymbol{\mu}}_h^0(\cdot) = \big[\delta_{s_f}(\cdot),\boldsymbol{\mu}_h^0(\cdot)^\top\big]^\top,
$$

where $\delta_{s_f}$ is the delta distribution with mass at $s_f$. Then the reward function $\{\tilde{r}_h\}_{h=1}^H$ and nominal transition kernel $\tilde{P}^0 = \{\tilde{P}_h^0\}_{h=1}^H$ have the following structures:

$$
\tilde{r}_h(s,a) = \langle\tilde{\boldsymbol{\phi}}(s,a),\tilde{\boldsymbol{\theta}}_h\rangle,\ \tilde{P}_h^0(\cdot|s,a) = \langle\tilde{\boldsymbol{\phi}}(s,a),\tilde{\boldsymbol{\mu}}_h^0(\cdot)\rangle, \quad \forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}. \tag{B.23}
$$

Given uncertainty level $\rho$, the uncertainty set centered around the nominal transition kernel $\{\tilde{P}_h^0\}_{h=1}^H$ is defined as

$$
\begin{aligned}
&\tilde{\mathcal{U}}^\rho(\tilde{P}^0) = \bigotimes_{h\in[H]} \tilde{\mathcal{U}}_h^\rho(\tilde{P}_h^0),\ \tilde{\mathcal{U}}_h^\rho(\tilde{P}_h^0) = \bigotimes_{(s,a)\in\mathcal{S}\times\mathcal{A}} \tilde{\mathcal{U}}_h^\rho(s,a;\tilde{\boldsymbol{\mu}}_h^0), \\
&\tilde{\mathcal{U}}_{h,i}^\rho(s,a;\tilde{\mu}_{h,i}^0) = \Big\{\sum_{i=1}^{d+1}\tilde{\phi}_i(s,a)\tilde{\mu}_{h,i}(\cdot) : \tilde{\mu}_{h,i} \in \tilde{\mathcal{U}}_{h,i}^\rho(\tilde{\mu}_{h,i}^0), \forall i \in [d+1]\Big\}, \\
&\tilde{\mathcal{U}}_{h,1}^\rho(\tilde{\mu}_{h,1}^0) = \delta(s_f),\ \tilde{\mathcal{U}}_{h,i}^\rho(\tilde{\mu}_{h,i}^0) = \big\{\tilde{\mu}_{h,i} : \tilde{\mu}_{h,i} \in \Delta(\mathcal{S}), D_{TV}(\tilde{\mu}_{h,i}||\tilde{\mu}_{h,i}^0) \le \rho\big\}, \quad i \in [d+1]/\{1\}.
\end{aligned}
$$

Further, we denote $[x_i]_{i\in[d]}$ as a vector with the $i$-th entry being $x_i$. Using these notions, we are ready to prove Proposition 4.3.

*Proof.* Based on the Proposition 3.4 and the linear MDP structure in (B.23), the robust Bellman equation can be written as

$$
\begin{aligned}
Q_h^{\pi,\rho}(s,a) &= \tilde{r}_h(s,a) + \inf_{\tilde{P}_h(\cdot|s,a)\in\tilde{\mathcal{U}}_h^\rho(s,a;\tilde{\boldsymbol{\mu}}_h^0)} \mathbb{E}_{s'\sim\tilde{P}_h(\cdot|s,a)} V_{h+1}^{\pi,\rho}(s') \\
&= \langle\tilde{\boldsymbol{\phi}}(s,a),\tilde{\boldsymbol{\theta}}_h\rangle + \inf_{\tilde{\mu}_{h,i}\in\tilde{\mathcal{U}}_{h,i}^\rho(\tilde{\mu}_{h,i}^0),\ i\in[d+1]} \Big\langle\tilde{\boldsymbol{\phi}}(s,a),\big[\mathbb{E}_{s'\sim\tilde{\mu}_{h,i}} V_{h+1}^{\pi,\rho}(s')\big]_{i\in[d+1]}\Big\rangle
\end{aligned}
$$

$$
\begin{aligned}
&= \left\langle \tilde{\boldsymbol{\phi}}(s,a), \tilde{\boldsymbol{\theta}}_h + \left[ \inf_{\tilde{\mu}_{h,i} \in \tilde{\mathcal{U}}^{\rho}_{h,i}(\tilde{\mu}^0_{h,i})} \mathbb{E}_{s' \sim \tilde{\mu}_{h,i}} V^{\pi,\rho}_{h+1}(s') \right]_{i \in [d+1]} \right\rangle \qquad \text{(B.24)} \\
&= \left\langle \tilde{\boldsymbol{\phi}}(s,a), \tilde{\boldsymbol{\theta}}_h + \left[ \max_{\alpha \in [0,H]} \left\{ \mathbb{E}^{\tilde{\mu}^0_{h,i}} \left[ V^{\pi,\rho}_{h+1} \right]_{\alpha} - \rho\alpha \right\} \right]_{i \in [d+1]} \right\rangle \\
&= \left\langle \tilde{\boldsymbol{\phi}}(s,a), \tilde{\boldsymbol{\theta}}_h + \tilde{\boldsymbol{\nu}}^{\pi,\rho}_h \right\rangle,
\end{aligned}
$$

where $\tilde{\boldsymbol{\nu}}^{\pi,\rho}_h = [\tilde{\nu}^{\pi,\rho}_{h,i}]_{i \in [d+1]}$, $\tilde{\nu}^{\pi,\rho}_{h,i} = \max_{\alpha \in [0,H]} \{\tilde{z}^{\pi}_{h,i}(\alpha) - \rho\alpha\}$, $\tilde{z}^{\pi}_{h,i}(\alpha) = \mathbb{E}^{\tilde{\mu}^0_{h,i}}[V^{\pi,\rho}_{h+1}(s')]_{\alpha}$, and (B.24) holds due to the fact that $\tilde{\boldsymbol{\phi}}(s,a) \geq 0$ and $\{\tilde{\mu}_{h,i}\}_{i \in [d+1]}$ are independent across dimensions, and thus the infimum can be moved elementwisely into the inner product. Note that $\tilde{\theta}_{h,1} = 0$ and $\tilde{\nu}^{\pi,\rho}_{h,1} = 0$, we have

$$
Q^{\pi,\rho}_h(s,a) = \left\langle \phi(s,a), \boldsymbol{\theta}_h + \boldsymbol{\nu}^{\pi,\rho}_h \right\rangle \mathbb{1}\{s \neq s_f\},
$$

where $\boldsymbol{\nu}^{\pi,\rho}_h = [\nu^{\pi,\rho}_{h,i}]_{i \in [d]}$, $\nu^{\pi,\rho}_{h,i} = \max_{\alpha \in [0,H]} \{z^{\pi}_{h,i}(\alpha) - \rho\alpha\}$, and $z^{\pi}_{h,i}(\alpha) = \mathbb{E}^{\mu^0_{h,i}}[V^{\pi,\rho}_{h+1}(s')]_{\alpha}$. □

# C PROOF OF THE MAIN RESULTS

In this section, we provide the proofs of our main theoretical results presented in Section 5.

**Notation:** Throughout this section, we denote value function as $V^{k,\rho}_h(s) = \max_a Q^{k,\rho}_h(s,a)$, feature vector $\boldsymbol{\phi}^k_h = \boldsymbol{\phi}(s^k_h, a^k_h)$. For a vector $\boldsymbol{x}$, we denote $(\boldsymbol{x})_j$ as its $j$-th entry. And we denote $[x_i]_{i \in [d]}$ as a vector with the $i$-th entry being $x_i$. For two $d$ dimensional vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, we denote $\boldsymbol{a} \leq \boldsymbol{b}$ as the fact that $a_i - b_i \leq 0, \forall i \in [d]$. For a matrix $A$, denote $\lambda_i(A)$ as the $i$-th eigenvalue of $A$. For two matrices $A$ and $B$, we denote $A \leq B$ as the fact that $B - A$ is a positive semidefinite matrix.

## C.1 Proof of Theorem 5.1

To begin with, we provide the technical lemmas that will be useful in our proof. The following concentration lemma bounds the error of the least-squares value iteration.

**Lemma C.1.** Under the setting of Theorem 5.1, let $c_\beta$ be the constant in our definition of $\beta$. There exists an absolute constant $C$ that is independent of $c_\beta$ such that for any $p \in [0,1]$, if we let $\mathcal{E}$ be the event that for any $(k,h) \in [K] \times [H]$,

$$
\left\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}^{\tau}_h \left[ \left[ V^{k,\rho}_{h+1}(s^{\tau}_{h+1}) \right]_{\alpha} - \left[ \mathbb{P}^0_h \left[ V^{k,\rho}_{h+1} \right]_{\alpha} \right](s^{\tau}_h, a^{\tau}_h) \right] \right\|^2_{(\Lambda^k_h)^{-1}} \leq C \cdot d^2 H^2 \log[3(c_\beta + 1)dT/p],
$$

then $\mathbb{P}(\mathcal{E}) \geq 1 - p/3$.

The following lemma states that $Q^{k,\rho}_h$ in Algorithm 1 can always be an upper bound of $Q^{\star,\rho}_h$ with high confidence.

**Lemma C.2.** (UCB) Under the setting of Theorem 5.1, on the event $\mathcal{E}$ defined in Lemma C.1, we have

$$
\forall (s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K], \ Q^{k,\rho}_h(s,a) \geq Q^{\star,\rho}_h(s,a).
$$

Next, we present a recursive formula, which is useful in proving Theorem 5.1.

**Lemma C.3.** (Recursive Formula) Let $\delta^{k,\rho}_h = V^{k,\rho}_h(s^k_h) - V^{\pi^k,\rho}_h(s^k_h)$, and

$$
\zeta^{k,\rho}_{h+1} = \mathbb{E}_{s \sim P_h(\cdot|s^k_h, a^k_h)} \left[ V^{k,\rho}_{h+1}(s) - V^{\pi^k,\rho}_{h+1}(s) \right] - \delta^{k,\rho}_{h+1}.
$$

Then on the event defined in Lemma C.1, we have the following: for any $(k,h) \in [K] \times [H]$:

$$
\delta^{k,\rho}_h \leq \delta^{k,\rho}_{h+1} + \zeta^{k,\rho}_{h+1} + 2\beta \sum_{i=1}^{d} \sqrt{\phi^k_{h,i} \mathbf{1}_i^{\top} (\Lambda^k_h)^{-1} \phi^k_{h,i} \mathbf{1}_i}.
$$

Finally, we are ready to prove the main theorem.

*Proof of Theorem 5.1.* Condition on the event $\mathcal{E}$ defined in Lemma C.1, by Lemma C.2 and Lemma C.3 we have:

$$\begin{aligned}\text{AveSubopt}(K) &= \frac{1}{K}\sum_{k=1}^{K}\left[V_1^{\star,\rho}(s_1^k) - V_1^{\pi^k,\rho}(s_1^k)\right] \\ &\le \underbrace{\frac{1}{K}\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^{k,\rho}}_{\text{(i)}} + \underbrace{\frac{2\beta}{K}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{d}\sqrt{\phi_{h,i}^k \mathbf{1}_i^\top (\Lambda_h^k)^{-1}\phi_{h,i}^k \mathbf{1}_i}}_{\text{(ii)}}. \end{aligned} \tag{C.1}$$

For the first term (i), $\{\zeta_h^{k,\rho}\}$ is a martingale difference sequence satisfying $|\zeta_h^{k,\rho}| \le H$ for all $(k,h) \in [K]\times[H]$. Therefore, by the Azuma-Hoeffding inequality, for any $t > 0$, we have

$$\mathbb{P}\left(\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^{k,\rho} > t\right) \le \exp\left(\frac{-t^2}{2KH\cdot H^2}\right).$$

Hence with probability at least $1 - p/3$, we have

$$\text{(i)} = \frac{1}{K}\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^{k,\rho} \le H\sqrt{\frac{2H\log(3/p)}{K}}. \tag{C.2}$$

Maintaining the second term, thus we have

$$\text{AveSubopt}(K) \le H\sqrt{\frac{2H\log(3/p)}{K}} + \frac{2\beta}{K}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{d}\sqrt{\phi_{h,i}^k \mathbf{1}_i^\top (\Lambda_h^k)^{-1}\phi_{h,i}^k \mathbf{1}_i}. \tag{C.3}$$

This completes the proof of Theorem 5.1. □

In the rest of this section, we prove Corollaries 5.2 and 5.3 respectively to further bound the term (ii) in (C.1).

### C.2 Proof of Corollary 5.2

*Proof.* To prove Corollary 5.2, it remains to bound the term (ii) in (C.1) using the structure of tabular MDP. Under tabular MDP, We set dimension $d = |\mathcal{S}|\times|\mathcal{A}|$ and the feature mapping $\boldsymbol{\phi}(s,a) = \boldsymbol{e}_{(s,a)}$ as the canonical basis in $\mathbb{R}^d$. Define

$$N_h^k(s,a) = \sum_{\tau=1}^{k-1}\mathbb{1}\{(s_h^\tau, a_h^\tau) = (s,a)\}, \quad \boldsymbol{N}_h^k = [N_h^k(s,a)]_{(s,a)\in\mathcal{S}\times\mathcal{A}}.$$

By the definition of feature mapping and $\Lambda_h^k$, we have

$$\Lambda_h^k = \sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau(\boldsymbol{\phi}_h^\tau)^\top + \lambda I = \text{diag}(\boldsymbol{N}_h^k + \lambda\mathbf{1}),$$

where $\mathbf{1}$ is the vector with all entries being 1. By our choice of $\lambda$, we have

$$\begin{aligned}\text{(ii)} &= \frac{2\beta}{K}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{N_h^k(s_h^k,a_h^k)+1}} \\ &\le \frac{2\beta}{K}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{1}{\sqrt{N_h^k(s_h^k,a_h^k)}} \\ &= \frac{2\beta}{K}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{i=1}^{N_h^K(s,a)}\frac{1}{\sqrt{i}}\end{aligned}$$

$$\leq \frac{4\beta}{K}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{N_h^K(s,a)} \tag{C.4}$$

$$\leq \frac{4\beta}{K}\sum_{h=1}^{H}\sqrt{SA\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}N_h^K(s,a)} \tag{C.5}$$

$$= \frac{4\beta}{K}H\sqrt{SAK}, \tag{C.6}$$

where (C.4) follows from the fact that $\sum_{i=1}^{N}\frac{1}{\sqrt{i}}\leq\sqrt{N}$, (C.5) follows from Cauchy-Schwarz inequality. Substitute (C.6) into (C.3) and with our choice of $\beta = c_\beta\cdot dH\sqrt{\log 3dHK/p}$ and the fact $d=SA$ we have

$$\begin{aligned}\mathrm{AveSubopt}(K) &\leq H\sqrt{\frac{2H\log(3/p)}{K}} + \frac{4c_\beta\cdot SAH\sqrt{\log 3SAHK/p}}{K}H\sqrt{SAK}\\ &\leq \frac{c(SA)^{3/2}H^2\sqrt{\log 3SAHK/p}}{\sqrt{K}},\end{aligned}$$

which completes the proof. □

### C.3 Proof of Corollary 5.3

The proof of this corollary requires the following concentration inequality.

**Lemma C.4.** (Tropp, 2012, Matrix Azuma inequality) Consider a finite adapted sequence $\{X_k\}$ of self-adjoint matrices in dimension $d$, and a fixed sequence $\{A_k\}$ of self-adjoint matrices that satisfy

$$\mathbb{E}_{k-1}[X_k] = 0 \text{ and } X_k^2 \leq A_k^2 \text{ almost surely.}$$

Compute the variance parameter

$$\sigma^2 := \left\|\sum_k A_k^2\right\|.$$

Then, for all $t\geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k X_k\right)\geq t\right\}\leq d\cdot e^{-t^2/8\sigma^2}.$$

*Proof of Corollary 5.3.* Based on the proof of Theorem 5.1, it remains to bound the term (ii) in (C.1) using the condition in (5.2). By Cauchy–Schwarz inequality we have

$$\begin{aligned}\text{(ii)}\cdot K &= 2\beta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{d}\sqrt{\phi_{h,i}^k\mathbf{1}_i^\top(\Lambda_h^k)^{-1}\phi_{h,i}^k\mathbf{1}_i}\\ &= 2\beta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{d}\phi_{h,i}^k\sqrt{\mathbf{1}_i^\top(\Lambda_h^k)^{-1}\mathbf{1}_i}\\ &\leq 2\beta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{d}\phi_{h,i}^k\sqrt{\lambda_{\max}\big((\Lambda_h^k)^{-1}\big)} \qquad \text{(C.7)}\\ &= 2\beta\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\lambda_{\max}\big((\Lambda_h^k)^{-1}\big)}\\ &= 2\beta\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{\lambda_{\min}(\Lambda_h^k)}}\end{aligned}$$

$$\leq 2\beta\sqrt{K}\sum_{h=1}^{H}\sqrt{\sum_{k=1}^{K}\frac{1}{\lambda_{\min}(\Lambda_h^k)}},$$

where (C.7) follows by the fact for any matrix $\boldsymbol{A}$, $\lambda_{\min} \leq \boldsymbol{A}_{ii} \leq \lambda_{\max}$, where $\boldsymbol{A}_{ii}$ is the $i$-th diagonal element of $\boldsymbol{A}$.

Next we bound $\lambda_{\min}(\Lambda_h^k)$. First, fix $(k,h) \in [K] \times [H]$. Recall that $\Lambda_h^k = \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau (\boldsymbol{\phi}_h^\tau)^\top + \lambda I$, we have

$$\Lambda_h^k - \mathbb{E}[\Lambda_h^k] = \sum_{\tau=1}^{k-1}\big[\boldsymbol{\phi}_h^\tau(\boldsymbol{\phi}_h^\tau)^\top - \mathbb{E}_{\pi^\tau}\big[\boldsymbol{\phi}_h^\tau(\boldsymbol{\phi}_h^\tau)^\top\big]\big] = \sum_{\tau=1}^{k-1} X_h^\tau,$$

where $X_h^\tau = \boldsymbol{\phi}_h^\tau(\boldsymbol{\phi}_h^\tau)^\top - \mathbb{E}_{\pi^\tau}[\boldsymbol{\phi}_h^\tau(\boldsymbol{\phi}_h^\tau)^\top]$. Then $\{X_h^\tau\}$ is a matrix martingale difference sequence. Note that $\|\boldsymbol{\phi}_h^\tau(\boldsymbol{\phi}_h^\tau)^\top\|_{\mathrm{op}} \leq 1$, then we have

$$\|X_h^\tau\|_{\mathrm{op}} \leq \|\boldsymbol{\phi}_h^\tau(\boldsymbol{\phi}_h^\tau)^\top\|_{\mathrm{op}} + \|\mathbb{E}_{\pi^\tau}\big[\boldsymbol{\phi}_h^\tau(\boldsymbol{\phi}_h^\tau)^\top\big]\|_{\mathrm{op}} \leq 1 + \mathbb{E}_{\pi^\tau}\big[\|\boldsymbol{\phi}_h^\tau(\boldsymbol{\phi}_h^\tau)^\top\|_{\mathrm{op}}\big] \leq 2,$$

so $\|(X_h^\tau)^2\|_{\mathrm{op}} \leq \|X_h^\tau\|_{\mathrm{op}}^2 \leq 4$. Then we have $(X_h^\tau)^2 \leq 4I$ and $\sigma^2 := \|\sum_{\tau=1}^{k-1} 4I\|_{\mathrm{op}} = 4(k-1)$. By Lemma C.4, for any $t_k \geq 0$ we have

$$\mathbb{P}\bigg\{\lambda_{\max}\bigg(-\sum_{\tau=1}^{k-1} X_h^\tau\bigg) \geq t_k\bigg\} \leq d \cdot e^{-t_k^2/32(k-1)}.$$

Let $t_k = \sqrt{32k\log(3d/\delta)}$, then with probability at least $1-\delta/3$, we have

$$\sum_{\tau=1}^{k-1} X_h^\tau \geq -t_k I.$$

Let $\delta = p/KH$ and define

$$\mathcal{E}^\dagger = \bigg\{\sum_{\tau=1}^{k-1} X_h^\tau \geq -t_k I : \forall (k,h) \in [K]\times[H]\bigg\},$$

then by union bound we have $\mathbb{P}(\mathcal{E}^\dagger) \geq 1 - p/3$.

By (5.2), we have

$$\mathbb{E}[\Lambda_h^k] = \sum_{\tau=1}^{k}\mathbb{E}_{\pi^\tau}\big[\boldsymbol{\phi}_h^\tau(\boldsymbol{\phi}_h^\tau)^\top + \lambda I\big] \geq \alpha(k-1)I + \lambda I.$$

Condition on $\mathcal{E}^\dagger$, we have

$$\Lambda_h^k = \Lambda_h^k - \mathbb{E}\Lambda_h^k + \mathbb{E}\Lambda_h^k \geq -t_k I + \mathbb{E}\Lambda_h^k.$$

Thus, we have

$$\lambda_{\min}(\Lambda_h^k) \geq \max\big\{\alpha(k-1) + \lambda - \sqrt{32k\log(3dKH/p)}, \lambda\big\}.$$

By our choice of $\lambda$, then we have

$$\sum_{k=1}^{K}\frac{1}{\lambda_{\min}(\Lambda_h^k)} \leq \sum_{k=1}^{K}\frac{1}{\max\{\alpha(k-1)+1-\sqrt{32k\log(3dHK/p)}, 1\}} \tag{C.8}$$

Next, we bound the RHS of (C.8). In particular, our goal is to discuss when $\alpha(k-1)+1-\sqrt{32k\log(3dHK/p)}$ is the larger term compared to 1 and can be lower bounded by $\alpha k/2$. To this end, we solve the following inequality

$$\frac{\alpha k}{2} \leq \alpha(k-1) - \sqrt{32k\log(3dHK/p)}. \tag{C.9}$$

It turns out that when $k \geq \frac{128}{\alpha^2}\log\frac{3dHK}{p}$, (C.9) holds. Thus, we further bound (C.8) as follows

$$\sum_{k=1}^{K}\frac{1}{\lambda_{\min}(\Lambda_h^k)} \leq \frac{128}{\alpha^2}\log\frac{3dHK}{p} + \sum_{k=1}^{K}\frac{2}{\alpha\cdot k} \leq \frac{128}{\alpha^2}\log\frac{3dHK}{p} + \frac{2}{\alpha}\log K, \tag{C.10}$$

where (C.10) follows from the fact that $\sum_{k=1}^{K} 1/k \leq \log K$. Therefore the term (ii) can be bounded as

$$\text{(ii)} \leq 2\beta\sum_{h=1}^{H}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\lambda_{\min}(\Lambda_h^k)}} \leq 2H\frac{\beta}{\sqrt{K}}\sqrt{\frac{128}{\alpha^2}\log\frac{3dHK}{p} + \frac{2}{\alpha}\log K}. \tag{C.11}$$

Finally combining (C.1), (C.2) and (C.11) and with our choice of $\beta = c_\beta \cdot dH\sqrt{\log 3dKH/p}$, we conclude that with probability $1-p$:

$$\text{AveSubopt}(K) \leq \frac{2H\sqrt{H}\log(3/p)}{\sqrt{K}} + \frac{2\beta H}{\sqrt{K}}\sqrt{\frac{128}{\alpha^2}\log\frac{3dHK}{p} + \frac{2}{\alpha}\log K} \leq \frac{cdH^2\log(3dHK/p)}{\alpha\sqrt{K}},$$

for some absolute constant $c$. This concludes the proof. □

# D PROOF OF TECHNICAL LEMMAS

## D.1 Proof of Lemma C.1

In this section, we prove Lemma C.1. Before the proof, we first present several auxiliary lemmas.

The following lemma states that the linear weights in Algorithm 1 are bounded.

**Lemma D.1.** For any $(k,h) \in [K]\times[H]$, denote the weight $\boldsymbol{w}_h^{\rho,k} = \boldsymbol{\theta}_h + \boldsymbol{\nu}_h^{\rho,k}$ in Algorithm 1, then $\boldsymbol{w}_h^{\rho,k}$ satisfies

$$\|\boldsymbol{w}_h^{\rho,k}\|_2 \leq 2H\sqrt{dk/\lambda}.$$

The following lemma presents a uniform self-normalized concentration over all value functions $V$ within a function class $\mathcal{V}$ and all parameters $\alpha$ with the interval $[0,H]$.

**Lemma D.2.** Let $\{x_\tau\}_{\tau=1}^{\infty}$ be a stochastic process on the state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_\tau\}_{\tau=0}^{\infty}$. Let $\{\boldsymbol{\phi}_\tau\}_{\tau=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process with $\boldsymbol{\phi}_\tau \in \mathcal{F}_{\tau-1}$, and $\|\boldsymbol{\phi}_\tau\| \leq 1$. Let $\Lambda_k = \lambda I + \sum_{\tau=1}^{k-1}\boldsymbol{\phi}_\tau\boldsymbol{\phi}_\tau^\top$, then for any $\delta > 0$, with probability at least $1-\delta$, for all $k \geq 0$, any $\alpha \in [0,H]$ and any $V \in \mathcal{V}$ such that $\sup_x |V(x)| \leq H$, we have

$$\Big\|\sum_{\tau=1}^{k}\boldsymbol{\phi}_\tau\big\{\big[V(x_\tau)\big]_\alpha - \mathbb{E}\big[\big[V(x_\tau)\big]_\alpha|\mathcal{F}_{\tau-1}\big]\big\}\Big\|_{\Lambda_k^{-1}}^2 \leq 8H^2\left[\frac{d}{2}\log\frac{k+\lambda}{\lambda} + \log\frac{\mathcal{N}_{\epsilon_1}}{\delta} + \log\frac{\mathcal{N}_{\epsilon_2}}{\delta}\right] + \frac{16k^2\epsilon_1^2}{\lambda} + \frac{8k^2\epsilon_2^2}{\lambda},$$

where $\mathcal{N}_{\epsilon_1}$ is the $\epsilon_1$ covering number of the interval $[0,H]$ with respect to the distance $\text{dist}(\alpha_1,\alpha_2) = |\alpha_1-\alpha_2|$, and $\mathcal{N}_{\epsilon_2}$ is the $\epsilon_2$ covering number of $\mathcal{V}$ with respect to the distance $\text{dist}(V_1,V_2) = \sup_x|V_1(x)-V_2(x)|$.

**Lemma D.3.** (Covering number of the function class $\mathcal{V}$) Let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following parametric form

$$V(\cdot) = \min\left\{\max_a\left\{w^\top\boldsymbol{\phi}(\cdot,a) + \beta\sum_{i=1}^{d}\sqrt{\phi_i(\cdot,a)\mathbf{1}_i^\top\Lambda^{-1}\phi_i(\cdot,a)\mathbf{1}_i}\right\}, H\right\},$$

where the parameters $(w,\beta,\Lambda,\alpha)$ satisfy $\|w\| \leq L$, $\beta \in [0,B]$, $\lambda_{\min}(\Lambda) \geq \lambda$ and $\alpha \in [0,H]$. Assume $\|\boldsymbol{\phi}(s,a)\| \leq 1$ for all (s,a) pairs, and let $\mathcal{N}_\epsilon$ be the $\epsilon$-covering number of $\mathcal{V}$ with respect to the distance $\text{dist}(V_1,V_2) = \sup_x|V_1(x)-V_2(x)|$. Then

$$\log\mathcal{N}_\epsilon \leq d\log(1+4L/\epsilon) + d^2\log\big[1+8d^{1/2}B^2/(\lambda\epsilon^2)\big].$$

**Lemma D.4.** (Vershynin, 2018, Covering number of an interval) Denote the $\epsilon$-covering number of the closed interval $[a, b]$ for some real number $b > a$ with respect to the distance metric $d(\alpha_1, \alpha_2) = |\alpha_1 - \alpha_2|$ as $\mathcal{N}_\epsilon([a, b])$. Then we have $\mathcal{N}_\epsilon([a, b]) \leq 3(b - a)/\epsilon$.

*Proof of Lemma C.1.* For all $(k, h) \in [K]\times[H]$, by Lemma D.1 we have $\|w_h^{\rho,k}\| \leq 2H\sqrt{dk/\lambda}$. By the construction of $\Lambda_h^k$, the minimum eigenvalue of $\Lambda_h^k$ is lower bounded by $\lambda$. By combining Lemmas D.2 to D.4, for any fix $\epsilon > 0$ , set $\epsilon_1 = \epsilon_2 = \epsilon$, we have

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau\left[\left[V_{h+1}^{k,\rho}(s_{h+1}^\tau)\right]_\alpha - \left[\mathbb{P}_h^0\left[V_{h+1}^{k,\rho}\right]_\alpha\right](s_h^\tau, a_h^\tau)\right]\right\|^2_{(\Lambda_h^k)^{-1}}\\
&\leq 4H^2\left[\frac{d}{2}\log\frac{k+\lambda}{\lambda} + d\log\left(1+\frac{8H\sqrt{dk}}{\epsilon\sqrt{\lambda}}\right) + d^2\log\left(1+\frac{8d^{1/2}\beta^2}{\epsilon^2\lambda}\right) + \log\frac{3H}{\epsilon} + \log\frac{3}{p}\right] + \frac{24k^2\epsilon^2}{\lambda}. \qquad \text{(D.1)}
\end{aligned}$$

In Algorithm 1, we choose parameters $\lambda = 1$ and $\beta = c_\beta dH\iota$, where $c_\beta$ is an absolute constant. Finally, picking $\epsilon = dH/k$, by (D.1), there exists an absolute $C > 0$ that is independent of $c_\beta$ such that

$$\left\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau\left[\left[V_{h+1}^{k,\rho}(s_{h+1}^\tau)\right]_\alpha - \left[\mathbb{P}_h^0\left[V_{h+1}^{k,\rho}\right]_\alpha\right](s_h^\tau, a_h^\tau)\right]\right\|^2_{(\Lambda_h^k)^{-1}} \leq C\cdot d^2H^2\log\frac{3(c_\beta+1)dKH}{p},$$

which completes the proof. □

## D.2 Proof of Lemma C.2

Before the proof of Lemma C.2, we present a lemma bounding the difference between the value function maintained in Algorithm 1 (without bonus) and the true value function of any policy $\pi$.

**Lemma D.5.** For any fixed policy $\pi$, on the event $\mathcal{E}$ defined in Lemma C.1, we have for all $(s, a, h, k) \in \mathcal{S}/\{s_f\}\times\mathcal{A}\times[H]\times[K]$ that:

$$\begin{aligned}
\langle\boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h + \boldsymbol{\nu}_h^{\rho,k}\rangle - Q_h^{\pi,\rho}(s,a) = &\inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)}\left[\mathbb{P}_hV_{h+1}^{k,\rho}\right](s,a)\\
&- \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)}\left[\mathbb{P}_hV_{h+1}^{\pi,\rho}\right](s,a) + \Delta_h^k(s,a),
\end{aligned}$$

for some $\Delta_h^k(s,a)$ that satisfies $|\Delta_h^k(s,a)| \leq \beta\sum_{i=1}^d\sqrt{\phi_i(s,a)\mathbf{1}_i^\top(\Lambda_h^k)^{-1}\phi_i(s,a)\mathbf{1}_i}$.

*Proof of Lemma C.2.* We prove this lemma by induction. Starting at step $H-1$. Since $V_H^{k,\rho}(s) = V_H^{\star,\rho}(s) = \max_a r_H(s,a)$, by Lemma D.5 we have

$$\left|\left\langle\boldsymbol{\phi}(s,a), \boldsymbol{\theta}_{H-1} + \boldsymbol{\nu}_{H-1}^{\rho,k}\right\rangle - Q_{H-1}^{\star,\rho}(s,a)\right| \leq \Gamma_{H-1}^k(s,a),$$

where $\Gamma_{H-1}^k(s,a)$ is the bonus at step $H-1$ used in Algorithm 1. Therefore, we know

$$Q_{H-1}^{k,\rho} = \min\left\{\left\langle\boldsymbol{\phi}(s,a), \boldsymbol{\theta}_{H-1} + \boldsymbol{\nu}_{H-1}^{\rho,k}\right\rangle + \Gamma_{H-1}^k(s,a), H\right\} \geq Q_{H-1}^{\star,\rho}(s,a).$$

Suppose the statement holds at stage $h+1$, $Q_{h+1}^{k,\rho}(s,a) \geq Q_{h+1}^{\star,\rho}(s,a)$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, then we have

$$V_{h+1}^{k,\rho}(s) = Q_{h+1}^{k,\rho}(s,\pi_{h+1}^k(s)) \geq Q_{h+1}^{k,\rho}(s,\pi_{h+1}^\star(s)) \geq Q_{h+1}^{\star,\rho}(s,\pi_{h+1}^\star(s)) = V_{h+1}^{\star,\rho}(s), \quad \forall s\in\mathcal{S},$$

where the first inequality holds by the fact that $\pi_{h+1}^k$ is the greedy policy with respect to $Q_{h+1}^{k,\rho}$, and the second inequality holds by the induction assumption that $Q_{h+1}^{k,\rho}(s,a) \geq Q_{h+1}^{\star,\rho}(s,a)$, $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$. Thus, we have

$$\inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)}\left[\mathbb{P}_hV_{h+1}^{k,\rho}\right](s,a) - \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)}\left[\mathbb{P}_hV_{h+1}^{\star,\rho}\right](s,a) \geq 0. \qquad \text{(D.2)}$$

Again by Lemma D.5 we have

$$\left|\left\langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h + \boldsymbol{\nu}_h^{\rho,k} \right\rangle - Q_h^{\star,\rho}(s,a) - \left( \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \left[\mathbb{P}_h V_{h+1}^{k,\rho}\right](s,a) - \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} \left[\mathbb{P}_h V_{h+1}^{\star,\rho}\right](s,a) \right) \right|$$
$$\leq \beta \sum_{i=1}^d \sqrt{\phi_i(s,a)\mathbf{1}_i^\top (\Lambda_h^k)^{-1}\phi_i(s,a)\mathbf{1}_i}.$$

By (D.2) we have

$$Q_h^{k,\rho}(s,a) = \min\left\{ \left\langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h + \boldsymbol{\nu}_h^{\rho,k} \right\rangle + \Gamma_h^k(s,a), H-h+1 \right\} \geq Q_h^{\star,\rho}(s,a),$$

which concludes the proof. □

### D.3 Proof of Lemma C.3

*Proof.* By Algorithm 1 and the definition of $\pi^k$, we have

$$\delta_h^{k,\rho} = V_h^{k,\rho}(s_h^k) - V_h^{\pi^k,\rho}(s_h^k) = Q_h^{k,\rho}(s_h^k, a_h^k) - Q_h^{\pi^k,\rho}(s_h^k, a_h^k).$$

By Lemma D.5 we have

$$\delta_h^{k,\rho} \leq \inf_{P_h(\cdot|s_h^k,a_h^k)\in\mathcal{U}_h^\rho(s_h^k,a_h^k;\boldsymbol{\mu}_h^0)} \left[\mathbb{P}_h V_{h+1}^{k,\rho}\right](s_h^k,a_h^k) - \inf_{P_h(\cdot|s_h^k,a_h^k)\in\mathcal{U}_h^\rho(s_h^k,a_h^k;\boldsymbol{\mu}_h^0)} \left[\mathbb{P}_h V_{h+1}^{\pi^k,\rho}\right](s_h^k,a_h^k)$$
$$+ 2\beta \sum_{i=1}^d \sqrt{\phi_{h,i}^k \mathbf{1}_i^\top (\Lambda_h^k)^{-1}\phi_{h,i}^k \mathbf{1}_i}. \tag{D.3}$$

For the difference on the RHS, we have

$$\begin{aligned}
&\inf_{P_h(\cdot|s_h^k,a_h^k)\in\mathcal{U}_h^\rho(s_h^k,a_h^k;\boldsymbol{\mu}_h^0)} \left[\mathbb{P}_h V_{h+1}^{k,\rho}\right](s_h^k,a_h^k) - \inf_{P_h(\cdot|s_h^k,a_h^k)\in\mathcal{U}_h^\rho(s_h^k,a_h^k;\boldsymbol{\mu}_h^0)} \left[\mathbb{P}_h V_{h+1}^{\pi^k,\rho}\right](s_h^k,a_h^k) \\
&= \left\langle \boldsymbol{\phi}(s_h^k,a_h^k), \left[ \max_{\alpha_i\in[0,H]} \left\{ \mathbb{E}^{\mu_{h,i}^0}\left[V_{h+1}^{k,\rho}(s)\right]_{\alpha_i} - \rho\alpha_i \right\} \right]_{i\in[d]} \right\rangle \\
&\quad - \left\langle \boldsymbol{\phi}(s_h^k,a_h^k), \left[ \max_{\alpha_i\in[0,H]} \left\{ \mathbb{E}^{\mu_{h,i}^0}\left[V_{h+1}^{\pi^k,\rho}(s)\right]_{\alpha_i} - \rho\alpha_i \right\} \right]_{i\in[d]} \right\rangle \\
&\leq \left\langle \boldsymbol{\phi}(s_h^k,a_h^k), \left[ \max_{\alpha_i\in[0,H]} \left\{ \mathbb{E}^{\mu_{h,i}^0}\left[V_{h+1}^{k,\rho}(s)\right]_{\alpha_i} - \mathbb{E}^{\mu_{h,i}^0}\left[V_{h+1}^{\pi^k,\rho}(s)\right]_{\alpha_i} \right\} \right]_{i\in[d]} \right\rangle.
\end{aligned}$$

By Lemma C.2, we have for all $s\in\mathcal{S}$,

$$V_{h+1}^{k,\rho}(s) = Q_{h+1}^{k,\rho}(s,\pi_{h+1}^k(s)) \geq Q_{h+1}^{k,\rho}(s,\pi_{h+1}^\star(s)) \geq Q_{h+1}^{\star,\rho}(s,\pi_{h+1}^\star(s)).$$

Since $\pi^\star$ is the greedy policy with respect to $Q_{h+1}^{\star,\rho}$, we have

$$V_{h+1}^{k,\rho}(s) \geq Q_{h+1}^{\star,\rho}(s,\pi_{h+1}^k(s)) \geq Q_{h+1}^{\pi^k,\rho}(s,\pi_{h+1}^k(s)) = V_{h+1}^{\pi^k,\rho}(s).$$

Then we have,

$$\begin{aligned}
&\inf_{P_h(\cdot|s_h^k,a_h^k)\in\mathcal{U}_h^\rho(s_h^k,a_h^k;\boldsymbol{\mu}_h^0)} \left[\mathbb{P}_h V_{h+1}^{k,\rho}\right](s_h^k,a_h^k) - \inf_{P_h(\cdot|s_h^k,a_h^k)\in\mathcal{U}_h^\rho(s_h^k,a_h^k;\boldsymbol{\mu}_h^0)} \left[\mathbb{P}_h V_{h+1}^{\pi^k,\rho}\right](s_h^k,a_h^k) \\
&\leq \left\langle \boldsymbol{\phi}(s_h^k,a_h^k), \mathbb{E}^{\boldsymbol{\mu}_h^0}\left[V_{h+1}^{k,\rho}(s) - V_{h+1}^{\pi^k,\rho}(s)\right] \right\rangle \\
&= \left[\mathbb{P}_h\left[V_{h+1}^{k,\rho} - V_{h+1}^{\pi^k,\rho}\right]\right](s_h^k,a_h^k) \\
&= \left[\mathbb{P}_h\left[V_{h+1}^{k,\rho} - V_{h+1}^{\pi^k,\rho}\right]\right](s_h^k,a_h^k) - \left[V_{h+1}^{k,\rho}(s_{h+1}^k) - V_{h+1}^{\pi^k,\rho}(s_{h+1}^k)\right] + \left[V_{h+1}^{k,\rho}(s_{h+1}^k) - V_{h+1}^{\pi^k,\rho}(s_{h+1}^k)\right] \\
&= \zeta_{h+1}^{k,\rho} + \delta_{h+1}^{k,\rho}.
\end{aligned} \tag{D.4}$$

Then we complete the proof by substituting (D.4) into (D.3). □

# E PROOF OF SUPPORTING LEMMAS

In this section, we provide the proofs of the supporting lemmas we used in Appendix D.

## E.1 Proof of Lemma D.1

The proof of Lemma D.1 will use the following fact.

**Lemma E.1.** (Jin et al., 2020, Lemma D.1) Let $\Lambda_t = \lambda \boldsymbol{I} + \sum_{i=1}^t \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top$, where $\boldsymbol{\phi}_i \in \mathbb{R}^d$ and $\lambda > 0$. Then:

$$\sum_{i=1}^t \boldsymbol{\phi}_i^\top (\Lambda_t)^{-1} \boldsymbol{\phi}_i \leq d.$$

*Proof of Lemma D.1.* Denote $\alpha_i = \text{argmax}_{\alpha \in [0,H]} \{z_{h,i}^k(\alpha) - \rho\alpha\}, i \in [d]$. For any vector $\boldsymbol{v} \in \mathbb{R}^d$, we have

$$\begin{aligned}
\big|\boldsymbol{v}^\top \boldsymbol{w}_h^{\rho,k}\big| &= \bigg|\boldsymbol{v}^\top \boldsymbol{\theta}_h + \boldsymbol{v}^\top \bigg[\max_{\alpha\in[0,H]} \{z_{h,i}^k(\alpha) - \rho\alpha\}\bigg]_{i\in[d]}\bigg| \\
&\leq \big|\boldsymbol{v}^\top \boldsymbol{\theta}_h\big| + \bigg|\boldsymbol{v}^\top \bigg[\max_{\alpha\in[0,H]} \{z_{h,i}^k(\alpha) - \rho\alpha\}\bigg]_{i\in[d]}\bigg| \\
&\leq \sqrt{d}\|\boldsymbol{v}\|_2 + H\|\boldsymbol{v}\|_1 + \bigg|\boldsymbol{v}^\top \bigg[\bigg((\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau [\max_a Q_{h+1}^{k,\rho}(s_{h+1}^\tau, a)]_{\alpha_i}\bigg)_i\bigg]_{i\in[d]}\bigg| \qquad \text{(E.1)} \\
&\leq \sqrt{d}\|\boldsymbol{v}\|_2 + H\sqrt{d}\|\boldsymbol{v}\|_2 + \sqrt{\bigg[\sum_{\tau=1}^{k-1} \boldsymbol{v}^\top (\Lambda_h^k)^{-1} \boldsymbol{v}\bigg]\bigg[\sum_{\tau=1}^{k-1} (\boldsymbol{\phi}_h^\tau)^\top (\Lambda_h^k)^{-1} (\boldsymbol{\phi}_h^\tau)\bigg]} \cdot H \qquad \text{(E.2)} \\
&\leq 2H\|\boldsymbol{v}\|_2 \sqrt{dk/\lambda}. \qquad \text{(E.3)}
\end{aligned}$$

We note that the term $[((\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau [\max_a Q_{h+1}^{k,\rho}(s_{h+1}^\tau, a)]_{\alpha_i})_i]_{i\in[d]}$ in (E.1) is constructed by first taking out the $i$-th coordinate of the ridge solution vector, $(\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau [\max_a Q_{h+1}^{k,\rho}(s_{h+1}^\tau, a)]_{\alpha_i} \in \mathbb{R}^d$, $\forall i \in [d]$, and then concatenating all $d$ values into a vector. Inequality (E.1) is due to the fact that $\rho \leq 1$, (E.2) is due to the fact that $Q_h^{k,\rho} \leq H$, and (E.3) is due to Lemma E.1 and the fact that the minimum eigenvalue of $\Lambda_h^k$ is lower bounded by $\lambda$. The remainder of the proof follows from the fact that $\|\boldsymbol{w}_h^{\rho,k}\|_2 = \max_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1} |\boldsymbol{v}^\top \boldsymbol{w}_h^{\rho,k}|$. □

## E.2 Proof of Lemma D.2

The proof of Lemma D.2 requires the following results on the concentration of self-normalized processes.

**Lemma E.2** (Concentration of Self-Normalized Processes)**.** (Abbasi-Yadkori et al., 2011, Theorem 1) Let $\{\epsilon_t\}_{t=1}^\infty$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^\infty$. Let $\epsilon_t|\mathcal{F}_{t-1}$ be mean-zero and $\sigma$-subGaussian; i.e. $\mathbb{E}[\epsilon_t|\mathcal{F}_{t-1}] = 0$, and

$$\forall \lambda \in \mathbb{R}, \qquad \mathbb{E}[e^{\lambda\epsilon_t}|\mathcal{F}_{t-1}] \leq e^{\lambda^2\sigma^2/2}.$$

Let $\{\boldsymbol{\phi}_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process where $\phi_t$ is $\mathcal{F}_{t-1}$ measurable. Assume $\Lambda_0$ is a $d \times d$ positive definite matrix, and let $\Lambda_t = \Lambda_0 + \sum_{s=1}^t \boldsymbol{\phi}_s \boldsymbol{\phi}_s^\top$. Then for any $\delta > 0$, with probability at least $1 - \delta$, we have for all $t \geq 0$:

$$\bigg\|\sum_{s=1}^t \boldsymbol{\phi}_s \epsilon_s\bigg\|_{\Lambda_t^{-1}}^2 \leq 2\sigma^2 \log\bigg[\frac{\det(\Lambda_t)^{1/2} \det(\Lambda_0)^{-1/2}}{\delta}\bigg].$$

*Proof of Lemma D.2.* For any $V \in \mathcal{V}$ and $\alpha \in [0, H]$, we know there exists a $\tilde{\alpha}$ in the $\epsilon_1$-covering and a $\tilde{V}$ in the $\epsilon_2$ covering such that

$$\begin{aligned}
&V = \tilde{V} + \Delta_V, \ \sup_x |\Delta_V(x)| \leq \epsilon, \\
&\alpha = \tilde{\alpha} + \Delta_\alpha, \ |\Delta_\alpha| \leq \epsilon.
\end{aligned}$$

This gives the following decomposition:

$$
\begin{aligned}
&\Big\|\sum_{\tau=1}^{k}\boldsymbol{\phi}_\tau\big\{\big[V(x_\tau)\big]_\alpha-\mathbb{E}\big[\big[V(x_\tau)\big]_\alpha|\mathcal{F}_{\tau-1}\big]\big\}\Big\|^2_{\Lambda_k^{-1}}\\
&\le 2\Big\|\sum_{\tau=1}^{k}\boldsymbol{\phi}_\tau\big\{\big[\tilde{V}(x_\tau)\big]_\alpha-\mathbb{E}\big[\big[\tilde{V}(x_\tau)\big]_\alpha|\mathcal{F}_{\tau-1}\big]\big\}\Big\|^2_{\Lambda_k^{-1}}\\
&\quad+2\Big\|\sum_{\tau=1}^{k}\boldsymbol{\phi}_\tau\big\{\big[V(x_\tau)\big]_\alpha-\big[\tilde{V}(x_\tau)\big]_\alpha-\mathbb{E}\big[\big[V(x_\tau)\big]_\alpha-\big[\tilde{V}(x_\tau)\big]_\alpha|\mathcal{F}_{\tau-1}\big]\big\}\Big\|^2_{\Lambda_k^{-1}}\\
&\le 4\Big\|\sum_{\tau=1}^{k}\boldsymbol{\phi}_\tau\big\{\big[\tilde{V}(x_\tau)\big]_{\tilde{\alpha}}-\mathbb{E}\big[\big[\tilde{V}(x_\tau)\big]_{\tilde{\alpha}}|\mathcal{F}_{\tau-1}\big]\big\}\Big\|^2_{\Lambda_k^{-1}}\\
&\quad+4\Big\|\sum_{\tau=1}^{k}\boldsymbol{\phi}_\tau\big\{\big[\tilde{V}(x_\tau)\big]_\alpha-\big[\tilde{V}(x_\tau)\big]_{\tilde{\alpha}}-\mathbb{E}\big[\big[\tilde{V}(x_\tau)\big]_\alpha-\big[\tilde{V}(x_\tau)\big]_{\tilde{\alpha}}|\mathcal{F}_{\tau-1}\big]\big\}\Big\|^2_{\Lambda_k^{-1}}\\
&\quad+2\Big\|\sum_{\tau=1}^{k}\boldsymbol{\phi}_\tau\big\{\big[V(x_\tau)\big]_\alpha-\big[\tilde{V}(x_\tau)\big]_\alpha-\mathbb{E}\big[\big[V(x_\tau)\big]_\alpha-\big[\tilde{V}(x_\tau)\big]_\alpha|\mathcal{F}_{\tau-1}\big]\big\}\Big\|^2_{\Lambda_k^{-1}}.
\end{aligned}
$$

We can apply Lemma E.2 and a union bound to the first term, and the second and the third term can be bounded by $16k^2\epsilon_1^2/\lambda$ and $8k^2\epsilon_2^2/\lambda$, respectively. Therefore we complete the proof. □

### E.3 Proof of Lemma D.3

The proof of Lemma D.3 will use the following fact.

**Lemma E.3.** (Jin et al., 2020, Covering Number of Euclidean Ball) For any $\epsilon>0$, the $\epsilon$-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R>0$ is upper bounded by $(1+2R/\epsilon)^d$.

*Proof of Lemma D.3.* The argument is similar to the proof of Lemma D.6 in Jin et al. (2020). Denote $A=\beta^2\Lambda^{-1}$, so we have

$$
V(\cdot)=\min\left\{\max_a\left\{\mathbf{w}^\top\boldsymbol{\phi}(\cdot,a)+\sum_{i=1}^{d}\sqrt{\phi_i(\cdot,a)\mathbf{1}_i^\top A\phi_i(\cdot,a)\mathbf{1}_i}\right\},H\right\},\tag{E.4}
$$

for $\|w\|\le L$ and $\|A\|\le B^2\lambda^{-1}$. For any two functions $V_1,V_2\in\mathcal{V}$, let them take the form in (E.4) with parameters $(\mathbf{w}_1,A_1)$ and $(\mathbf{w}_2,A_2)$, respectively. Then since both $\min\{\cdot,H\}$ and $\max_a$ are contraction maps, we have

$$
\begin{aligned}
\mathrm{dist}(V_1,V_2)&\le\sup_{x,a}\left|\left[\mathbf{w}_1^\top\boldsymbol{\phi}(x,a)+\sum_{i=1}^{d}\sqrt{\phi_i(x,a)\mathbf{1}_i^\top A_1\phi_i(x,a)\mathbf{1}_i}\right]-\left[\mathbf{w}_2^\top\boldsymbol{\phi}(x,a)+\sum_{i=1}^{d}\sqrt{\phi_i(x,a)\mathbf{1}_i^\top A_2\phi_i(x,a)\mathbf{1}_i}\right]\right|\\
&\le\sup_{\boldsymbol{\phi}:\|\boldsymbol{\phi}\|\le1}\left|\left[\mathbf{w}_1^\top\boldsymbol{\phi}+\sum_{i=1}^{d}\sqrt{\phi_i\mathbf{1}_i^\top A_1\phi_i\mathbf{1}_i}\right]-\left[\mathbf{w}_2^\top\boldsymbol{\phi}+\sum_{i=1}^{d}\sqrt{\phi_i\mathbf{1}_i^\top A_2\phi_i\mathbf{1}_i}\right]\right|\\
&\le\sup_{\boldsymbol{\phi}:\|\boldsymbol{\phi}\|\le1}\left|(\mathbf{w}_1-\mathbf{w}_2)^\top\boldsymbol{\phi}\right|+\sup_{\boldsymbol{\phi}:\|\boldsymbol{\phi}\|\le1}\sum_{i=1}^{d}\sqrt{\phi_i\mathbf{1}_i^\top(A_1-A_2)\phi_i\mathbf{1}_i}\qquad\text{(E.5)}\\
&\le\|\mathbf{w}_1-\mathbf{w}_2\|+\sqrt{\|A_1-A_2\|}\sup_{\boldsymbol{\phi}:\|\boldsymbol{\phi}\|\le1}\sum_{i=1}^{d}\|\phi_i\mathbf{1}_i\|\\
&\le\|\mathbf{w}_1-\mathbf{w}_2\|+\sqrt{\|A_1-A_2\|_F},\qquad\text{(E.6)}
\end{aligned}
$$

where (E.5) follows from triangular inequlaity and the fact that $|\sqrt{x}-\sqrt{y}|\le\sqrt{|x-y|}$, $\forall x,y\ge0$. For matrices, $\|\cdot\|$ and $\|\cdot\|_F$ denote the matrix operator norm and Frobenius norm respectively.

Let $\mathcal{C}_{\mathbf{w}}$ be an $\epsilon/2$-cover of $\{\mathbf{w} \in \mathbb{R}^d | \|\mathbf{w}\|_2 \leq L\}$ with respect to the 2-norm, and $\mathcal{C}_A$ be an $\epsilon^2/4$-cover of $\{A \in \mathbb{R}^{d\times d} | \|A\|_F \leq d^{1/2}B^2\lambda^{-1}\}$ with respect to the Frobenius norm. By Lemma E.3, we know:

$$\big|\mathcal{C}_{\mathbf{w}}\big| \leq \big(1 + 4L/\epsilon\big)^d, \quad \big|\mathcal{C}_A\big| \leq \big[1 + 8d^{1/2}B^2/(\lambda\epsilon^2)\big]^{d^2}.$$

By (E.6), for any $V_1 \in \mathcal{V}$, there exists $\mathbf{w}_2 \in \mathcal{C}_{\mathbf{w}}$ and $A_2 \in \mathcal{C}_A$ such that $V_2$ parametrized by $(\mathbf{w}_2, A_2)$ satisfies $\text{dist}(V_1, V_2) \leq \epsilon$. Hence, it holds that $\mathcal{N}_\epsilon \leq |\mathcal{C}_{\mathbf{w}}| \cdot |\mathcal{C}_A|$, which leads to

$$\log \mathcal{N}_\epsilon \leq \log|\mathcal{C}_{\mathbf{w}}| + \log|\mathcal{C}_A| \leq d\log(1 + 4L/\epsilon) + d^2 \log\big[1 + 8d^{1/2}B^2/(\lambda\epsilon^2)\big].$$

This concludes the proof. □

### E.4 Proof of Lemma D.5

*Proof.* For all $(s, a, h) \in \mathcal{S}/\{s_f\} \times \mathcal{A} \times [H]$, we have

$$Q_h^{\pi,\rho}(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h + \boldsymbol{\nu}_h^{\pi,\rho}\rangle = r_h(s,a) + \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)} [\mathbb{P}_h V_{h+1}^{\pi,\rho}](s,a).$$

This gives

$$(\boldsymbol{\theta}_h + \boldsymbol{\nu}_h^{\rho,k}) - (\boldsymbol{\theta}_h + \boldsymbol{\nu}_h^{\pi,\rho}) = \boldsymbol{\nu}_h^{\rho,k} - \boldsymbol{\nu}_h^{\pi,\rho} = \underbrace{\boldsymbol{\nu}_h^{\rho,k} - \tilde{\boldsymbol{\nu}}_h^{k,\rho}}_{I} + \underbrace{\tilde{\boldsymbol{\nu}}_h^{k,\rho} - \boldsymbol{\nu}_h^{\pi,\rho}}_{II}, \tag{E.7}$$

where $\tilde{\boldsymbol{\nu}}_h^{k,\rho} = \big[\tilde{\nu}_{h,i}^{k,\rho}\big]_{i\in[d]}$, and $\tilde{\nu}_{h,i}^{k,\rho} = \max_{\alpha\in[0,H]}\big\{\mathbb{E}^{\mu_{h,i}^0}\big[V_{h+1}^{k,\rho}(s)\big]_\alpha - \rho\alpha\big\}$. In what follows, we will bound these two terms separately.

For term $I$ in (E.7), we have

$$\boldsymbol{\nu}_h^{\rho,k} - \tilde{\boldsymbol{\nu}}_h^{k,\rho} \leq \Big[\max_{\alpha\in[0,H]}\Big\{\mathbb{E}^{\mu_{h,i}^0}\big[\widehat{V_{h+1}^{k,\rho}(s)}\big]_\alpha - \mathbb{E}^{\mu_{h,i}^0}\big[V_{h+1}^{k,\rho}(s)\big]_\alpha\Big\}\Big]_{i\in[d]}.$$

Denote $\alpha_i^k = \text{argmax}_{\alpha\in[0,H]}\big\{\mathbb{E}^{\mu_{h,i}^0}\big[\widehat{V_{h+1}^{k,\rho}(s)}\big]_\alpha - \mathbb{E}^{\mu_{h,i}^0}\big[V_{h+1}^{k,\rho}(s)\big]_\alpha\big\}$, $i = 1, \cdots, d$. Then we have

$$\begin{aligned}
&\boldsymbol{\nu}_h^{\rho,k} - \tilde{\boldsymbol{\nu}}_h^{k,\rho} \\
&\leq \Bigg[\Bigg(\big(\Lambda_h^k\big)^{-1}\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau\Big[V_{h+1}^{k,\rho}(s_{h+1}^\tau)\Big]_{\alpha_i^k}\Bigg)_i - \Big(\mathbb{E}^{\boldsymbol{\mu}_h^0}\Big[V_{h+1}^{k,\rho}(s)\Big]_{\alpha_i^k}\Big)_i\Bigg]_{i\in[d]} \\
&= \Bigg[\Bigg(-\lambda\big(\Lambda_h^k\big)^{-1}\mathbb{E}^{\boldsymbol{\mu}_h^0}\Big[V_{h+1}^{k,\rho}(s)\Big]_{\alpha_i^k}\Bigg)_i + \Bigg(\big(\Lambda_h^k\big)^{-1}\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau\Big[\Big[V_{h+1}^{k,\rho}(s_{h+1}^\tau)\Big]_{\alpha_i^k} - \Big[\mathbb{P}_h^0\Big[V_{h+1}^{k,\rho}\Big]_{\alpha_i^k}\Big](s_h^\tau, a_h^\tau)\Big]\Bigg)_i\Bigg]_{i\in[d]}.
\end{aligned} \tag{E.8}$$

For the first term on the RHS of (E.8),

$$\begin{aligned}
&\Bigg|\Bigg\langle\boldsymbol{\phi}(s,a), \Bigg[\Bigg(-\lambda\big(\Lambda_h^k\big)^{-1}\mathbb{E}^{\boldsymbol{\mu}_h^0}\Big[V_{h+1}^{k,\rho}(s)\Big]_{\alpha_i^k}\Bigg)_i\Bigg]_{i\in[d]}\Bigg\rangle\Bigg| \\
&= \Bigg|\sum_{i=1}^d \phi_i(s,a)\mathbf{1}_i^\top(-\lambda)\big(\Lambda_h^k\big)^{-1}\mathbb{E}^{\boldsymbol{\mu}_h^0}\Big[V_{h+1}^{k,\rho}(s)\Big]_{\alpha_i^k}\Bigg| \\
&\leq \lambda\sum_{i=1}^d\sqrt{\phi_i(s,a)\mathbf{1}_i^\top\big(\Lambda_h^k\big)^{-1}\phi_i(s,a)\mathbf{1}_i}\cdot\Bigg\|\mathbb{E}^{\boldsymbol{\mu}_h^0}\Big[V_{h+1}^{k,\rho}(s)\Big]_{\alpha_i^k}\Bigg\|_{(\Lambda_h^k)^{-1}} \\
&\leq \sqrt{\lambda}H\sum_{i=1}^d\sqrt{\phi_i(s,a)\mathbf{1}_i^\top\big(\Lambda_h^k\big)^{-1}\phi_i(s,a)\mathbf{1}_i},
\end{aligned} \tag{E.9}$$

where $\mathbf{1}_i$ is the vector with the $i$-th entry being 1 and else being 0. The first inequality holds due to the Cauchy-Schwarz inequality. For the second term on the RHS of (E.8), given the event $\mathcal{E}$ defined in Lemma C.1 we have,

$$\begin{aligned}
&\left|\left\langle \boldsymbol{\phi}(s,a), \left[\left((\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau\left[\left[V_{h+1}^{k,\rho}(s_{h+1}^\tau)\right]_{\alpha_i^k} - \left[\mathbb{P}_h^0\left[V_{h+1}^{k,\rho}\right]_{\alpha_i^k}\right](s_h^\tau,a_h^\tau)\right]\right)_i\right]_{i\in[d]}\right\rangle\right| \\
&= \left|\sum_{i=1}^d \phi_i(s,a)\mathbf{1}_i^\top(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau\left[\left[V_{h+1}^{k,\rho}(s_{h+1}^\tau)\right]_{\alpha_i^k} - \left[\mathbb{P}_h^0\left[V_{h+1}^{k,\rho}\right]_{\alpha_i^k}\right](s_h^\tau,a_h^\tau)\right]\right| \\
&\le \sum_{i=1}^d \sqrt{\phi_i(s,a)\mathbf{1}_i^\top\left(\Lambda_h^k\right)^{-1}\phi_i(s,a)\mathbf{1}_i}\cdot\left\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau\left[\left[V_{h+1}^{k,\rho}(s_{h+1}^\tau)\right]_{\alpha_i^k} - \left[\mathbb{P}_h^0\left[V_{h+1}^{k,\rho}\right]_{\alpha_i^k}\right](s_h^\tau,a_h^\tau)\right]\right\|_{(\Lambda_h^k)^{-1}} \\
&\le C\cdot dH\sqrt{\chi}\sum_{i=1}^d\sqrt{\phi_i(s,a)\mathbf{1}_i^\top\left(\Lambda_h^k\right)^{-1}\phi_i(s,a)\mathbf{1}_i}, \qquad \text{(E.10)}
\end{aligned}$$

for an absolute constant $C$ independent of $c_\beta$, and $\chi = \log[3(c_\beta+1)dT/p]$. Combining (E.8), (E.9) and (E.10), we have

$$\left\langle\boldsymbol{\phi}(s,a), \boldsymbol{\nu}_h^{\rho,k} - \tilde{\boldsymbol{\nu}}_h^{k,\rho}\right\rangle \le c\cdot dH\sqrt{\chi}\sum_{i=1}^d\sqrt{\phi_i(s,a)\mathbf{1}_i^\top\left(\Lambda_h^k\right)^{-1}\phi_i(s,a)\mathbf{1}_i},$$

for an absolute constant $c$ independent of $c_\beta$. On the other hand, we can similarly deduce $\langle\boldsymbol{\phi}(s,a), \tilde{\boldsymbol{\nu}}_h^{k,\rho} - \boldsymbol{\nu}_h^{\rho,k}\rangle \le c\cdot dH\sqrt{\chi}\sum_{i=1}^d\sqrt{\phi_i(s,a)\mathbf{1}_i^\top(\Lambda_h^k)^{-1}\phi_i(s,a)\mathbf{1}_i}$. Thus, we have

$$\left|\left\langle\boldsymbol{\phi}(s,a), \boldsymbol{\nu}_h^{\rho,k} - \tilde{\boldsymbol{\nu}}_h^{k,\rho}\right\rangle\right| \le c\cdot dH\sqrt{\chi}\sum_{i=1}^d\sqrt{\phi_i(s,a)\mathbf{1}_i^\top\left(\Lambda_h^k\right)^{-1}\phi_i(s,a)\mathbf{1}_i}. \qquad \text{(E.11)}$$

For term $II$ in (E.7), we have

$$\left\langle\boldsymbol{\phi}(s,a), \tilde{\boldsymbol{\nu}}_h^{k,\rho} - \boldsymbol{\nu}_h^{\pi,\rho}\right\rangle = \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)}\left[\mathbb{P}_h V_{h+1}^{k,\rho}\right](s,a) - \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)}\left[\mathbb{P}_h V_{h+1}^{\pi,\rho}\right](s,a).$$

Finally, since $\langle\boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h + \boldsymbol{\nu}_h^{\rho,k}\rangle - Q_h^{\pi,\rho}(s,a) = \langle\boldsymbol{\phi}(s,a), \boldsymbol{\nu}_h^{\rho,k} - \tilde{\boldsymbol{\nu}}_h^{k,\rho} + \tilde{\boldsymbol{\nu}}_h^{k,\rho} - \boldsymbol{\nu}_h^{\pi,\rho}\rangle$, by our choice of $\lambda$ and (E.11) we have

$$\begin{aligned}
&\left|\left\langle\boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h + \boldsymbol{\nu}_h^{\rho,k}\right\rangle - Q_h^{\pi,\rho}(s,a) - \left(\inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)}\left[\mathbb{P}_h V_{h+1}^{k,\rho}\right](s,a) - \inf_{P_h(\cdot|s,a)\in\mathcal{U}_h^\rho(s,a;\boldsymbol{\mu}_h^0)}\left[\mathbb{P}_h V_{h+1}^{\pi,\rho}\right](s,a)\right)\right| \\
&= \left|\left\langle\boldsymbol{\phi}(s,a), \boldsymbol{\nu}_h^{\rho,k} - \tilde{\boldsymbol{\nu}}_h^{k,\rho}\right\rangle\right| \\
&\le c\cdot dH\sqrt{\chi}\sum_{i=1}^d\sqrt{\phi_i(s,a)\mathbf{1}_i^\top\left(\Lambda_h^k\right)^{-1}\phi_i(s,a)\mathbf{1}_i}.
\end{aligned}$$

Finally, to prove this lemma, we only need to show that there exists a choice of absolute value $c_\beta$ so that

$$c'\sqrt{\iota + \log(c_\beta+1)} \le c_\beta\sqrt{\iota}, \qquad \text{(E.12)}$$

where $\iota = \log 3dT/p$. We know $\iota \in [\log 3, \infty)$ by its definition, and $c'$ is an absolute constant independent of $c_\beta$. Therefore we can pick an absolute constant $c_\beta$ which satisfies $c'\sqrt{\log 3 + \log(c_\beta+1)} \le c_\beta\sqrt{\log 3}$. This choice of $c_\beta$ will ensure (E.12) hold for all $\iota \in [\log 3, \infty)$, which finishes the proof. □