Learning Adversarial Low-rank Markov Decision Processes with Unknown Transition and Full-information Feedback

Canzhe Zhao
Shanghai Jiao Tong University
canzhezhao@sjtu.edu.cn

Ruofeng Yang
Shanghai Jiao Tong University
wanshuiyin@sjtu.edu.cn

Baoxiang Wang
The Chinese University of Hong Kong, Shenzhen
bxiangwang@cuhk.edu.cn

Xuezhou Zhang
Boston University
xuezhouz@bu.edu

Shuai Li[*]
Shanghai Jiao Tong University
shuaili8@sjtu.edu.cn

Abstract

In this work, we study the low-rank MDPs with adversarially changed losses in the full-information feedback setting. In particular, the unknown transition probability kernel admits a low-rank matrix decomposition [Uehara et al., 2022], and the loss functions may change adversarially but are revealed to the learner at the end of each episode. We propose a policy optimization-based algorithm POLO, and we prove that it attains the $\widetilde{O}(K^{5/6}A^{1/2}d\ln(1+M)/(1-\gamma)^2)$ regret guarantee, where d is rank of the transition kernel (and hence the dimension of the unknown representations), A is the cardinality of the action space, M is the cardinality of the model class that contains all the plausible representations, and γ is the discounted factor. Notably, our algorithm is oracle-efficient and has a regret guarantee with no dependence on the size of potentially arbitrarily large state space. Furthermore, we also prove an $\Omega(\frac{\gamma^2}{1-\gamma}\sqrt{dAK})$ regret lower bound for this problem, showing that low-rank MDPs are statistically more difficult to learn than linear MDPs in the regret minimization setting. To the best of our knowledge, we present the first algorithm that interleaves representation learning, exploration, and exploitation to achieve the sublinear regret guarantee for RL with nonlinear function approximation and adversarial losses.

1 Introduction

In reinforcement learning (RL), the goal is to learn a (near) optimal policy through the interactions between the learner and the environment, which is typically modeled as the Markov decision processes (MDPs) [Feinberg, 1996]. When the state and action spaces are finite, many works have established the minimax (near) optimal regret guarantees for MDPs with finite horizon [Azar et al., 2017] and MDPs with infinite horizon [Tossou et al., 2019, He et al., 2021b]. In real applications of RL, however, the state and action spaces may be arbitrarily large and even infinite, which may lead to the curse of dimensionality. To tackle this issue, a common approach is *function approximation*, which

---

[*]Corresponding author.

approximates the value functions of given policies with the leverage of feature mappings. Assuming that the feature mapping which embeds the state-action pairs to a low dimensional embedding space is known, RL with linear function approximation has been well studied recently. In particular, linear mixture MDPs [Ayoub et al., 2020] and linear MDPs [Jin et al., 2020b] are the two models of RL with linear function approximation that have been extensively studied. Notably, their (near) optimal regret guarantees are established by Zhou et al. [2021] and He et al. [2023] respectively. Nevertheless, in scenarios with complex and large-scale data, attaining the true underlying feature mappings might be unrealistic, and thus representation learning is needed. Many empirical works have shown that representation learning can accelerate the sample and computation efficiency of RL [Silver et al., 2018, Laskin et al., 2020, Yang and Nachum, 2021, Stooke et al., 2021, Schwarzer et al., 2021, Xie et al., 2022]. However, representation learning is provably more difficult in RL and other sequential decision-making problems than in non-sequential and non-interactive problems (*e.g.*, supervised learning) [Du et al., 2020, Wang et al., 2021a, Weisz et al., 2021, Uehara et al., 2022]. To pursue sample-efficient RL in the presence of representation learning, recent works have made initial attempts to study the theoretical guarantees of representation learning in RL under the fixed or stochastic loss functions [Uehara et al., 2022, Zhang et al., 2022].

In practice, however, it might be stringent to assume that the loss functions are stochastic. To tackle this issue, Even-Dar et al. [2009] and Yu et al. [2009] propose the first algorithms with provably theoretical guarantees that can handle adversarial MDPs, where the loss functions may change adversarially in each episode. Subsequently, most of the works in this line of research focus on learning tabular MDPs with adversarial loss functions [Neu et al., 2010a,b, 2012, Arora et al., 2012, Zimin and Neu, 2013, Dekel and Hazan, 2013, Dick et al., 2014, Rosenberg and Mansour, 2019a,b, Jin and Luo, 2020, Jin et al., 2020a, Shani et al., 2020b, Chen et al., 2021, Ghasemi et al., 2021, Rosenberg and Mansour, 2021, Jin et al., 2021b, Dai et al., 2022, Chen et al., 2022]. To learn adversarial MDPs with large state and action spaces, some recent works study RL with adversarial loss functions and linear function approximation [Cai et al., 2020, Neu and Olkhovskaya, 2021, Luo et al., 2021b,a, He et al., 2022, Zhao et al., 2023, Kong et al., 2023]. However, all these existing works assume that the state-action feature representations are known. As aforementioned, in complex and high-dimensional environments, the application of these algorithms may be still hindered due to the potential difficulty of knowing the true feature mappings a priori. Therefore, the following question naturally remains open:

Can we devise an algorithm to simultaneously tackle the representation learning and adversarially changed loss functions in reinforcement learning?

In this work, we give an affirmative answer to the above question in the setting of adversarial low-rank MDPs with full-information feedback. Specifically, in this problem, the unknown transition probability kernel admits a low-rank matrix decomposition but the true representations regarding the transitions are not known a priori. Meanwhile, the loss functions are arbitrarily chosen by an adversary across episodes and are revealed to the learner after each episode.

To solve this problem, we propose a policy optimization-based algorithm, which we call **P**olicy **O**ptimization for **LO**w-rank MDPs (POLO). Specifically, our POLO algorithm obtains an $\widetilde{O}(K^{5/6}A^{1/2}d\ln(1+M)/(1-\gamma)^2)$ regret guarantee for adversarial low-rank MDPs in the full-information feedback setting and is oracle-efficient. Algorithmically, POLO follows similar ideas of optimistic policy optimization methods in that it first constructs optimistic value function estimates and then runs online mirror descent (OMD) over the optimistic value estimates to deal with the adversarially changed loss functions [Shani et al., 2020b, Cai et al., 2020, He et al., 2022, Chen et al., 2022]. However, in the presence of representation learning, the exploration and exploitation needed to learn the adversarial MDPs are more difficult than them in the tabular case [Shani et al., 2020b, Chen et al., 2022] and in the linear case [Cai et al., 2020, He et al., 2022], where the exploration and exploitation can be essentially well balanced by directly modifying the algorithms studying stochastic MDPs in the tabular and linear cases. Concretely, to learn stochastic low-rank MDPs, previous works perform maximum likelihood estimation (MLE) over the experienced transitions [Agarwal et al., 2020, Uehara et al., 2022, Zhang et al., 2022]. Though the balance of representation learning, exploration, and exploitation can be simultaneously handled by the algorithms in these works, these algorithms intrinsically have no regret guarantees but only sample complexity guarantees even in the stochastic case, since these algorithms need to take actions uniformly at certain steps

Table 1: Comparisons of regret bounds with most related works studying adversarial RL with function approximation under unknown transitions. K is the number of episodes, d is the ambient dimension of the feature mapping, γ is the discounted factor for infinite-horizon MDPs, and S, A, and M are the cardinality of the state space, action space, and model class, respectively. Note that the dependence on γ is not strictly comparable since some works originally studying finite-horizon MDPs and these results are translated into results for infinite-horizon MDPs by substituting horizon length H with $\Theta(1/(1-\gamma))$. The column of "unknown features" indicates whether the algorithm can work in the case when no true feature mappings are known a priori.

Algorithm	Model	Feedback	Regret	Unknown Features
OPPO [Cai et al., 2020]	Linear Mixture MDPs	Full-information	$\widetilde{O}\left(d\sqrt{K}/(1-\gamma)^2\right)$	✗
POWERS [He et al., 2022]	Linear Mixture MDPs	Full-information	$\widetilde{O}\left(d\sqrt{K}/(1-\gamma)^{3/2}\right)$	✗
LSUOB-REPS Zhao et al. [2023]	Linear Mixture MDPs	Bandit Feedback	$\widetilde{O}\left(dS^2\sqrt{K} + \sqrt{\frac{SAK}{(1-\gamma)}}\right)$	✗
Luo et al. [2021a]	Linear MDPs	Bandit Feedback	$\widetilde{O}\left(d^2 K^{14/15}/(1-\gamma)^4\right)$	✗
Dai et al. [2023]	Linear MDPs	Bandit Feedback	$\widetilde{O}\left(\frac{A^{1/9}d^{2/3}K^{8/9}}{(1-\gamma)^{20/9}}\right)$	✗
PO-LSBE Sherman et al. [2023]	Linear MDPs	Bandit Feedback	$\widetilde{O}\left(\frac{dK^{6/7}}{(1-\gamma)^2} + \frac{d^{3/2}K^{5/7}}{(1-\gamma)^4}\right)$	✗
GLAP Kong et al. [2023]	Linear MDPs	Bandit Feedback	$\widetilde{O}\left(d^{7/5}H^{12/5}K^{4/5}\right)$	✗
OPPO+ Zhong and Zhang [2023]	Linear MDPs	Full-information	$\widetilde{O}\left(\frac{d^{3/4}K^{3/4}+d^{5/2}\sqrt{K}}{(1-\gamma)^2}\right)$	✗
POLO (Ours)	Low-rank MDPs	Full-information	$\widetilde{O}\left(\frac{K^{5/6}A^{1/2}d\ln(1+M)}{(1-\gamma)^2}\right)$ $\Omega\left(\frac{\gamma^2}{1-\gamma}\sqrt{dAK}\right)$	✔

in each episode (*cf.*, Lemma 9 of Uehara et al. [2022]).[2] Hence, a straightforward adaption of their methods from the stochastic setting to the adversarial setting will fail to learn adversarial low-rank MDPs. To cope with this issue, we carefully devise an algorithm with a *doubled exploration and exploitation* scheme, which interleaves (a) the exploration over transitions required in representation learning; and (b) the exploration and exploitation suggested by the policy optimization. To this end, our algorithm adopts a mixed roll-out policy, which consists of a uniformly explorative policy and a policy optimized by OMD. Through carefully tuning the hyper-parameter of the mixing coefficient used in our mixed policy, we can avoid pulling actions uniformly at random to conduct exploration in each episode and only conduct uniform exploration at a certain fraction of all the episodes (see Section 3.1 for details). Besides, unlike tabular and linear (mixture) MDPs, it is in general hard to achieve the point-wise optimism for each state-action pair. Therefore, depart from previous methods [Shani et al., 2020b, Cai et al., 2020, He et al., 2022] conducting policy optimization in the *true* model (*i.e.*, the transition kernel characterized by the true representation), our algorithm conducts policy optimization in the fixed *learned* model with the epoch-based model update, which enables a new analysis scheme that only requires a *near optimism* at the initial state s_0 (see Section 3.2 for

---

[2]With the leverage of the common explore-then-commit (ETC) style conversion, the modified versions of these algorithms can obtain sublinear regret in the setting of low-rank MDPs with stochastic loss functions, but this conversion is still not able to deal with adversarial loss functions.

details). Also, we prove a regret lower bound of order $\Omega\left(\frac{\gamma^2}{1-\gamma}\sqrt{dAK}\right)$ for low-rank MDPs with fixed loss functions, which thus also serves as a regret lower bound for our problem and indicates that low-rank MDPs are statistically more difficult to learn than linear MDPs in the regret minimization setting. To the best of our knowledge, this work makes the first step to establish an algorithm with a sublinear regret guarantee for adversarial low-rank MDPs, which permits RL with both nonlinear function approximation and adversarial loss functions. The concrete comparisons between the results of this work and those of previous works are summarized in Table 1.

1.1 Additional Related Works

RL with Function Approximation Significant advances have emerged in RL with function approximation to cope with the curse of dimensionality in arbitrarily large state space or action space. In general, these results fall into two categories. The first category studies RL with linear function approximation, including linear MDPs [Yang and Wang, 2019, Jin et al., 2020b, Du et al., 2020, Zanette et al., 2020, Wang et al., 2020, 2021b, He et al., 2021a, Hu et al., 2022, He et al., 2023] and linear mixture MDPs [Ayoub et al., 2020, Zhang et al., 2021, Zhou et al., 2021, He et al., 2021a, Zhou and Gu, 2022, Wu et al., 2022, Min et al., 2022, Zhao et al., 2023]. Remarkably, He et al. [2023] and Zhou et al. [2021] obtain the nearly minimax optimal regret $\widetilde{O}(d\sqrt{H^3K})$ in linear MDPs and linear mixture MDPs respectively when the loss functions are fixed or stochastic. The other category studies RL with general function approximation. Amongst these works, Jiang et al. [2017], Dann et al. [2018], Sun et al. [2019], Du et al. [2019], and Jin et al. [2021a] study the MDPs satisfying the low Bellman-rank assumption, which assumes the Bellman error matrix has a low-rank factorization. Also, Du et al. [2021] consider a similar but slightly more general assumption termed as bounded bilinear rank. Besides, Russo and Roy [2013], Wang et al. [2020], Jin et al. [2021a], and Ishfaq et al. [2021] study low Eluder dimension assumption, which is originally proposed to characterize the complexity of function classes for bandit problems.

Representation learning in RL arises when the feature mapping that embeds the state-action pairs in RL with linear function approximation is no longer known a priori. Such a problem is typically studied in the setting of low-rank MDPs, which does not assume the feature mapping of state-action pairs is known. Consequently, the setting of low-rank MDPs strictly generalizes the setting of linear MDPs, but at the cost of being more difficult to learn due to potential nonlinear function approximation induced by representation learning. In this line of research, algorithms with provably sample complexity guarantees have been developed in both model-based methods [Agarwal et al., 2020, Ren et al., 2022, Uehara et al., 2022] and model-free methods [Modi et al., 2021, Zhang et al., 2022], respectively. The model-based algorithms of Agarwal et al. [2020], Ren et al. [2022] and Uehara et al. [2022] learn the representation from a given model class of transition probability kernels. In contrast, the model-free methods do not require model learning but may bear some limitations. In particular, Modi et al. [2021] assume the MDPs satisfying the minimal reachability assumption, and the sample complexity of the algorithm of Zhang et al. [2022] only holds for a special class of low-rank MDPs called block MDPs. Besides, representation learning in Markov games has also been investigated recently [Ni et al., 2022].

RL with Adversarial Losses Recent years have witnessed significant advances in learning RL with adversarial losses in the tabular case [Neu et al., 2010a,b, 2012, Arora et al., 2012, Zimin and Neu, 2013, Dekel and Hazan, 2013, Dick et al., 2014, Rosenberg and Mansour, 2019a,b, Jin and Luo, 2020, Jin et al., 2020a, Shani et al., 2020b, Chen et al., 2021, Ghasemi et al., 2021, Rosenberg and Mansour, 2021, Jin et al., 2021b, Dai et al., 2022, Chen et al., 2022]. When it comes to the setting of linear function approximation, various policy optimization-based methods have been established to solve adversarial linear mixture MDPs [Cai et al., 2020, He et al., 2022] and adversarial linear MDPs [Luo et al., 2021a,b, Dai et al., 2023, Sherman et al., 2023, Zhong and Zhang, 2023]. The other line of works studies RL with linear function approximation and adversarial losses using occupancy measure-based methods [Neu and Olkhovskaya, 2021, Zhao et al., 2023, Kong et al., 2023]. To the best of our knowledge, however, there are no works in existing literature studying RL with both nonlinear function approximation and adversarial loss functions.

2 Preliminaries

We consider episodic infinite horizon low-rank MDPs with adversarial loss functions, the preliminaries of which are introduced as follows.

Episodic Infinite-horizon Adversarial MDPs   An episodic infinite horizon adversarial MDP is denoted by a tuple $(\mathcal{S}, \mathcal{A}, P^\star, \{\ell_k\}_{k=1}^K, \gamma, d_0)$,[3] where $\mathcal{S}$ is the state space (with potentially infinitely many states), $\mathcal{A}$ is the finite action space with cardinality $|\mathcal{A}| = A$, $P^\star : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability kernel such that $P^\star(s' \mid s, a)$ is the probability of transferring to state s' from state s after executing action a, $\gamma \in [0, 1)$ is the discount factor, $d_0 \in \Delta(\mathcal{S})$ is the initial distribution over the state space, and $\ell_k : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the loss function of episode k chosen by the adversary. For the ease of exposition, we assume d_0 is known.

In this work, we consider a special class of MDPs called *low-rank MDPs* [Agarwal et al., 2020, Uehara et al., 2022, Zhang et al., 2022]. Specifically, instead of assuming a known true feature mapping, low-rank MDPs only assume that the transition probability kernel $P^\star$ admits a low-rank decomposition, with the formal definition given as follows.

Definition 2.1 (Low-rank MDPs). *An MDP is a low-rank MDP if there exist two feature embedding functions $\phi^\star : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\mu^\star : \mathcal{S} \to \mathbb{R}^d$ such that for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, $P^\star(s' \mid s, a) = \mu^\star(s')^\top \phi^\star(s, a)$, where $\|\phi^\star(s, a)\|_2 \leq 1$ and for any function $g : \mathcal{S} \to [0, 1], \left\| \int \mu^\star(s) g(s) \mathrm{d}(s) \right\|_2 \leq \sqrt{d}$.*

Note that the regularity assumption imposed over $\phi^\star$ and $\mu^\star$ is only for the purpose of normalization.

Function Approximation   When the state space is arbitrarily large, function approximation is usually considered to permit sample-efficient learning for MDPs. Since the true feature mapping of state-action pairs is not known a priori in the low-rank MDPs, to make this problem tractable, we assume the access to a *realizable* model class as previous works [Agarwal et al., 2020, Uehara et al., 2022], detailed in the following.

Assumption 2.1. *There exists a known model class $\mathcal{M} = \{(\mu, \phi) : \mu \in \Psi, \phi \in \Phi\}$ satisfying (a) $\|\phi(s, a)\|_2 \leq 1$ and $\int \mu^\top(s') \phi(s, a) \mathrm{d}(s') = 1$ for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, $\mu \in \Psi$, $\phi \in \Phi$; and (b) $\left\| \int \mu(s) g(s) \mathrm{d}(s) \right\|_2 \leq \sqrt{d}$ for any function $g : \mathcal{S} \to [0, 1]$. Moreover, it holds that $\mu^\star \in \Psi$ and $\phi^\star \in \Phi$.*

Throughout this paper, for the sake of brevity, we assume that the cardinality of Ψ and Φ are finite, meaning that $\mathcal{M}$ also has bounded cardinality $M = |\mathcal{M}|$. However, note that extending the analyses to the function classes with infinite cardinality but bounded statistical complexity (*e.g.*, VC dimension) is not technically difficult.

Interaction Protocol   We now introduce the interaction protocol between the learner and the environment. To begin with, denote by $d_P^\pi(s, a) = (1 - \gamma) \sum_{\tau=0}^\infty \gamma^\tau d_{P,\tau}^\pi(s, a)$ the state-action occupancy distribution, where $d_{P,\tau}^\pi(s, a)$ is the probability of visiting (s, a) at step τ under policy π and transition P. With a slight abuse of notation, let $d_P^\pi(s) = \sum_{a \in \mathcal{A}} d_P^\pi(s, a)$ be the state occupancy distribution, denoting the probability of visiting state s under π and P.

Ahead of time, an MDP is decided by the environment, and only the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are revealed to the learner. Meanwhile, the adversary secretly chooses K loss functions $\{\ell_k\}_{k=1}^K$, each of which will be used in one episode. The interaction will proceed in K episodes. At the beginning of episode k, the learner chooses a stochastic policy $\pi_k : \mathcal{S} \times \mathcal{A} \to [0, 1]$, where $\pi_k(a \mid s)$ is probability of taking a at state s. Starting from an initial state $s_0 \sim d_0$, the learner repeatedly executes policy π_k until reaching the termination. After episode k is terminated, the learner observes a trajectory $\{(s_{k,\tau}, a_{k,\tau})\}_\tau$ as well as the loss function ℓ_k.

For each episode k and each state-action pair (s, a), the state-action value $Q_k^\pi(s, a)$ is defined as $Q_k^\pi(s, a) = \mathbb{E}\left[\sum_{\tau=0}^\infty \gamma^\tau \ell_k(s_{k,\tau}, a_{k,\tau}) \Big| \pi, P^\star, (s_{k,0}, a_{k,0}) = (s, a)\right]$. Also define $V_k^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot | s)}[Q_k^\pi(s, a)]$ and $V_k^\pi = \mathbb{E}_{s_0 \sim d_0}[V_k^\pi(s_0)]$. The learning objective is to minimize the *pseudo regret* with respect to $\pi^\star$, defined as

$$\mathcal{R}_K = \mathbb{E}\left[\sum_{k=1}^K \left(V_k^{\pi_k} - V_k^{\pi^\star}\right)\right],$$

---

[3]Though we focus on episodic infinite-horizon MDPs in this work, note that it is not technically difficult to extend the analyses in this work to the case of episodic finite-horizon MDPs.

where the expectation is taken over the potential randomness of the algorithm, $\pi^\star \in \arg\min_{\pi \in \Pi} \sum_{k=1}^K V_k^\pi$ is the fixed optimal policy in hindsight and Π is the set of stochastic policies.

3 Algorithm

This section presents our POLO algorithm with the pseudocode illustrated in Algorithm 1. At a high level, POLO leverages a mixed roll-out policy to conduct doubled exploration and exploitation, *i.e.*, (a) the exploration over transitions required by representation learning; and (b) the exploration and exploitation over adversarially changed loss functions required by the policy optimization (Section 3.1). To deal with the issue that only the *near optimism* at the initial state s_0 is available in low-rank MDPs, POLO conducts policy optimization in fixed *learned* models with the epoch-based model update, which enables a new analysis scheme (Section 3.2).

3.1 Doubled Exploration and Exploitation

Let $\tilde{\pi}_k$ be the policy of episode k computed by policy optimization (*cf.*, Eq. (3)). At the beginning of episode k, our algorithm first collects a state $s_k \sim d_{P^\star}^{\tilde{\pi}_k}$ (Line 6) by invoking a geometric sampling *roll-in* procedure [Kakade and Langford, 2002, Agarwal et al., 2021, Uehara et al., 2022]. Starting from an initial state $s_0 \sim d_0$, at each step τ, this roll-in procedure will terminate and return state s_τ with probability $1 - \gamma$, and otherwise will take action $a_\tau \sim \tilde{\pi}_k(\cdot \mid s_\tau)$ and transfer to the next state $s_\tau \sim P^\star(\cdot \mid s_\tau, a_\tau)$.

Then our algorithm will further interact with the environment in successive two steps after collecting $s_k \sim d_{P^\star}^{\tilde{\pi}_k}$ (Line 7 - Line 11) like previous works studying low-rank MDPs [Agarwal et al., 2020, Uehara et al., 2022, Zhang et al., 2022]. One of the main differences between previous algorithms and ours lies in how to deal with exploration and exploitation when interacting with the environment. In the case of stochastic low-rank MDPs, the key in the analysis is to ensure the (near) optimism of the optimal policy $\pi^\star$ in the learned model, which essentially requires to bound the performance gap $\widehat{V}_k^{\pi^\star} - V_k^{\pi^\star}$ by relating it with the model error regarding $\pi^\star$, *i.e.*, $\mathbb{E}_{(s,a) \sim d_{P^\star}^{\pi^\star}}[\|\hat{P}_k(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1]$, where $\hat{P}_k$ is the MLE solution defined in Eq. (1). However, this model error is not directly controllable, as the algorithm does not know the optimal policy $\pi^\star$ and can not collect data following $\pi^\star$ to bound the model error. Fortunately, by empirical process theory [Geer, 2000, Zhang, 2006], the model error $\mathbb{E}_{(s,a) \sim \rho_k}[\|\hat{P}_k(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1]$ regarding the executed policies $\{\tilde{\pi}_i\}_{i=1}^k$ is bounded, where $\rho_k(s,a) = \frac{1}{k}\sum_{i=1}^k d_{P^\star}^{\tilde{\pi}_i}(s,a)$. Therefore, previous works [Agarwal et al., 2020, Uehara et al., 2022, Zhang et al., 2022] ensure the optimism of the optimal policy $\pi^\star$ by applying importance weighting to change the measure from $d_{P^\star}^{\pi^\star}$ to ρ_k when bounding the model error. The complication is that to control the ratio $\pi^\star(a \mid s)/\tilde{\pi}_i(a \mid s)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $i \in [K]$ in importance weighting, the algorithms in previous works take actions from the uniform distribution $U(\mathcal{A})$ over action space $\mathcal{A}$, which intuitively can be seen as conducting exploration over transitions required by representation learning. Consequently, though these algorithms enjoy excellent sample complexities, they intrinsically do not have regret guarantees due to the uniform exploration over action space, even in the stochastic setting. Moreover, to learn adversarial low-rank MDPs, it is required to take actions adaptively according to the observed loss functions in previous episodes instead of uniformly taking actions. To address this "conflict" so as to learn adversarial low-rank MDPs, we propose to use a mixed roll-out policy to interleave (a) the exploration over transitions required by representation learning; and (b) the exploration and exploitation over the adversarial loss functions by policy optimization, which we call *doubled exploration and exploitation* and is pivotal to achieving our regret bound as we will shortly see. Formally, our algorithm will conduct the exploration over the transitions with probability ξ and execute policy $\tilde{\pi}_k$ optimized by OMD with probability $1 - \xi$ respectively, as shown in Line 7 - Line 11.

After interacting with the environment, the newly collected data in these two steps will be used to update the datasets (Line 14), and the empirical transition $\widehat{P}_k$ will be updated by performing MLE over the updated datasets by solving (Line 16)

$$\left(\widehat{\mu}_k, \widehat{\phi}_k\right) = \underset{(\mu,\phi) \in \mathcal{M}}{\arg\max} \, \mathbb{E}_{\mathcal{D}_k \cup \mathcal{D}_k'}\left[\ln \mu^\top(s')\phi(s,a)\right], \tag{1}$$

where we denote $\mathbb{E}_{\mathcal{D}}[f(s,a,s')] = \frac{1}{|\mathcal{D}|}\sum_{(s,a,s') \in \mathcal{D}} f(s,a,s')$.

Algorithm 1 Policy Optimization for Low-rank MDPs (POLO)

1: **Input:** Mixing coefficient ξ, epoch length L, regularization coefficients $\{\lambda_k\}_{k=1}^K$, bonus coefficients $\{\alpha_k\}_{k=1}^K$, model class $\mathcal{M}$, number of episodes K, learning rate η.
2: **Initialization:** Set $\mathcal{D}_0 = \emptyset, \mathcal{D}'_0 = \emptyset$.
3: **for** $i = 1, 2, \ldots, \lceil K/L \rceil$ **do**
4:     Set $k_i = (i-1)L + 1$ and $\tilde{\pi}_{k_i}(\cdot \mid s)$ to be uniform for any $s \in \mathcal{S}$.
5:     **for** $k = k_i, k_i + 1, \ldots, k_i + L - 1$ **do**
6:         Sample s_k from $d_{P^\star}^{\tilde{\pi}_k}$.
7:         Sample $c_k \sim \text{Ber}(1 - \xi)$.
8:         **if** $c_k = 1$ **then**
9:             Sample $a_k \sim \tilde{\pi}_k(\cdot \mid s_k), s'_k \sim P^\star(\cdot \mid s_k, a_k), a'_k \sim \tilde{\pi}_k(\cdot \mid s'_k), s''_k \sim P^\star(\cdot \mid s'_k, a'_k)$.
10:         **else**
11:             Sample $a_k \sim U(\mathcal{A}), s'_k \sim P^\star(\cdot \mid s_k, a_k), a'_k \sim U(\mathcal{A}), s''_k \sim P^\star(\cdot \mid s'_k, a'_k)$.
12:         **end if**
13:         Observe the loss function ℓ_k.
14:         Update datasets $\mathcal{D}_k = \mathcal{D}_{k-1} \cup \{(s_k, a_k, s'_k)\}, \mathcal{D}'_k = \mathcal{D}'_{k-1} \cup \{(s'_k, a'_k, s''_k)\}$.
15:         **if** $k = k_i$ **then**
16:             Set the empirical transition $\widehat{P}_k(s' \mid s, a) = \widehat{\mu}_k(s')^\top \widehat{\phi}_k(s, a), \forall(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, via solving Eq. (1).
17:             Update the empirical covariance matrix $\widehat{\Sigma}_k = \sum_{(s,a) \in \mathcal{D}_k} \widehat{\phi}_k(s,a)\widehat{\phi}_k(s,a)^\top + \lambda_k I$.
18:             Set the bonus function $\widehat{b}_k(s,a) := \min(\alpha_k \|\widehat{\phi}_k(s,a)\|_{\widehat{\Sigma}_k^{-1}}, 2)/(1-\gamma), \forall(s,a) \in \mathcal{S} \times \mathcal{A}$.
19:         **else**
20:             Set the empirical transition $\widehat{P}_k = \widehat{P}_{k_i}$ and bonus function $\widehat{b}_k = \widehat{b}_{k_i}$.
21:         **end if**
22:         Compute $\widehat{Q}_k^{\tilde{\pi}_k}(\cdot, \cdot) = \text{Policy-Evaluation}(\widehat{P}_k, \ell_k - \widehat{b}_k, \tilde{\pi}_k)$.
23:         Update policy $\tilde{\pi}_{k+1}(\cdot \mid \cdot) \propto \tilde{\pi}_k(\cdot \mid \cdot) \exp(-\eta \widehat{Q}_k^{\tilde{\pi}_k}(\cdot, \cdot))$.
24:     **end for**
25: **end for**

3.2 Policy Optimization in Fixed Learned Models

It remains to compute the policy $\tilde{\pi}_{k+1}$ to be used in the next episode. To this end, we resort to the canonical OMD framework to perform policy optimization like previous methods [Shani et al., 2020a, Cai et al., 2020, He et al., 2022]. However, previous OMD-based policy optimization methods for tabular and linear (mixture) MDPs [Shani et al., 2020a, Cai et al., 2020, He et al., 2022] critically depend on the point-wise optimism for each state-action pair, *i.e.*, $\widehat{Q}_k^{\tilde{\pi}_k}(s, a) \leq \ell_k(s, a) + \gamma[P^\star \widehat{V}_k^{\tilde{\pi}_k}](s, a)$, to enable the decomposition (*cf.*, Lemma 1 by Shani et al. [2020a])

$$\widehat{V}_k^{\tilde{\pi}_k}(s_0) - V_k^{\pi^\star}(s_0) = \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \left\langle \tilde{\pi}_k(\cdot \mid s_\tau) - \pi^\star(\cdot \mid s_\tau), \widehat{Q}_k^{\tilde{\pi}_k}(s_\tau, \cdot)\right\rangle \middle| \pi^\star, P^\star, s_0\right]$$
$$+ \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \left(\widehat{Q}_k^{\tilde{\pi}_k}(s_\tau, a_\tau) - \ell_k(s_\tau, a_\tau) - \gamma\left[P^\star \widehat{V}_k^{\tilde{\pi}_k}\right](s_\tau, a_\tau)\right) \middle| \pi^\star, P^\star, s_0\right],$$

where $\widehat{V}_k^{\tilde{\pi}_k}$ and $\widehat{Q}_k^{\tilde{\pi}_k}$ are the state value and state-action value functions of $\tilde{\pi}_k$ on $(\widehat{P}_k, \ell_k - \widehat{b}_k)$ with $\widehat{b}_k$ as some bonus function and the expectation is taken over the randomness of sampling $a_\tau \sim \pi^\star(\cdot \mid s_\tau)$ and $s_{\tau+1} \sim P^\star(\cdot \mid s_\tau, a_\tau)$. The summation of the first term in the above display is contributed by competing with the optimal policy $\pi^\star$ in the *true* model $P^\star$ and can be bounded by common OMD analysis, which thus can be regarded as conducting policy optimization in the *true* model. The point-wise optimism guarantees that the second term is less than or equal to 0.

Nevertheless, in low-rank MDPs, due to the unknown representation, it is generally hard to obtain the above point-wise optimism, which leaves the second optimism term unbounded. To cope with this issue, we instead consider the following decomposition:

$$\widehat{V}_k^{\tilde{\pi}_k}(s_0) - V_k^{\pi^\star}(s_0)$$

$$=\widehat{V}_k^{\tilde{\pi}_k}(s_0) - \widehat{V}_k^{\pi^\star}(s_0) + \widehat{V}_k^{\pi^\star}(s_0) - V_k^{\pi^\star}(s_0)$$

$$=\mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \left\langle \tilde{\pi}_k(\cdot \mid s_{k,\tau}) - \pi^\star(\cdot \mid s_{k,\tau}), \widehat{Q}_k^{\tilde{\pi}_k}(s_{k,\tau}, \cdot) \right\rangle \,\middle|\, \pi^\star, \widehat{P}_k, s_0\right] + \widehat{V}_k^{\pi^\star}(s_0) - V_k^{\pi^\star}(s_0), \quad (2)$$

where the first term is contributed by competing against the optimal policy $\pi^\star$ in the *learned* model $\widehat{P}_k$ and can be seen as conducting policy optimization in *learned* models. This decomposition will be amenable as long as we can achieve a *near optimism* at the initial state s_0, *i.e.*, $\widehat{V}_k^{\pi^\star}(s_0) - V_k^{\pi^\star}(s_0) \lesssim 0$, which turns out to be feasible for low-rank MDPs [Uehara et al., 2022]. However, there remains one more caveat. The first term in Eq. (2) is now no longer directly bounded by OMD analysis, due to the local update nature of OMD-based policy optimization at each state and the state occupancy distribution $d_{\widehat{P}_k}^{\pi^\star}$ now varies across different episodes. To address this issue, Algorithm 1 adopts an epoch-based transition update, in which one epoch has L episodes and the model is only updated at the first episode in one epoch (Line 15 - Line 20).[4] Concretely, Algorithm 1 sets $\widehat{P}_k = \widehat{P}_{k_i}$ and $\widehat{b}_k = \widehat{b}_{k_i}$, where k_i is the first episode of the epoch to which the episode k belongs. In this manner, the learned model is *fixed* in one epoch, and thus the regret of dealing with the adversarial loss functions by competing against the optimal policy $\pi^\star$ can be bounded in one epoch. Subsequently, at the end of episode k, our algorithm first computes the optimistic value estimate $\widehat{Q}_k^{\tilde{\pi}_k}$ for current policy $\tilde{\pi}_k$ under $\widehat{P}_k$ together with the bonus-enhanced loss functions $\ell_k - \widehat{b}_k$ by policy evaluation (Line 22). Note that this boils down to planning in the setting of linear MDPs for given features in the learned model and this can be done computationally efficiently [Jin et al., 2020b]. Then the policy is updated by solving

$$\tilde{\pi}_{k+1}(\cdot \mid s) \in \underset{\pi(\cdot \mid s) \in \Delta(\mathcal{A})}{\arg\min}\ \eta \left\langle \pi(\cdot \mid s), \widehat{Q}_k^{\tilde{\pi}_k}(s, \cdot) \right\rangle + D_F(\pi(\cdot \mid s), \tilde{\pi}_k(\cdot \mid s)), \quad (3)$$

where $\eta > 0$ is the learning rate to be tuned later and $D_F(x, y) = F(x) - F(y) - \langle x - y, \nabla F(y)\rangle$ is the Bregman divergence induced by the regularizer F. With $F(\pi(\cdot \mid s)) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \ln \pi(a \mid s)$ as the negative entropy, the closed-form solution to the above display is shown in Line 23, which can be regarded as a kind of soft policy improvement.

4 Analysis

4.1 Regret Upper Bound

The regret upper bound of our POLO algorithm for learning adversarial low-rank MDPs is guaranteed by the following theorem.

Theorem 4.1. *Suppose $K > d^6 A^3/(1 - \gamma)^6$. For any adversarial low-rank MDP satisfying Definition 2.1, by setting the epoch length $L = K^{1/2} A^{-1/2} d^{-1} \xi (1 - \gamma)$, learning rate $\eta = (1 - \gamma)\sqrt{\ln A/(2L)}$, bonus coefficient $\alpha_k = O(\sqrt{\gamma(A/\xi + d^2)\ln(Mk/\delta)})$, regularization coefficient $\lambda_k = O(d \ln(Mk/\delta))$, mixing coefficient $\xi = K^{-1/6} A^{1/2} d/(1 - \gamma)$, and $\delta = 1/K$, then the regret of Algorithm 1 is upper bounded by*

$$\mathcal{R}_K = O\left(\frac{K^{\frac{5}{6}} A^{\frac{1}{2}} d \ln\left(1 + AMK^2\right)}{(1 - \gamma)^2}\right).$$

Remark 4.1. *Ignoring the dependence on all logarithmic factors but M, the regret upper bound can be simplified as $\widetilde{O}(K^{5/6} A^{1/2} d \ln(1 + M)/(1 - \gamma)^2)$. Comparing with the regret lower bound $\Omega(\frac{\gamma^2}{1-\gamma}\sqrt{dAK})$ in Theorem 4.2, the regret upper bound in Theorem 4.1 matches in A up to a logarithmic factor but looses in factors of K and d. Also, note that when K is large enough such that ξ and L can be chosen as $\xi = K^{-1/6} A^{1/3} d^{2/3}/(1 - \gamma)$ and $L = K^{1/2} A^{-1} d^{-2} \xi (1 - \gamma) = K^{1/3} A^{-2/3} d^{-4/3} \geq 1$, meaning that $K \geq d^4 A^2$, the regret upper bound can be further optimized to $\widetilde{O}(K^{5/6} A^{1/3} d^{2/3} \ln(1 + M)/(1 - \gamma)^2)$. Note that this does not conflict with the regret lower bound in Section 4.3 since the magnitude of this upper bound is still larger than that of the regret lower bound as long as $K \geq A^{1/2} d^{-1/2} \gamma^6 (1 - \gamma)^3$.*

---

[4]Throughout this paper, we suppose for simplicity that the number of episodes K is divisible by the epoch length L considered.

4.2 Proof of Regret Upper Bound

We now present the proof of Theorem 4.1. To begin with, recall that in each episode k, after state s_k is sampled from $d_{P^\star}^{\tilde\pi_k}$, the actual roll-out policy will be $\pi_k(\cdot \mid s) = \xi \cdot U(\mathcal{A}) + (1 - \xi) \cdot \tilde\pi_k(\cdot \mid s)$. Therefore, it holds that

$$
\begin{aligned}
\mathcal{R}_K &= \mathbb{E}\left[\sum_{k=1}^K \left(V_k^{\pi_k} - V_k^{\pi^\star}\right)\right] \\
&= \mathbb{E}\left[\sum_{k=1}^K \mathbb{I}\{c_k = 1\}\left(V_k^{\pi_k} - V_k^{\pi^\star}\right) + \mathbb{I}\{c_k = 0\}\left(V_k^{\pi_k} - V_k^{\pi^\star}\right)\right] \\
&\le \mathbb{E}\left[\sum_{k=1}^K \left(V_k^{\tilde\pi_k} - V_k^{\pi^\star}\right)\right] + \frac{\xi K}{(1 - \gamma)},
\end{aligned}
\tag{4}
$$

where the inequality is due to that $\sum_{\tau=0}^\infty \gamma^\tau \ell_k\left(s_\tau, a_\tau\right) \in [0, 1/(1 - \gamma)]$ holds for any episode k and any trajectory $\{(s_\tau, a_\tau)\}_{\tau=0}^\infty$. We now turn to bound the first term in Eq. (4) by decomposing it into the following three terms

$$
\mathbb{E}\left[\sum_{k=1}^K \left(V_k^{\tilde\pi_k} - V_k^{\pi^\star}\right)\right] = \mathbb{E}\left[\underbrace{\sum_{k=1}^K \left(V_k^{\tilde\pi_k} - \widehat{V}_k^{\tilde\pi_k}\right)}_{\text{ESTIMATION BIAS TERM}} + \underbrace{\sum_{k=1}^K \left(\widehat{V}_k^{\tilde\pi_k} - \widehat{V}_k^{\pi^\star}\right)}_{\text{OMD REGRET TERM}} + \underbrace{\sum_{k=1}^K \left(\widehat{V}_k^{\pi^\star} - V_k^{\pi^\star}\right)}_{\text{OPTIMISM TERM}}\right].
\tag{5}
$$

Bounding OMD REGRET TERM   The OMD REGRET TERM is contributed by competing against $\pi^\star$ using $\tilde\pi_k$ with $\widehat{Q}_k^{\tilde\pi_k}$ as loss function in the learned model $\widehat{P}_{k_i}$. This term is thus bounded by standard OMD analysis, detailed in the following lemma.

Lemma 4.1. *By setting* $\eta = (1 - \gamma)\sqrt{\ln A/(2L)}$, *the* OMD REGRET TERM *is bounded as* $\mathbb{E}\left[\sum_{k=1}^K \left(\widehat{V}_k^{\tilde\pi_k} - \widehat{V}_k^{\pi^\star}\right)\right] \le \frac{K\sqrt{2\ln A}}{\sqrt{L}(1-\gamma)^2}$.

Bounding OPTIMISM TERM   The OPTIMISM TERM is controlled by choosing appropriate bonus coefficient α_k. Note that different from tabular and linear cases, the bonus functions and coefficients here are not devised to control the optimism for each state-action pair. Instead, they are devised to provide a (near) optimism only at the initial state s_0.

Lemma 4.2. *By setting* $\alpha_k = O(\sqrt{\gamma(A/\xi + d^2)\ln(Mk/\delta)})$ *and* $\lambda_k = O(d\ln(Mk/\delta))$, *with probability* $1 - \delta$, *the* OPTIMISM TERM *is bounded as* $\sum_{k=1}^K \left(\widehat{V}_k^{\pi^\star} - V_k^{\pi^\star}\right) \le (L + \sqrt{K})\sqrt{\frac{A\ln(MN/\delta)}{\xi(1-\gamma)^3}}$.

Bounding ESTIMATION BIAS TERM   It remains to bound the ESTIMATION BIAS TERM, which comes from the difference between the values of running the same policy $\tilde\pi_k$ in the true model (*i.e.*, $P^\star$ and ℓ_k) and the learned empirical model (*i.e.*, $\widehat{P}_k$ and $\ell_k - \widehat{b}_k$), respectively. This term can be translated into the error between the true model and the learned model using the common simulation lemma, which is thus bounded by the summation of bonus functions. Note that since the empirical features used to construct our bonus functions vary in each episode, we first relate the bonus functions with the fixed true feature $\phi^\star$ using the one-step trick [Uehara et al., 2022, Zhang et al., 2022], and finally bound this term with the leverage of the canonical elliptical potential lemma. The result is shown in the following lemma.

Lemma 4.3. *By setting* $\alpha_k = O(\sqrt{\gamma(A/\xi + d^2)\ln(Mk/\delta)})$ *and* $\lambda_k = O(d\ln(Mk/\delta))$, *with probability* $1 - \delta$, *the* ESTIMATION BIAS TERM *is bounded as* $\sum_{k=1}^K \left(V_k^{\tilde\pi_k} - \widehat{V}_k^{\tilde\pi_k}\right) \le O\left(\frac{d^2 A\sqrt{KL}}{\xi(1-\gamma)^3}\sqrt{\ln(1 + K)\ln(MK/\delta)}\right)$.

We refer the readers to Appendix A for the proof of the above lemmas. The proof of Theorem 4.1 is now concluded by first combining Eq. (4), Eq. (5), Lemma 4.1, 4.2, and 4.3 and then choosing $L = K^{1/2}A^{-1/2}d^{-1}\xi(1 - \gamma)$, $\xi = K^{-1/6}A^{1/2}d/(1 - \gamma)$, and $\delta = 1/K$.

Intuitively, the epoch length L illustrates a trade-off between dealing with the adversarial losses and the representation learning over the unknown transitions. When L is large, there will be fewer restarts in the running of OMD and thus the learner will suffer less regret contributed by dealing with the adversarial losses as shown by Lemma 4.1. In contrast, a smaller L enables more frequent model updates, which leads to more accurate model estimation and less regret contributed by the representation learning as shown by Lemma 4.2 and 4.3.

4.3 Regret Lower Bound

This section presents the regret lower bound for learning adversarial low-rank MDPs with fixed loss functions in Theorem 4.2, which thus also serves as a regret lower bound for learning adversarial low-rank MDPs with full-information feedback.

Theorem 4.2. *Suppose $d \geq 8$, $S \geq d + 1$, $A \geq d - 3$, and $K \geq 2(d - 4)A$. Then for any algorithm* Alg, *there exists an episodic infinite-horizon low-rank MDP $\mathcal{M}_{\mathrm{Alg}}$ with fixed loss function such that the regret for this MDP is lower bounded by $\Omega(\frac{\gamma^2}{1-\gamma}\sqrt{dAK})$.*

Proof Sketch. At a high level, we construct dA hard-to-learn low-rank MDP instances, which are difficult to distinguish in KL divergence but have very different optimal policies. In particular, all the constructed low-rank MDP instances have three levels of states, in which the only state in the first level is a fixed initial state and the states in the third level are absorbing states. Moreover, only one unique absorbing state in the third level has the lowest loss, which is termed as the "good state". In the constructed low-rank MDP instance $\mathcal{M}_{(i^\star, a^\star)}$, the learner can only take specific action to transfer to state $s_{2,i^\star}$ in the second level and then take the other specific action to transfer to the unique good state. Due to the unknown representations of state-action pairs, the learner needs to distinguish all these dA low-rank MDP instances, which is essentially equivalent to dealing with a bandit problem with dA "arms". The detailed proof of Theorem 4.2 is postponed to Appendix B. $\square$

Remark 4.2. *Theorem 4.2, to the best of our knowledge, provides the first regret lower bound for learning low-rank MDPs with fixed loss functions. We note that this regret lower bound can hold when $d \ll S$ and $d \ll A$, which thus means that this lower bound is non-trivial. Besides, the regret upper bound in our Theorem 4.1 matches the regret lower bound in A up to a logarithmic factor but looses a factor of $\widetilde{O}(K^{1/3}d^{1/2}/((1-\gamma)\gamma^2))$. Importantly, compared with the regret upper bound $\widetilde{O}(d\sqrt{K/(1-\gamma)^3})$ of linear MDPs [He et al., 2023] (the finite horizon H is substituted by the effective horizon $\Theta(1/(1-\gamma))$ in our infinite-horizon setting for a fair comparison), the dependence on A in the regret lower bound of low-rank MDP shows a clear separation between low-rank MDPs and linear MDPs, which demonstrates that low-rank MDPs are statistically more difficult to learn than linear MDPs in the regret minimization setting.*

5 Conclusions

In this work, we study learning adversarial low-rank MDPs with unknown transition and full-information feedback. We prove that our proposed algorithm POLO achieves the $\widetilde{O}(K^{5/6}A^{1/2}d\ln(1 + M)/(1-\gamma)^2)$ regret, which is the first sublinear regret guarantee for this challenging problem. The design of our proposed algorithm features (a) a doubled exploration and exploitation scheme to simultaneously learn the transitions and adversarial loss functions; and (b) policy optimization in the fixed learned models with epoch-based model update to enable a new analysis scheme that only requires the near optimism at the initial state instead of the point-wise optimism. Also, we prove an $\Omega(\frac{\gamma^2}{1-\gamma}\sqrt{dAK})$ regret lower bound for this problem, serving as the first regret lower bound for learning low-rank MDPs in the regret minimization setting. Besides, there also remain several interesting future directions to be explored. One natural question is whether it is possible to further optimize the dependence of our regret guarantee on the number of episodes K. The other question is how to also learn adversarial low-rank MDPs with only the bandit feedback available. This is also challenging since the current occupancy measure-based methods and policy optimization-based methods tackling adversarial MDPs with bandit feedback both depend on the point-wise optimism provided by the true feature mapping, which seems not feasible in low-rank MDPs. We hope our results may shed light on better understandings of RL with both nonlinear function approximation and adversarial losses and we leave the above extensions as our future works.

Limitations

We note that in general our algorithm is oracle-efficient (given access to the MLE computation oracle in Eq. (1)) but may not be computationally efficient as previous works studying low-rank MDPs [Agarwal et al., 2020, Uehara et al., 2022, Zhang et al., 2022, Ni et al., 2022]. However, we also remark that in practice, these algorithms including ours are computationally feasible since the computation of MLE is only a standard supervised learning problem and can be implemented using gradient descent methods. The other limitation is that throughout this paper, we assume model class $\mathcal{M}$ with bounded cardinality M and the regret upper bound of our algorithm has a logarithmic dependence on M. We remark that similar assumptions and dependence have also appeared in previous theoretical works studying RL with general function approximation [Jiang et al., 2017, Sun et al., 2019]. Also, extending the analyses to an infinite hypothesis class is possible if the hypothesis class has bounded statistical complexity [Agarwal et al., 2020].

Acknowledgement

References

Pieter Abbeel and Andrew Y. Ng. Exploration and apprenticeship learning in reinforcement learning. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 1–8. ACM, 2005.

Alekh Agarwal, Sham M. Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: structural complexity and representation learning of low rank mdps. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22: 98:1–98:76, 2021.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Deterministic mdps with adversarial rewards and bandit feedback. In Nando de Freitas and Kevin P. Murphy, editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*, pages 93–101. AUAI Press, 2012.

Alex Ayoub, Zeyu Jia, Csaba Szepesvári, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 463–474. PMLR, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 2017.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1283–1294. PMLR, 2020.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. In *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 1180–1215. PMLR, 2021.

Liyu Chen, Haipeng Luo, and Aviv Rosenberg. Policy optimization for stochastic shortest path. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 982–1046. PMLR, 2022.

Yuan Cheng, Ruiquan Huang, Yingbin Liang, and Jing Yang. Improved sample complexity for reward-free reinforcement learning under low-rank mdps. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Yan Dai, Haipeng Luo, and Liyu Chen. Follow-the-perturbed-leader for adversarial markov decision processes with bandit feedback. *CoRR*, abs/2205.13451, 2022. doi: 10.48550/arXiv.2205.13451.

Yan Dai, Haipeng Luo, Chen-Yu Wei, and Julian Zimmert. Refined regret for adversarial mdps with linear function approximation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 6726–6759. PMLR, 2023.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. On oracle-efficient PAC RL with rich observations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1429–1439, 2018.

Ofer Dekel and Elad Hazan. Better rates for any adversarial deterministic MDP. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 675–683. JMLR.org, 2013.

Travis Dick, András György, and Csaba Szepesvári. Online learning in markov decision processes with changing cost sequences. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 512–520. JMLR.org, 2014.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, pages 578–598. PMLR, 2021.

Simon S. Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1665–1674. PMLR, 2019.

Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Simon S. Du, Sham M. Kakade, Jason D. Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in RL. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2826–2836. PMLR, 2021.

Eyal Even-Dar, Sham M. Kakade, and Yishay Mansour. Online markov decision processes. *Math. Oper. Res.*, 34(3):726–736, 2009.

A. Feinberg. Markov decision processes: Discrete stochastic dynamic programming (martin l. puterman). *SIAM Rev.*, 38(4):689, 1996.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Math. Oper. Res.*, 44(2):377–399, 2019.

Sara A Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Mahsa Ghasemi, Abolfazl Hashemi, Haris Vikalo, and Ufuk Topcu. No-regret learning with high-probability in adversarial markov decision processes. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, volume 161 of *Proceedings of Machine Learning Research*, pages 992–1001. AUAI Press, 2021.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4171–4180. PMLR, 2021a.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted mdps. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 22288–22300, 2021b.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 4259–4280. PMLR, 2022.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 12790–12822. PMLR, 2023.

Pihe Hu, Yu Chen, and Longbo Huang. Nearly minimax optimal reinforcement learning with linear function approximation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 8971–9019. PMLR, 2022.

Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, and Lin Yang. Randomized exploration in reinforcement learning with general value function approximation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4607–4616. PMLR, 2021.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 2017.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4860–4869. PMLR, 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 2020b.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 13406–13418, 2021a.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic mdps with known transition. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 20491–20502, 2021b.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In Claude Sammut and Achim G. Hoffmann, editors, *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8-12, 2002*, pages 267–274. Morgan Kaufmann, 2002.

Fang Kong, Xiangcheng Zhang, Baoxiang Wang, and Shuai Li. Improved regret bounds for linear adversarial mdps via linear optimization. *CoRR*, abs/2302.06834, 2023.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: contrastive unsupervised representations for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5639–5650. PMLR, 2020.

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

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *CoRR*, abs/2107.08346, 2021a.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 22931–22942, 2021b.

Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Learning stochastic shortest path with linear function approximation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 15584–15629. PMLR, 2022.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *CoRR*, abs/2102.07035, 2021.

Gergely Neu and Julia Olkhovskaya. Online learning in mdps with linear function approximation and bandit feedback. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 10407–10417, 2021.

Gergely Neu, András György, and Csaba Szepesvári. The online loop-free stochastic shortest-path problem. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 231–243. Omnipress, 2010a.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 1804–1812. Curran Associates, Inc., 2010b.

Gergely Neu, András György, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, Spain, April 21-23, 2012*, volume 22 of *JMLR Proceedings*, pages 805–813. JMLR.org, 2012.

Chengzhuo Ni, Yuda Song, Xuezhou Zhang, Chi Jin, and Mengdi Wang. Representation learning for general-sum low-rank markov games. *CoRR*, abs/2210.16976, 2022.

Tongzheng Ren, Tianjun Zhang, Csaba Szepesvári, and Bo Dai. A free lunch from the noise: Provable and practical exploration for representation learning. In *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*, volume 180 of *Proceedings of Machine Learning Research*, pages 1686–1696. PMLR, 2022.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2209–2218, 2019a.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5478–5486. PMLR, 2019b.

Aviv Rosenberg and Yishay Mansour. Stochastic shortest path with adversarially changing costs. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 2936–2942. ijcai.org, 2021.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2256–2264, 2013.

Max Schwarzer, Nitarshan Rajkumar, Michael Noukhovitch, Ankesh Anand, Laurent Charlin, R. Devon Hjelm, Philip Bachman, and Aaron C. Courville. Pretraining representations for data-efficient reinforcement learning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12686–12699, 2021.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8604–8613. PMLR, 2020a.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8604–8613. PMLR, 2020b.

Uri Sherman, Tomer Koren, and Yishay Mansour. Improved regret for efficient online reinforcement learning with linear function approximation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31117–31150. PMLR, 2023.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419): 1140–1144, 2018.

Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9870–9879. PMLR, 2021.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 2898–2933. PMLR, 2019.

Aristide C. Y. Tossou, Debabrota Basu, and Christos Dimitrakakis. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. *CoRR*, abs/1905.12425, 2019.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank mdps. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F. Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Ruosong Wang, Dean P. Foster, and Sham M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a.

Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, pages 1237–1264. PMLR, 2021.

Yue Wu, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 3883–3913. PMLR, 2022.

Zhihui Xie, Zichuan Lin, Junyou Li, Shuai Li, and Deheng Ye. Pretraining in deep reinforcement learning: A survey. *CoRR*, abs/2211.03959, 2022.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6995–7004. PMLR, 2019.

Mengjiao Yang and Ofir Nachum. Representation matters: Offline pretraining for sequential decision making. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11784–11794. PMLR, 2021.

Jia Yuan Yu, Shie Mannor, and Nahum Shimkin. Markov decision processes with arbitrary reward processes. *Math. Oper. Res.*, 34(3):737–757, 2009.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 1954–1964. PMLR, 2020.

Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 4473–4525. PMLR, 2021.

Tong Zhang. From ε-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient reinforcement learning in block mdps: A model-free representation learning approach. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 26517–26547. PMLR, 2022.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S. Du. Improved variance-aware confidence sets for linear bandits and linear mixture MDP. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 4342–4355, 2021.

Canzhe Zhao, Ruofeng Yang, Baoxiang Wang, and Shuai Li. Learning adversarial linear mixture markov decision processes with bandit feedback and unknown transition. In *The Eleventh International Conference on Learning Representations*, 2023.

Han Zhong and Tong Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *arXiv preprint arXiv:2305.08841*, 2023.

Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *CoRR*, abs/2205.11507, 2022.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvári. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 4532–4576. PMLR, 2021.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 1583–1591, 2013.

Appendix

A Omitted Analysis of The Regret Upper Bound

In this section, we first introduce some further notations and then present the detailed analysis of Theorem 4.1. For the proofs in this section, we assume that K is divisible by the epoch length L considered for simplicity.

To begin with, denote by $\rho_k(s) = 1/k \sum_{i=1}^k d_{P^\star}^{\tilde{\pi}_i}(s)$ the averaged state occupancy distribution under $\{\tilde{\pi}_i\}_{i=1}^k$ and $P^\star$. Analogously, let $\rho_k'(s') = \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho_k(s) \bar{\pi}_k(a \mid s) P^\star(s' \mid s, a)$ be the average next-state occupancy distribution after s is sampled from $d_{P^\star}^{\tilde{\pi}_i}$, where $\bar{\pi}_k(\cdot \mid s) = \xi \cdot U(\mathcal{A}) + (1 - \xi) \cdot 1/k \sum_{i=1}^k \tilde{\pi}_i(\cdot \mid s)$ is the averaged mixed roll-out policy. Define $f_k(s, a) = \left\| \widehat{P}_k(\cdot \mid s, a) - P^\star(\cdot \mid s, a) \right\|_1$ as the ℓ_1-error between the estimate transition kernel $\widehat{P}_k$ and $P^\star$. Further, we define the following feature covariance matrices:

- $\widehat{\Sigma}_{k,\phi} = k \mathbb{E}_{(s,a) \sim \mathcal{D}_k} \left[\phi(s,a) \phi(s,a)^\top \right] + \lambda_k I$,
- $\Sigma_{\rho_k, \phi} = k \mathbb{E}_{(s,a) \sim \rho_k} \left[\phi(s,a) \phi(s,a)^\top \right] + \lambda_k I$,
- $\Sigma_{\rho_k \times \bar{\pi}_k, \phi} = k \mathbb{E}_{s \sim \rho_k, a \sim \bar{\pi}_k} \left[\phi(s,a) \phi(s,a)^\top \right] + \lambda_k I$.

For notational convenience, we abbreviate $\widehat{\Sigma}_{k, \phi_k}$ as $\widehat{\Sigma}_k^{-1}$. Also, note that $\widehat{\Sigma}_{k, \phi}$ is an unbiased estimate of $\Sigma_{\rho_k \times \bar{\pi}_k, \phi}$.

A.1 Bounding OMD REGRET TERM

We first present the proof of Lemma 4.1, which follows from the standard OMD analysis.

Proof of Lemma 4.1. We first consider fixed initial state s_0. For some fixed $i \in [N]$, one can see that the OMD regret in episode $k \in \{k_i, k_i + 1, \ldots, k_i + L - 1\}$ can be written recursively as

$$
\begin{aligned}
& \widehat{V}_k^{\tilde{\pi}_k}(s_0) - \widehat{V}_k^{\pi^\star}(s_0) \\
&= \left\langle \tilde{\pi}_k(\cdot \mid s_0), \widehat{Q}_k^{\tilde{\pi}_k}(s_0, \cdot) \right\rangle - \left\langle \pi^\star(\cdot \mid s_0), \widehat{Q}_k^{\pi^\star}(s_0, \cdot) \right\rangle \\
&= \left\langle \tilde{\pi}_k(\cdot \mid s_0) - \pi^\star(\cdot \mid s_0), \widehat{Q}_k^{\tilde{\pi}_k}(s_0, \cdot) \right\rangle + \left\langle \pi^\star(\cdot \mid s_0), \widehat{Q}_k^{\tilde{\pi}_k}(s_0, \cdot) - \widehat{Q}_k^{\pi^\star}(s_0, \cdot) \right\rangle \\
&= \left\langle \tilde{\pi}_k(\cdot \mid s_0) - \pi^\star(\cdot \mid s_0), \widehat{Q}_k^{\tilde{\pi}_k}(s_0, \cdot) \right\rangle + \left\langle \pi^\star(\cdot \mid s_0), \gamma \left[\widehat{P}_{k_i} \left(\widehat{V}_k^{\tilde{\pi}_k} - \widehat{V}_k^{\pi^\star} \right) \right](s_0, \cdot) \right\rangle \\
&= \mathbb{E} \left[\sum_{\tau=0}^\infty \gamma^\tau \left\langle \tilde{\pi}_k(\cdot \mid s_{k,\tau}) - \pi^\star(\cdot \mid s_{k,\tau}), \widehat{Q}_k^{\tilde{\pi}_k}(s_{k,\tau}, \cdot) \right\rangle \,\middle|\, \pi^\star, \widehat{P}_{k_i}, s_0 \right],
\end{aligned}
$$

where recall $\widehat{Q}_k^\pi(s, a) = \mathbb{E} \left[\sum_{\tau=0}^\infty \gamma^t [\ell_k - \widehat{b}_k](s_{k,\tau}, a_{k,\tau}) \,\middle|\, \pi, \widehat{P}_{k_i}, (s_{k,0}, a_{k,0}) = (s, a) \right]$ is the state-action value of policy π under the empirical model $(\widehat{P}_{k_i}, \ell_k - \widehat{b}_{k_i})$ and the expectation $\mathbb{E}[\cdot \mid \pi^\star, \widehat{P}_{k_i}, s_0]$ is taken over the randomness of the state-action sequence $\{(s_{k,\tau}, a_{k,\tau})\}_{\tau=0}^\infty$ with $a_{k,\tau} \sim \pi^\star(\cdot \mid s_{k,\tau})$, $s_{k,\tau+1} \sim \widehat{P}_{k_i}(\cdot \mid s_{k,\tau}, a_{k,\tau})$, and $s_{k,0} = s_0$.

Taking summation of the above display over $k \in \{k_i, k_i + 1, \ldots, k_i + L - 1\}$ and re-arranging show that

$$
\begin{aligned}
& \mathbb{E} \left[\sum_{k=k_i}^{k_i+L-1} \left(\widehat{V}_k^{\tilde{\pi}_k}(s_0) - \widehat{V}_k^{\pi^\star}(s_0) \right) \right] \\
&= \mathbb{E} \left[\mathbb{E} \left[\sum_{k=k_i}^{k_i+L-1} \sum_{\tau=0}^\infty \gamma^\tau \left\langle \tilde{\pi}_k(\cdot \mid s_{k,\tau}) - \pi^\star(\cdot \mid s_{k,\tau}), \widehat{Q}_k^{\tilde{\pi}_k}(s_{k,\tau}, \cdot) \right\rangle \,\middle|\, \pi^\star, \widehat{P}_{k_i}, s_0 \right] \right] \\
&= \sum_{\tau=0}^\infty \gamma^\tau \mathbb{E} \left[\mathbb{E} \left[\sum_{k=k_i}^{k_i+L-1} \left\langle \tilde{\pi}_k(\cdot \mid s_{k,\tau}) - \pi^\star(\cdot \mid s_{k,\tau}), \widehat{Q}_k^{\tilde{\pi}_k}(s_{k,\tau}, \cdot) \right\rangle \,\middle|\, \pi^\star, \widehat{P}_{k_i}, s_0 \right] \right]. \quad (6)
\end{aligned}
$$

Further, note that the update process of policy in Eq. (3) can be solved by the following two-step procedure [Lattimore and Szepesvári, 2020]:

$$\widehat{\pi}_{k+1}(\cdot \mid s) \in \underset{\pi(\cdot \mid s) \in \mathbb{R}_+^A}{\arg\min} \, \eta \left\langle \pi(\cdot \mid s), \widehat{Q}_k^{\tilde{\pi}_k}(s, \cdot) \right\rangle + D_F(\pi(\cdot \mid s), \tilde{\pi}_k(\cdot \mid s)) \quad \text{and} \qquad (7)$$

$$\tilde{\pi}_{k+1}(\cdot \mid s) \in \underset{\pi(\cdot \mid s) \in \Delta(\mathcal{A})}{\arg\min} \, D_F(\pi(\cdot \mid s), \widehat{\pi}_{k+1}(\cdot \mid s)), \qquad (8)$$

for any $s \in \mathcal{S}$. Eq. (7) combined with the first-order optimality condition implies that

$$\widehat{Q}_k^{\tilde{\pi}_k}(s, \cdot) = -\frac{1}{\eta}(\nabla F(\widehat{\pi}_{k+1}(\cdot \mid s)) - \nabla F(\tilde{\pi}_k(\cdot \mid s))), \qquad (9)$$

for any $s \in \mathcal{S}$. This display shows that $\widehat{\pi}_{k+1}(a \mid s) = \tilde{\pi}_k(a \mid s) \exp(-\eta \widehat{Q}_k^{\tilde{\pi}_k}(s, \cdot))$, and $\widehat{\pi}_{k+1}(a \mid s) \le \tilde{\pi}_k(a \mid s)$ since $\widehat{Q}_k^{\tilde{\pi}_k}(s, a) \ge 0$, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Therefore, one can see that

$$\mathbb{E}\left[\left\langle \tilde{\pi}_k(\cdot \mid s_{k,\tau}) - \pi^\star(\cdot \mid s_{k,\tau}), \widehat{Q}_k^{\tilde{\pi}_k}(s_{k,\tau}, \cdot) \right\rangle \middle| \pi^\star, \widehat{P}_{k_i}, s_0\right]$$

$$=\mathbb{E}\left[\frac{1}{\eta}\langle \pi^\star(\cdot \mid s_{k,\tau}) - \tilde{\pi}_k(\cdot \mid s_{k,\tau}), \nabla F(\widehat{\pi}_{k+1}(\cdot \mid s_{k,\tau})) - \nabla F(\tilde{\pi}_k(\cdot \mid s_{k,\tau}))\rangle \middle| \pi^\star, \widehat{P}_{k_i}, s_0\right]$$

$$=\mathbb{E}\left[\frac{1}{\eta}(D_F(\pi^\star(\cdot \mid s_{k,\tau}), \tilde{\pi}_k(\cdot \mid s_{k,\tau})) + D_F(\tilde{\pi}_k(\cdot \mid s_{k,\tau}), \widehat{\pi}_{k+1}(\cdot \mid s_{k,\tau}))\right.$$

$$\left. -D_F(\pi^\star(\cdot \mid s_{k,\tau}), \widehat{\pi}_{k+1}(\cdot \mid s_{k,\tau}))) \middle| \pi^\star, \widehat{P}_{k_i}, s_0\right]$$

$$\le\mathbb{E}\left[\frac{1}{\eta}(D_F(\pi^\star(\cdot \mid s_{k,\tau}), \tilde{\pi}_k(\cdot \mid s_{k,\tau})) + D_F(\tilde{\pi}_k(\cdot \mid s_{k,\tau}), \widehat{\pi}_{k+1}(\cdot \mid s_{k,\tau}))\right.$$

$$\left. -D_F(\pi^\star(\cdot \mid s_{k,\tau}), \tilde{\pi}_{k+1}(\cdot \mid s_{k,\tau}))) \middle| \pi^\star, \widehat{P}_{k_i}, s_0\right], \qquad (10)$$

where the first equality comes from Eq. (9), the second equality is due to the three-point lemma, and the last inequality follows the generalized Pythagorean theorem.

Taking summation of Eq. (10) over $k \in \{k_i, k_i + 1, \ldots, k_i + L - 1\}$ leads to

$$\mathbb{E}\left[\sum_{k=k_i}^{k_i+L-1} \left\langle \tilde{\pi}_k(\cdot \mid s_{k,\tau}) - \pi^\star(\cdot \mid s_{k,\tau}), \widehat{Q}_k^{\tilde{\pi}_k}(s_{k,\tau}, \cdot) \right\rangle \middle| \pi^\star, \widehat{P}_{k_i}, s_0\right]$$

$$\le\mathbb{E}\left[\frac{1}{\eta}\left(D_F(\pi^\star(\cdot \mid s_{k_i,\tau}), \tilde{\pi}_{k_i}(\cdot \mid s_{k_i,\tau})) + \sum_{k=k_i}^{k_i+L-1} D_F(\tilde{\pi}_k(\cdot \mid s_{k,\tau}), \widehat{\pi}_{k+1}(\cdot \mid s_{k,\tau}))\right) \middle| \pi^\star, \widehat{P}_{k_i}, s_0\right].$$

$$(11)$$

The first term in Eq. (11) can be bounded as follows:

$$\mathbb{E}\left[D_F(\pi^\star(\cdot \mid s_{k_i,\tau}), \tilde{\pi}_{k_i}(\cdot \mid s_{k_i,\tau})) \middle| \pi^\star, \widehat{P}_{k_i}, s_0\right]$$

$$=\mathbb{E}\left[\sum_{a \in \mathcal{A}} \pi^\star(a \mid s_{k_i,\tau}) \ln \frac{\pi^\star(a \mid s_{k_i,\tau})}{\tilde{\pi}_{k_i}(a \mid s_{k_i,\tau})} \middle| \pi^\star, \widehat{P}_{k_i}, s_0\right]$$

$$\le\mathbb{E}\left[\sum_{a \in \mathcal{A}} \pi^\star(a \mid s_{k_i,\tau}) \ln A \middle| \pi^\star, \widehat{P}_{k_i}, s_0\right]$$

$$= \ln A, \qquad (12)$$

where the inequality is because we choose $\tilde{\pi}_{k_i}(\cdot \mid s) = U(\mathcal{A})$ for any $s \in \mathcal{S}$ in Algorithm 1.

It remains to bound $\mathbb{E}\left[\sum_{k=k_i}^{k_i+L-1} D_F(\tilde{\pi}_k(\cdot \mid s_{k,\tau}), \widehat{\pi}_{k+1}(\cdot \mid s_{k,\tau})) \middle| \pi^\star, \widehat{P}_{k_i}, s_0\right]$ in Eq. (11):

$$\mathbb{E}\left[\sum_{k=k_i}^{k_i+L-1} D_F(\tilde{\pi}_k(\cdot \mid s_{k,\tau}), \widehat{\pi}_{k+1}(\cdot \mid s_{k,\tau})) \middle| \pi^\star, \widehat{P}_{k_i}, s_0\right]$$

$$
\begin{aligned}
=&\mathbb{E}\left[\sum_{k=k_i}^{k_i+L-1}\left(-D_F(\widehat{\pi}_{k+1}(\cdot\mid s_{k,\tau}),\tilde{\pi}_k(\cdot\mid s_{k,\tau}))\right.\right.\\
&\left.\left.+\langle\nabla F(\tilde{\pi}_k(\cdot\mid s_{k,\tau}))-\nabla F(\widehat{\pi}_{k+1}(\cdot\mid s_{k,\tau})),\tilde{\pi}_k(\cdot\mid s_{k,\tau})-\widehat{\pi}_{k+1}(\cdot\mid s_{k,\tau})\rangle\right)\Big|\pi^\star,\widehat{P}_{k_i},s_0\right]\\
\leq&\mathbb{E}\left[\sum_{k=k_i}^{k_i+L-1}\left(-D_F(\widehat{\pi}_{k+1}(\cdot\mid s_{k,\tau}),\tilde{\pi}_k(\cdot\mid s_{k,\tau}))+\frac{1}{2}\|\nabla F(\tilde{\pi}_k(\cdot\mid s_{k,\tau}))-\nabla F(\widehat{\pi}_{k+1}(\cdot\mid s_{k,\tau}))\|^2_{\nabla^{-2}F(z_k(\cdot\mid s_{k,\tau}))}\right.\right.\\
&\left.\left.+\frac{1}{2}\|\tilde{\pi}_k(\cdot\mid s_{k,\tau})-\widehat{\pi}_{k+1}(\cdot\mid s_{k,\tau})\|^2_{\nabla^2 F(z_k(\cdot\mid s_{k,\tau}))}\right)\Big|\pi^\star,\widehat{P}_{k_i},s_0\right]\\
=&\mathbb{E}\left[\sum_{k=k_i}^{k_i+L-1}\left(-D_F(\widehat{\pi}_{k+1}(\cdot\mid s_{k,\tau}),\tilde{\pi}_k(\cdot\mid s_{k,\tau}))+\frac{1}{2}\|\eta\widehat{Q}_k^{\tilde{\pi}_k}(s_{k,\tau},\cdot)\|^2_{\nabla^{-2}F(z_k(\cdot\mid s_{k,\tau}))}\right.\right.\\
&\left.\left.+\frac{1}{2}\|\tilde{\pi}_k(\cdot\mid s_{k,\tau})-\widehat{\pi}_{k+1}(\cdot\mid s_{k,\tau})\|^2_{\nabla^2 F(z_k(\cdot\mid s_{k,\tau}))}\right)\Big|\pi^\star,\widehat{P}_{k_i},s_0\right]\\
=&\mathbb{E}\left[\sum_{k=k_i}^{k_i+L-1}\left(-\frac{1}{2}\|\tilde{\pi}_k(\cdot\mid s_{k,\tau})-\widehat{\pi}_{k+1}(\cdot\mid s_{k,\tau})\|^2_{\nabla^2 F(\omega_k(\cdot\mid s_{k,\tau}))}+\frac{1}{2}\|\eta\widehat{Q}_k^{\tilde{\pi}_k}(s_{k,\tau},\cdot)\|^2_{\nabla^{-2}F(z_k(\cdot\mid s_{k,\tau}))}\right.\right.\\
&\left.\left.+\frac{1}{2}\|\tilde{\pi}_k(\cdot\mid s_{k,\tau})-\widehat{\pi}_{k+1}(\cdot\mid s_{k,\tau})\|^2_{\nabla^2 F(z_k(\cdot\mid s_{k,\tau}))}\right)\Big|\pi^\star,\widehat{P}_{k_i},s_0\right]\\
=&\mathbb{E}\left[\frac{\eta^2}{2}\sum_{k=k_i}^{k_i+L-1}\sum_{a\in\mathcal{A}}z_k(a\mid s_{k,\tau})\widehat{Q}_k^{\tilde{\pi}_k}(s_{k,\tau},a)^2\Big|\pi^\star,\widehat{P}_{k_i},s_0\right]\\
\leq&\mathbb{E}\left[\frac{\eta^2}{2}\sum_{k=k_i}^{k_i+L-1}\sum_{a\in\mathcal{A}}\tilde{\pi}_k(a\mid s_{k,\tau})\widehat{Q}_k^{\tilde{\pi}_k}(s_{k,\tau},a)^2\Big|\pi^\star,\widehat{P}_{k_i},s_0\right]\\
\leq&2\eta^2 L/(1-\gamma)^2\,,
\end{aligned}
\tag{13}
$$

where the second line comes from the Young-Fenchel inequality for all $z_k(\cdot\mid s_{k,\tau})\in[\widehat{\pi}_{k+1}(\cdot\mid s_{k,\tau}),\tilde{\pi}_k(\cdot\mid s_{k,\tau})]$ arbitrarily, the third line follows by the first-order optimality condition in Eq. (9), the fourth line is by the mean value theorem of the second derivative for some fixed $\omega_k(\cdot\mid s_{k,\tau})\in[\widehat{\pi}_{k+1}(\cdot\mid s_{k,\tau}),\tilde{\pi}_k(\cdot\mid s_{k,\tau})]$, the fifth line comes from fixing $z_k(\cdot\mid s_{k,\tau})=\omega_k(\cdot\mid s_{k,\tau})$, the sixth line comes from the fact that $z_k(\cdot\mid s_{k,\tau})\leq\tilde{\pi}_k(\cdot\mid s_{k,\tau})$ and the last line is due to that $|\widehat{Q}_k^{\tilde{\pi}_k}(s_{k,\tau},a)|\leq 2/(1-\gamma)$.

Substituting Eq. (12) and Eq. (13) into Eq. (11), along with Eq. (6), shows that

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K}\left(\widehat{V}_k^{\tilde{\pi}_k}-\widehat{V}_k^{\pi^\star}\right)\right]\\
=&\mathbb{E}\left[\mathbb{E}_{s_0\sim d_0}\left[\sum_{k=1}^{K}\left(\widehat{V}_k^{\tilde{\pi}_k}(s_0)-\widehat{V}_k^{\pi^\star}(s_0)\right)\right]\right]\\
=&\mathbb{E}\left[\mathbb{E}_{s_0\sim d_0}\left[\sum_{i=1}^{N}\sum_{k=k_i}^{k_i+L-1}\left(\widehat{V}_k^{\tilde{\pi}_k}(s_0)-\widehat{V}_k^{\pi^\star}(s_0)\right)\right]\right]\\
=&\sum_{\tau=0}^{\infty}\gamma^\tau\mathbb{E}\left[\mathbb{E}_{s_0\sim d_0}\left[\sum_{i=1}^{N}\mathbb{E}\left[\sum_{k=k_i}^{k_i+L-1}\left\langle\tilde{\pi}_k(\cdot\mid s_{k,\tau})-\pi^\star(\cdot\mid s_{k,\tau}),\widehat{Q}_k^{\tilde{\pi}_k}(s_{k,\tau},\cdot)\right\rangle\Big|\pi^\star,\widehat{P}_{k_i},s_0\right]\right]\right]\\
\leq&\sum_{\tau=0}^{\infty}\gamma^\tau N\cdot\frac{1}{\eta}\left(\ln A+2\eta^2 L/(1-\gamma)^2\right)\leq\frac{K\sqrt{2\ln A}}{\sqrt{L}(1-\gamma)^2}\,,
\end{aligned}
$$

which completes the proof. $\qquad\square$

20

A.2 Bounding OPTIMISM TERM

We now turn to prove Lemma 4.2, which provides a (near) optimism at the initial state distribution.

Proof of Lemma 4.2. Recall $f_{k_i}(s,a) = \left\| \widehat{P}_{k_i}(\cdot \mid s,a) - P^\star(\cdot \mid s,a) \right\|_1$. In the following, we condition on the good event

$$\mathcal{E} = \left\{ \forall i \in [N], \mathbb{E}_{s \sim \rho_{k_i}, a \sim \bar{\pi}_{k_i}} \left[f_{k_i}^2(s,a) \right] \leq \zeta_{k_i}, \mathbb{E}_{s \sim \rho'_{k_i}, a \sim \bar{\pi}_{k_i}} \left[f_{k_i}^2(s,a) \right] \leq \zeta_{k_i}; \right.$$

$$\left. \forall i \in [N], \forall \phi, \|\phi(s,a)\|_{\widehat{\Sigma}_{k_i,\phi}^{-1}} = \Theta \left(\|\phi(s,a)\|_{\Sigma_{\rho_{k_i} \times \bar{\pi}_{k_i},\phi}^{-1}} \right) \right\},$$

which is guaranteed to hold with probability $1 - \delta$ by using union bound with Lemma C.1 and Lemma C.3.

We first consider fixed initial state s_0. For some epoch $i \in [N]$ and episode $k \in \{k_i, k_i + 1, \ldots, k_i + L - 1\}$, applying Eq. (32) in Lemma C.4 implies that

$$\widehat{V}_k^{\pi^\star}(s_0) - V_k^{\pi^\star}(s_0)$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\widehat{P}_{k_i}}^{\pi^\star}} \left[-\widehat{b}_{k_i}(s,a) + \gamma \left(\widehat{P}_{k_i}(\cdot \mid s,a) - P^\star(\cdot \mid s,a) \right)^\top V_k^{\pi^\star} \right]$$

$$\leq \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\widehat{P}_{k_i}}^{\pi^\star}} \left[-\widehat{b}_{k_i}(s,a) + \frac{\gamma}{1-\gamma} f_{k_i}(s,a) \right]$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{\widehat{P}_{k_i}}^{\pi^\star}} \left[- \min \left(\alpha_{k_i} \left\| \widehat{\phi}_{k_i}(s,a) \right\|_{\Sigma_{\rho_{k_i} \times \bar{\pi}_{k_i}, \widehat{\phi}_{k_i}}^{-1}}, 2 \right) \frac{1}{(1-\gamma)} \right]$$

$$+ \frac{\gamma}{(1-\gamma)^2} \left(\underbrace{\gamma \mathbb{E}_{(\tilde{s},\tilde{a}) \sim d_{\widehat{P}_{k_i}}^{\pi^\star}, s \sim \widehat{P}_{k_i}(\cdot|\tilde{s},\tilde{a}), a \sim \pi^\star(s)} \left[f_{k_i}(s,a) \right]}_{\text{TERM}_1} + \underbrace{(1-\gamma) \mathbb{E}_{s \sim d_0, a \sim \pi^\star(s)} \left[f_{k_i}(s,a) \right]}_{\text{TERM}_2} \right), \tag{14}$$

where the inequality comes from the Cauchy–Schwarz inequality together with $\|V_k^{\pi^\star}\|_\infty = 1/(1-\gamma)$, and the second equality is due to the fact that $\widehat{\Sigma}_{k_i, \widehat{\phi}_{k_i}}$ is an unbiased estimate of $\Sigma_{\rho_{k_i} \times \bar{\pi}_{k_i}, \widehat{\phi}_{k_i}}$ and Lemma A.4.

We first bound TERM$_2$ as follows:

$$(1-\gamma) \mathbb{E}_{s \sim d_0, a \sim \pi^\star(s)} \left[f_{k_i}(s,a) \right] \leq \sqrt{\frac{(1-\gamma)A}{\xi} \mathbb{E}_{s \sim \rho_{k_i}, a \sim \bar{\pi}_{k_i}} \left[f_{k_i}^2(s,a) \right]} \leq \sqrt{\frac{(1-\gamma)A\zeta_{k_i}}{\xi}}, \tag{15}$$

where the first inequality is due to Lemma A.1 and the second inequality is by the definition of event $\mathcal{E}$.

It remains to bound TERM$_1$:

$$\gamma \mathbb{E}_{(\tilde{s},\tilde{a}) \sim d_{\widehat{P}_{k_i}}^{\pi^\star}, s \sim \widehat{P}_{k_i}(\cdot|\tilde{s},\tilde{a}), a \sim \pi^\star(s)} \left[f_{k_i}(s,a) \right]$$

$$\leq \gamma \mathbb{E}_{(\tilde{s},\tilde{a}) \sim d_{\widehat{P}_{k_i}}^{\pi^\star}} \left[\left\| \widehat{\phi}_{k_i}(\tilde{s},\tilde{a}) \right\|_{\Sigma_{\rho_{k_i} \times \bar{\pi}_{k_i}, \widehat{\phi}_{k_i}}^{-1}} \right] \sqrt{\frac{k_i A}{\xi} \mathbb{E}_{s \sim \rho'_{k_i}, a \sim \bar{\pi}_{k_i}} \left[f_{k_i}^2(s,a) \right] + 4\lambda_{k_i} d + 4k_i \zeta_{k_i}}$$

$$\leq \gamma \mathbb{E}_{(\tilde{s},\tilde{a}) \sim d_{\widehat{P}_{k_i}}^{\pi^\star}} \left[\left\| \widehat{\phi}_{k_i}(\tilde{s},\tilde{a}) \right\|_{\Sigma_{\rho_{k_i} \times \bar{\pi}_{k_i}, \widehat{\phi}_{k_i}}^{-1}} \right] \sqrt{\frac{k_i A}{\xi} \zeta_{k_i} + 4\lambda_{k_i} d + 4k_i \zeta_{k_i}}$$

$$\lesssim \alpha_{k_i} \mathbb{E}_{(\tilde{s},\tilde{a}) \sim d_{\widehat{P}_{k_i}}^{\pi^\star}} \left[\left\| \widehat{\phi}_{k_i}(\tilde{s},\tilde{a}) \right\|_{\Sigma_{\rho_{k_i} \times \bar{\pi}_{k_i}, \widehat{\phi}_{k_i}}^{-1}} \right], \tag{16}$$

21

where the first inequality follows by Lemma A.3 as well as $\|f_{k_i}\|_\infty \le 2$, the second inequality is again due to the definition of the good event $\mathcal{E}$, and the last inequality comes from the definition of α_{k_i}.

Now substituting Eq. (15) and Eq. (16) into Eq. (14) shows that

$$\sum_{k=1}^{K}\left(\widehat{V}_k^{\pi^\star} - V_k^{\pi^\star}\right)$$

$$=\mathbb{E}_{s_0 \sim d_0}\left[\sum_{k=1}^{K}\left(\widehat{V}_k^{\pi^\star}(s_0) - V_k^{\pi^\star}(s_0)\right)\right]$$

$$=\mathbb{E}_{s_0 \sim d_0}\left[\sum_{i=1}^{N}\sum_{k=k_i}^{k_i+L-1}\left(\widehat{V}_k^{\pi^\star}(s_0) - V_k^{\pi^\star}(s_0)\right)\right]$$

$$\le \sum_{i=1}^{N}\sum_{k=k_i}^{k_i+L-1}\frac{1}{(1-\gamma)^2}\mathbb{E}_{(s,a)\sim d_{\widehat{P}_{k_i}}^{\pi^\star}}\left[-\min\left(\alpha_{k_i}\left\|\widehat{\phi}_{k_i}(s,a)\right\|_{\Sigma^{-1}_{\rho_{k_i}\times\bar{\pi}_{k_i},\widehat{\phi}_{k_i}}}, 2\right)\right.$$

$$\left.+\min\left(\alpha_{k_i}\left\|\widehat{\phi}_{k_i}(s,a)\right\|_{\Sigma^{-1}_{\rho_{k_i}\times\bar{\pi}_{k_i},\widehat{\phi}_{k_i}}} + \sqrt{\frac{(1-\gamma)A\zeta_{k_i}}{\xi}}, 2\right)\right]$$

$$\le \sum_{i=1}^{N}L\sqrt{\frac{A\ln(MN/\delta)}{\xi(1-\gamma)^3 k_i}}$$

$$=\sum_{i=1}^{N}L\sqrt{\frac{A\ln(MN/\delta)}{\xi(1-\gamma)^3((i-1)L+1)}}$$

$$\le (L+\sqrt{K})\sqrt{\frac{A\ln(MN/\delta)}{\xi(1-\gamma)^3}},$$

where the first inequality follows by $\|f_{k_i}\|_\infty \le 2$ for any $i \in [N]$ and the second inequality is due to the definition of ζ_{k_i} in Lemma C.1. $\qquad\square$

A.3 Bounding ESTIMATION BIAS TERM

We now give the proof of Lemma 4.3, which controls the estimation bias term.

Proof of Lemma 4.3. Similar to the proof of Lemma 4.2, in what follows, we condition on the good event

$$\mathcal{E} = \left\{\forall i \in [N], \mathbb{E}_{s\sim\rho_{k_i},a\sim\bar{\pi}_{k_i}}\left[f_{k_i}^2(s,a)\right] \le \zeta_{k_i}, \mathbb{E}_{s\sim\rho'_{k_i},a\sim\bar{\pi}_{k_i}}\left[f_{k_i}^2(s,a)\right] \le \zeta_{k_i};\right.$$

$$\left.\forall i \in [N], \forall\phi, \|\phi(s,a)\|_{\widehat{\Sigma}^{-1}_{k_i,\phi}} = \Theta\left(\|\phi(s,a)\|_{\Sigma^{-1}_{\rho_{k_i}\times\bar{\pi}_{k_i},\phi}}\right)\right\}.$$

We first consider fixed initial state s_0. For some epoch $i \in [N]$ and episode $k \in \{k_i, k_i+1, \ldots, k_i + L-1\}$, applying Lemma C.4 shows that

$$V_k^{\tilde{\pi}_k}(s_0) - \widehat{V}_k^{\tilde{\pi}_k}(s_0) \tag{17}$$

$$=(1-\gamma)^{-1}\mathbb{E}_{(s,a)\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\widehat{b}_{k_i}(s,a) - \gamma\mathbb{E}_{\widehat{P}_{k_i}(s'|s,a)}\left[\widehat{V}_k^{\tilde{\pi}_k}(s')\right] + \gamma\mathbb{E}_{P^\star(s'|s,a)}\left[\widehat{V}_k^{\tilde{\pi}_k}(s')\right]\right]$$

$$\le \mathbb{E}_{(s,a)\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\frac{1}{1-\gamma}\widehat{b}_{k_i}(s,a) + \frac{2}{(1-\gamma)^3}f_{k_i}(s,a)\right]$$

$$=\underbrace{\gamma\mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\tilde{\pi}_k},s\sim P^\star(\cdot|\tilde{s},\tilde{a}),a\sim\tilde{\pi}_k(s)}\left[\frac{1}{1-\gamma}\widehat{b}_{k_i}(s,a) + \frac{2}{(1-\gamma)^3}f_{k_i}(s,a)\right]}_{\text{TERM}_1}.$$

$$+ (1-\gamma)\mathbb{E}_{s\sim d_0, a\sim\tilde{\pi}_k(s_0)}\underbrace{\left[\frac{1}{1-\gamma}\widehat{b}_{k_i}(s,a) + \frac{2}{(1-\gamma)^3}f_{k_i}(s,a)\right]}_{\text{TERM}_2}, \tag{18}$$

where the inequality follows from the fact that $\|\widehat{b}_{k_i}\|_\infty \leq 2/(1-\gamma)$ and $\|\widehat{V}_{k_i}^{\tilde{\pi}_k}\|_\infty \leq 2/(1-\gamma)^2$ and the second equality is due to Lemma A.4.

TERM_2 can be bounded as follows:

$$\begin{aligned}
\text{TERM}_2 &= \mathbb{E}_{s\sim d_0, a\sim\tilde{\pi}_k(s_0)}\left[\widehat{b}_{k_i}(s,a) + \frac{2}{(1-\gamma)^2}f_{k_i}(s,a)\right] \\
&\leq \sqrt{\frac{A}{(1-\gamma)\xi}\mathbb{E}_{s\sim\rho_{k_i}, a\sim\bar{\pi}_{k_i}(s)}\left[\widehat{b}_{k_i}^2(s,a) + \frac{4}{(1-\gamma)^4}f_{k_i}^2(s,a)\right]} \\
&\lesssim \sqrt{\frac{dA\alpha_{k_i}^2}{(1-\gamma)^3 k_i\xi} + \frac{A\zeta_{k_i}}{(1-\gamma)^5\xi}} \\
&\leq \sqrt{\frac{dA\alpha_{k_i}^2}{(1-\gamma)^3 k_i\xi}} + \sqrt{\frac{A\zeta_{k_i}}{(1-\gamma)^5\xi}},
\end{aligned} \tag{19}$$

where the first inequality follows from Lemma A.1, and the second inequality comes from Lemma C.1 as well as the following inequality:

$$\begin{aligned}
&\mathbb{E}_{s\sim\rho_{k_i}, a\sim\bar{\pi}_{k_i}(s)}\left[k_i\widehat{b}_{k_i}^2(s,a)\right] \\
\leq& \frac{k_i\alpha_{k_i}^2}{(1-\gamma)^2}\mathbb{E}_{s\sim\rho_{k_i}, a\sim\bar{\pi}_{k_i}(s)}\left[\left\|\widehat{\phi}_{k_i}(s,a)\right\|_{\Sigma_{\rho_{k_i}\times\bar{\pi}_{k_i}, \widehat{\phi}_{k_i}}^{-1}}^2\right] \\
=& \frac{k_i\alpha_{k_i}^2}{(1-\gamma)^2}\text{Tr}\left(\mathbb{E}_{s\sim\rho_{k_i}, a\sim\bar{\pi}_{k_i}}\left[\widehat{\phi}_{k_i}(s,a)\widehat{\phi}_{k_i}(s,a)^\top\right]\left\{k_i\mathbb{E}_{s\sim\rho_{k_i}, a\sim\bar{\pi}_{k_i}}\left[\widehat{\phi}_{k_i}(s,a)\widehat{\phi}_{k_i}(s,a)^\top\right] + \lambda_{k_i}I\right\}^{-1}\right) \\
\leq& \frac{\alpha_{k_i}^2 d}{(1-\gamma)^2},
\end{aligned} \tag{20}$$

where the first inequality is because $\widehat{\Sigma}_{k_i, \phi_{k_i}}$ is an unbiased estimate of $\Sigma_{\rho_{k_i}\times\bar{\pi}_{k_i}, \phi_{k_i}}$ and the second inequality is by $\text{Tr}(AB) \geq \text{Tr}(AC)$ for positive semi-definite matrices A, B, C and $B - C \succeq 0$.

To bound TERM_1, we note that

$$\begin{aligned}
&\mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\tilde{\pi}_k}, s\sim P^\star(\cdot|\tilde{s},\tilde{a}), a\sim\tilde{\pi}_k(s)}\left[\widehat{b}_{k_i}(s,a) + \frac{2}{(1-\gamma)^2}f_{k_i}(s,a)\right] \\
\leq& \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma_{\rho_{k_i}, \phi^\star}^{-1}}\right]\sqrt{\frac{2k_iA}{\xi\gamma}\mathbb{E}_{s\sim\rho_{k_i}, a\sim\bar{\pi}_{k_i}(s)}\left[\widehat{b}_{k_i}^2(s,a) + \frac{4}{(1-\gamma)^4}f_{k_i}^2(s,a)\right] + \lambda_{k_i}d\frac{36}{(1-\gamma)^4}} \\
\leq& \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma_{\rho_{k_i}, \phi^\star}^{-1}}\right]\sqrt{\frac{2k_iA}{\xi\gamma}\left(\frac{\alpha_{k_i}^2 d}{k_i(1-\gamma)^2} + \frac{4}{(1-\gamma)^4}\zeta_{k_i}\right) + \lambda_{k_i}d\frac{36}{(1-\gamma)^4}} \\
\lesssim& \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma_{\rho_{k_i}, \phi^\star}^{-1}}\right]\sqrt{\frac{dA\alpha_{k_i}^2}{\xi\gamma(1-\gamma)^2}} + \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma_{\rho_{k_i}, \phi^\star}^{-1}}\right]\sqrt{\frac{k_iA\zeta_{k_i}}{\xi\gamma(1-\gamma)^4} + \frac{\lambda_{k_i}d}{\gamma(1-\gamma)^4}},
\end{aligned}$$

where the first inequality follows from Lemma A.2, the AM-GM inequality, $\|\widehat{b}_{k_i}\|_\infty \leq 2(1-\gamma)$, and $\|f_{k_i}\|_\infty \leq 2$ and the second inequality is due to Eq. (20).

The above display implies that TERM_1 can be bounded as follows:

$$\begin{aligned}
&\gamma\mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\tilde{\pi}_k}, s\sim P^\star(\cdot|\tilde{s},\tilde{a}), a\sim\tilde{\pi}_k(s)}\left[\frac{1}{1-\gamma}\widehat{b}_{k_i}(s,a) + \frac{2}{(1-\gamma)^3}f_{k_i}(s,a)\right] \\
\lesssim& \frac{1}{(1-\gamma)^2}\mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma_{\rho_{k_i}, \phi^\star}^{-1}}\right]\sqrt{\gamma\frac{dA\alpha_{k_i}^2}{\xi}} + \frac{1}{(1-\gamma)^3}\mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma_{\rho_{k_i}, \phi^\star}^{-1}}\right]
\end{aligned}$$

$$\cdot \sqrt{\gamma\left(\frac{k_i A \zeta_{k_i}}{\xi} + \lambda_{k_i} d\right)}$$

$$\leq \frac{1}{(1-\gamma)^2} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma_{\rho_{k_i},\phi^\star}^{-1}}\right] \sqrt{\frac{dA\alpha_K^2}{\xi}} + \frac{1}{(1-\gamma)^3} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma_{\rho_{k_i},\phi^\star}^{-1}}\ \alpha_K\right]$$

$$\lesssim \frac{1}{(1-\gamma)^3} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma_{\rho_{k_i},\phi^\star}^{-1}}\right] \sqrt{\frac{dA\alpha_K^2}{\xi}}\,, \tag{21}$$

where the second inequality is due to that $\alpha_k = O(\sqrt{\gamma(A/\xi + d^2)\ln(Mk/\delta)}) = O(\sqrt{\gamma(A/\xi + d^2)k\zeta_k}) = O(\sqrt{\gamma(Ak\zeta_k/\xi + \lambda_k d)}) = O(\sqrt{\gamma(Ak\zeta_k/\xi + \lambda_k d + k\zeta_k)})$.

Substituting Eq. (19) and Eq. (21) into Eq. (17), and taking summation over $k \in [K]$ leads to

$$\sum_{k=1}^K \left(V_k^{\tilde{\pi}_k} - \widehat{V}_k^{\tilde{\pi}_k}\right)$$

$$= \mathbb{E}_{s_0 \sim d_0}\left[\sum_{k=1}^K \left(V_k^{\tilde{\pi}_k}(s_0) - \widehat{V}_k^{\tilde{\pi}_k}(s_0)\right)\right]$$

$$\lesssim \sum_{i=1}^N \sum_{k=k_i}^{k_i+L-1} \left(\frac{1}{(1-\gamma)^3} \mathbb{E}_{(s,a)\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\|\phi^\star(s,a)\|_{\Sigma_{\rho_{k_i},\phi^\star}^{-1}}\right]\sqrt{\frac{dA\alpha_K^2}{\xi}} + \sqrt{\frac{dA\alpha_K^2}{(1-\gamma)^3 k_i \xi}} + \sqrt{\frac{A\zeta_{k_i}}{(1-\gamma)^5 \xi}}\right)$$

$$\lesssim \frac{\alpha_K}{(1-\gamma)^3}\sqrt{\frac{dA}{\xi}} \cdot \sum_{i=1}^N \sum_{k=k_i}^{k_i+L-1} \mathbb{E}_{(s,a)\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\|\phi^\star(s,a)\|_{\Sigma_{\rho_{k_i},\phi^\star}^{-1}}\right]$$

$$\leq \frac{\alpha_K}{(1-\gamma)^3}\sqrt{\frac{dA}{\xi}} \cdot \sqrt{K\sum_{i=1}^N \sum_{k=k_i}^{k_i+L-1} \mathbb{E}_{(s,a)\sim d_{P^\star}^{\tilde{\pi}_k}}\left[\|\phi^\star(s,a)\|_{\Sigma_{\rho_{k_i},\phi^\star}^{-1}}^2\right]}$$

$$= \frac{\alpha_K}{(1-\gamma)^3}\sqrt{\frac{dA}{\xi}} \cdot \sqrt{K\sum_{i=1}^N \sum_{k=k_i}^{k_i+L-1} \mathrm{Tr}\left(\mathbb{E}_{(s,a)\sim d_{P^\star}^{\tilde{\pi}_k}}[\phi^\star(s,a)\phi^\star(s,a)^\top]\Sigma_{\rho_{k_i},\phi^\star}^{-1}\right)}$$

$$= \frac{\alpha_K}{(1-\gamma)^3}\sqrt{\frac{dA}{\xi}} \cdot \sqrt{K\sum_{j=1}^L \sum_{i=1}^N \mathrm{Tr}\left(\mathbb{E}_{(s,a)\sim d_{P^\star}^{\tilde{\pi}_{(i-1)L+j}}}[\phi^\star(s,a)\phi^\star(s,a)^\top]\Sigma_{\rho_{k_i},\phi^\star}^{-1}\right)}$$

$$\leq \frac{\alpha_K}{(1-\gamma)^3}\sqrt{\frac{dA}{\xi}}$$

$$\cdot \sqrt{K\sum_{j=1}^L \sum_{i=1}^N \mathrm{Tr}\left(\mathbb{E}_{(s,a)\sim d_{P^\star}^{\tilde{\pi}_{(i-1)L+j}}}[\phi^\star(s,a)\phi^\star(s,a)^\top]\left(\sum_{q=1}^{i-1}\mathbb{E}_{(s,a)\sim d_{P^\star}^{\tilde{\pi}_{(q-1)L+j}}}[\phi^\star(s,a)\phi^\star(s,a)^\top] + \lambda_1 I\right)^{-1}\right)}$$

$$\lesssim \frac{\alpha_K}{(1-\gamma)^3}\sqrt{\frac{dA}{\xi}} \cdot \sqrt{K\sum_{j=1}^L d\ln\left(1+\frac{N}{d\lambda_1}\right)}$$

$$\lesssim \frac{\alpha_K}{(1-\gamma)^3}\sqrt{\frac{dA}{\xi}} \cdot \sqrt{dLK\ln\left(1+\frac{K}{d\lambda_1}\right)}$$

$$\lesssim \frac{d^2 A\sqrt{KL}}{\xi(1-\gamma)^3}\sqrt{\ln(1+K)\ln(MK/\delta)}\,,$$

where the third inequality follows from Cauchy–Schwarz inequality together with Jensen's inequality, the fourth inequality is by $\mathrm{Tr}(AB) \geq \mathrm{Tr}(AC)$ for positive semi-definite matrices A, B, C and $B - C \succeq 0$, and the fifth inequality is due to Lemma C.2. The proof is now completed. $\qquad\square$

A.4 One-step-back Inequalities

We first present the following lemma, which bounds the quantity under the initial state distribution, for any policy π. Note that this lemma holds for any $k \in [K]$.

Lemma A.1. *For any $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that $\|g\|_\infty \leq B$ and any policy π, it holds that*

$$\mathbb{E}_{s \sim d_0, a \sim \pi(s_0)} [g(s,a)] \leq \sqrt{\frac{A}{(1-\gamma)\xi} \mathbb{E}_{s \sim \rho_k, a \sim \bar{\pi}_k} [g^2(s,a)]}.$$

Proof.

$$\mathbb{E}_{s \sim d_0, a \sim \pi(s_0)} [g(s,a)] \leq \sqrt{\mathbb{E}_{s \sim d_0, a \sim \pi(s_0)} [g^2(s,a)]}$$

$$\leq \sqrt{\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{d_0(s)\pi(a|s)}{\rho_k(s)\bar{\pi}_k(a|s)} \mathbb{E}_{s \sim \rho_k, a \sim \bar{\pi}_k} [g^2(s,a)]}$$

$$\leq \sqrt{\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{d_0(s)\pi(a|s)}{(1-\gamma)d_0(s)\bar{\pi}_k(a|s)} \mathbb{E}_{s \sim \rho_k, a \sim \bar{\pi}_k} [g^2(s,a)]}$$

$$\leq \sqrt{\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{1}{(1-\gamma)\xi \cdot U(a)} \mathbb{E}_{s \sim \rho_k, a \sim \bar{\pi}_k} [g^2(s,a)]}$$

$$\leq \sqrt{\frac{A}{(1-\gamma)\xi} \mathbb{E}_{s \sim \rho_k, a \sim \bar{\pi}_k} [g^2(s,a)]},$$

where the first inequality follows from Jensen's inequality, the second inequality is by importance sampling, and the fourth inequality is due to the definition of $\bar{\pi}_k$. $\square$

The following lemma shows that

$$\mathbb{E}_{(s,a) \sim d_{P^\star}^{\tilde{\pi}_k}} [g(s,a)] \lesssim \mathbb{E}_{(s,a) \sim d_{P^\star}^{\tilde{\pi}_k}} \left[\|\phi^\star(s,a)\|_{\Sigma_{\rho_k, \phi^\star}^{-1}} \right],$$

if $\mathbb{E}_{s \sim \rho_k, a \sim \bar{\pi}_k} [g^2(s,a)]$ is upper bounded.

Lemma A.2 (One-step-back inequality in the true model). *For any $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that $\|g\|_\infty \leq B$, any epoch $i \in [N]$ and any episode $k \in \{k_i, \ldots, k_i + L - 1\}$, it holds that*

$$\mathbb{E}_{(\tilde{s},\tilde{a}) \sim d_{P^\star}^{\tilde{\pi}_k}, s \sim P^\star(\cdot|\tilde{s},\tilde{a}), a \sim \tilde{\pi}_k(s)} [g(s,a)]$$

$$\leq \mathbb{E}_{(\tilde{s},\tilde{a}) \sim d_{P^\star}^{\tilde{\pi}_k}} \left[\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma_{\rho_{k_i}, \phi^\star}^{-1}} \right] \sqrt{\frac{k_i A}{\xi \gamma} \mathbb{E}_{s \sim \rho_{k_i}, a \sim \bar{\pi}_{k_i}} [g^2(s,a)] + \lambda_{k_i} d B^2}.$$

Proof. To begin with, applying the Cauchy–Schwarz inequality shows that

$$\mathbb{E}_{(\tilde{s},\tilde{a}) \sim d_{P^\star}^{\tilde{\pi}_k}, s \sim P^\star(\cdot|\tilde{s},\tilde{a}), a \sim \tilde{\pi}_k(s)} [g(s,a)]$$

$$= \mathbb{E}_{(\tilde{s},\tilde{a}) \sim d_{P^\star}^{\tilde{\pi}_k}} \left[\phi^\star(\tilde{s},\tilde{a})^\top \int \sum_a \mu^\star(s)\tilde{\pi}_k(a \mid s) g(s,a) d(s) \right]$$

$$\leq \mathbb{E}_{(\tilde{s},\tilde{a}) \sim d_{P^\star}^{\tilde{\pi}_k}} \left[\|\phi^\star(\tilde{s},\tilde{a})\|_{\Sigma_{\rho_{k_i}, \phi^\star}^{-1}} \left\| \int \sum_a \mu^\star(s)\tilde{\pi}_k(a \mid s) g(s,a) d(s) \right\|_{\Sigma_{\rho_{k_i}, \phi^\star}} \right]. \quad (22)$$

We bound the second quadratic form w.r.t. $\Sigma_{\rho_{k_i}, \phi^\star}$ in Eq. (22) as follows:

$$\left\| \int \sum_a \mu^\star(s)\tilde{\pi}_k(a \mid s) g(s,a) d(s) \right\|_{\Sigma_{\rho_{k_i}, \phi^\star}}^2$$

25

$$= \left[\int \sum_a \mu^\star(s) \tilde{\pi}_k(a \mid s) g(s,a) d(s) \right]^\top \left\{ k_i \mathbb{E}_{(s,a)\sim\rho_{k_i}} \left[\phi^\star(s,a) \phi^\star(s,a)^\top \right] + \lambda_k I \right\}$$

$$\left[\int \sum_a \mu^\star(s) \tilde{\pi}_k(a \mid s) g(s,a) d(s) \right]$$

$$\leq k_i \mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_{k_i}} \left\{ \left[\int \sum_a \mu^\star(s)^\top \phi^\star(\tilde{s},\tilde{a}) \tilde{\pi}_k(a \mid s) g(s,a) d(s) \right]^2 \right\} + \lambda_{k_i} dB^2$$

$$= k_i \mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_{k_i}} \left\{ \mathbb{E}_{s\sim P^\star(\cdot|\tilde{s},\tilde{a}), a\sim\tilde{\pi}_k(s)} \left[g(s,a) \right]^2 \right\} + \lambda_{k_i} dB^2$$

$$\leq k_i \mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_{k_i}, s\sim P^\star(\cdot|\tilde{s},\tilde{a}), a\sim\tilde{\pi}_k(s)} \left[g^2(s,a) \right] + \lambda_{k_i} dB^2 , \tag{23}$$

where the first inequality comes from $\|g(s,a)\|_\infty \leq B$ together with the regularity condition in Assumption 2.1 that $\left\| \int \mu^\star(s) h(s) \mathrm{d}(s) \right\|_2 \leq \sqrt{d}$ for any $h : \mathcal{S} \to [0,1]$, and the last inequality follows from the Jensen's inequality.

By importance sampling, it is clear that

$$k_i \mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_{k_i}, s\sim P^\star(\cdot|\tilde{s},\tilde{a}), a\sim\tilde{\pi}_k(s)} \left[g^2(s,a) \right] + \lambda_{k_i} dB^2$$

$$\leq \frac{k_i}{\gamma} \mathbb{E}_{s\sim\rho_{k_i}, a\sim\tilde{\pi}_k(s)} \left[g^2(s,a) \right] + \lambda_{k_i} dB^2$$

$$\leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{k_i}{\gamma} \frac{\tilde{\pi}_k(a \mid s)}{\bar{\pi}_{k_i}(a \mid s)} \mathbb{E}_{s\sim\rho_{k_i}, a\sim\bar{\pi}_{k_i}(s)} \left[g^2(s,a) \right] + \lambda_{k_i} dB^2$$

$$\leq \frac{k_i A}{\xi\gamma} \mathbb{E}_{s\sim\rho_{k_i}, a\sim\bar{\pi}_{k_i}(s)} \left[g^2(s,a) \right] + \lambda_{k_i} dB^2 , \tag{24}$$

where the third inequality is due to the definition of $\bar{\pi}_k$, and the first inequality is because

$$\gamma \mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_{k_i}, s\sim P^\star(\cdot|\tilde{s},\tilde{a}), a\sim\tilde{\pi}_k(s)} \left[g^2(s,a) \right]$$

$$\leq \gamma \mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_{k_i}, s\sim P^\star(\cdot|\tilde{s},\tilde{a}), a\sim\tilde{\pi}_k(s)} \left[g^2(s,a) \right] + (1-\gamma) \mathbb{E}_{s_0\sim d_0, a\sim\tilde{\pi}_k(s)} \left[g^2(s,a) \right]$$

$$= \mathbb{E}_{s\sim\rho_{k_i}, a\sim\tilde{\pi}_k(s)} \left[g^2(s,a) \right] ,$$

which comes from Lemma A.4.

The proof is concluded by substituting Eq. (23) and Eq. (24) into Eq. (22). $\qquad\square$

The following lemma is a counterpart of Lemma A.2, which shows that

$$\mathbb{E}_{(s,a)\sim d^\pi_{\widehat{P}_k}} \left[g(s,a) \right] \lesssim \mathbb{E}_{(s,a)\sim d^\pi_{\widehat{P}_k}} \left[\left\| \widehat{\phi}_k(s,a) \right\|_{\Sigma^{-1}_{\rho_k \times \bar{\pi}_k, \widehat{\phi}_k}} \right] ,$$

if $\mathbb{E}_{s\sim\rho'_k, a\sim\bar{\pi}_k} \left[g^2(s,a) \right]$ is upper bounded. Note that compared with Lemma A.2, this lemma additionally needs to condition on the event that the MLE guarantee (*cf.*, Lemma C.1) holds.

Lemma A.3 (One-step-back inequality in the learned model). *Conditioned on the event where the MLE guarantee in Lemma C.1 holds, i.e., $\mathbb{E}_{s\sim\rho_{k_i}, a\sim\bar{\pi}_{k_i}} \left[f_{k_i}(s,a)^2 \right] \lesssim \zeta_{k_i}$, for any epoch $i \in [N]$. Then for any $g : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ such that $\|g\|_\infty \leq B$, any epoch $i \in [N]$ and any policy π, it holds that*

$$\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\widehat{P}_{k_i}}, s\sim\widehat{P}_{k_i}(\cdot|\tilde{s},\tilde{a}), a\sim\pi(s)} \left[g(s,a) \right]$$

$$\leq \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\widehat{P}_{k_i}}} \left[\left\| \widehat{\phi}_{k_i}(\tilde{s},\tilde{a}) \right\|_{\Sigma^{-1}_{\rho_{k_i} \times \bar{\pi}_{k_i}, \widehat{\phi}_{k_i}}} \right] \sqrt{ \frac{k_i A}{\xi} \mathbb{E}_{s\sim\rho'_{k_i}, a\sim\bar{\pi}_{k_i}} \left[g^2(s,a) \right] + B^2 \lambda_{k_i} d + k_i B^2 \zeta_{k_i} } ,$$

where recall that $\bar{\pi}_k(\cdot \mid s) = \xi \cdot U(\mathcal{A}) + (1-\xi) \cdot 1/k \sum_{j=1}^k \tilde{\pi}_j(\cdot \mid s)$.

Proof. The proof of this lemma is generally similar to that of Lemma A.2. We start by applying the Cauchy–Schwarz inequality:

$$\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\widehat{P}_{k_i}}, s\sim\widehat{P}_{k_i}(\cdot|\tilde{s},\tilde{a}), a\sim\pi(s)} \left[g(s,a) \right]$$

$$=\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\widehat{P}_{k_i}}}\left[\widehat{\phi}_{k_i}(\tilde{s},\tilde{a})^{\top}\int\sum_a\widehat{\mu}_{k_i}(s)\pi(a\mid s)g(s,a)d(s)\right]$$

$$\le\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\widehat{P}_{k_i}}}\left[\left\|\widehat{\phi}_{k_i}(\tilde{s},\tilde{a})\right\|_{\Sigma^{-1}_{\rho_{k_i}\times\bar{\pi}_{k_i},\widehat{\phi}_{k_i}}}\left\|\int\sum_a\widehat{\mu}_{k_i}(s)\pi(a\mid s)g(s,a)d(s)\right\|_{\Sigma_{\rho_{k_i}\times\bar{\pi}_{k_i},\widehat{\phi}_{k_i}}}\right]. \quad (25)$$

We now bound the second quadratic form w.r.t. $\Sigma_{\rho_{k_i}\times\bar{\pi}_{k_i},\widehat{\phi}_{k_i}}$ in Eq. (25) as follows:

$$\left\|\int\sum_a\widehat{\mu}_{k_i}(s)\pi(a\mid s)g(s,a)d(s)\right\|^2_{\Sigma_{\rho_{k_i}\times\bar{\pi}_{k_i},\widehat{\phi}_{k_i}}}$$

$$=\left[\int\sum_a\widehat{\mu}_{k_i}(s)\pi(a\mid s)g(s,a)d(s)\right]^{\top}\cdot\left\{k_i\mathbb{E}_{s\sim\rho_{k_i},a\sim\bar{\pi}_{k_i}}\left[\widehat{\phi}_{k_i}(s,a)\widehat{\phi}_{k_i}(s,a)^{\top}\right]+\lambda_{k_i}I\right\}$$

$$\cdot\left[\int\sum_a\widehat{\mu}_{k_i}(s)\pi(a\mid s)g(s,a)d(s)\right]$$

$$\le k_i\mathbb{E}_{\tilde{s}\sim\rho_{k_i},\tilde{a}\sim\bar{\pi}_{k_i}}\left\{\left[\int\sum_a\widehat{\mu}_{k_i}(s)^{\top}\widehat{\phi}_{k_i}(\tilde{s},\tilde{a})\pi(a\mid s)g(s,a)d(s)\right]^2\right\}+B^2\lambda_{k_i}d$$

$$=k_i\mathbb{E}_{\tilde{s}\sim\rho_{k_i},\tilde{a}\sim\bar{\pi}_{k_i}}\left\{\mathbb{E}_{s\sim\widehat{P}_{k_i}(\cdot|\tilde{s},\tilde{a}),a\sim\pi(s)}[g(s,a)]^2\right\}+B^2\lambda_{k_i}d\,, \quad (26)$$

where the first inequality is because $\|g(s,a)\|_{\infty}\le B$ as well as the regularity condition in Assumption 2.1 that $\left\|\int\mu(s)h(s)\mathrm{d}(s)\right\|_2\le\sqrt{d}$ for any $h:\mathcal{S}\to[0,1]$ and any $\mu\in\Psi$.

Moreover, using the MLE guarantee in Lemma C.1, we have that

$$k_i\mathbb{E}_{\tilde{s}\sim\rho_{k_i},\tilde{a}\sim\bar{\pi}_{k_i}}\left\{\mathbb{E}_{s\sim\widehat{P}_{k_i}(\cdot|\tilde{s},\tilde{a}),a\sim\pi(s)}[g(s,a)]^2\right\}+B^2\lambda_{k_i}d$$

$$\le k_i\mathbb{E}_{\tilde{s}\sim\rho_{k_i},\tilde{a}\sim\bar{\pi}_{k_i}}\left\{\mathbb{E}_{s\sim P^{\star}(\cdot|\tilde{s},\tilde{a}),a\sim\pi(s)}[g(s,a)]^2\right\}+B^2\lambda_{k_i}d+k_iB^2\zeta_{k_i}$$

$$\le k_i\mathbb{E}_{\tilde{s}\sim\rho_{k_i},\tilde{a}\sim\bar{\pi}_{k_i},s\sim P^{\star}(\cdot|\tilde{s},\tilde{a}),a\sim\pi(s)}\left[g^2(s,a)\right]+B^2\lambda_{k_i}d+B^2k_i\zeta_{k_i}$$

$$\le\frac{k_iA}{\xi}\mathbb{E}_{\tilde{s}\sim\rho_{k_i},\tilde{a}\sim\bar{\pi}_{k_i},s\sim P^{\star}(\cdot|\tilde{s},\tilde{a}),a\sim\bar{\pi}_{k_i}}\left[g^2(s,a)\right]+B^2\lambda_{k_i}d+B^2k_i\zeta_{k_i}$$

$$\le\frac{k_iA}{\xi}\mathbb{E}_{s\sim\rho'_{k_i},a\sim\bar{\pi}_{k_i}}\left[g^2(s,a)\right]+B^2\lambda_{k_i}d+B^2k_i\zeta_{k_i}\,, \quad (27)$$

where the second inequality follows by Jensen's inequality, the third inequality comes from importance sampling and the definition of $\bar{\pi}_{k_i}$, and the last inequality is due to the definition of ρ'_{k_i}.

The proof is now concluded by substituting Eq. (26) and Eq. (27) into Eq. (25). $\qquad\square$

The following lemma shows that the expectation of any state-action function $g:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ w.r.t. $d^{\pi_1}_P$ and π_2 can be decomposed into (a) the one-step-back expectation of g w.r.t. $d^{\pi_1}_P$ and π_2; and (b) the expectation of g w.r.t. d_0 and π_2.

Lemma A.4. *For any P, and any policy π_1 and π_2, it holds that*

$$\mathbb{E}_{s\sim d^{\pi_1}_P,a\sim\pi_2(\cdot|s)}[g(s,a)]$$
$$=\gamma\mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi_1}_P,s\sim P(\cdot|\tilde{s},\tilde{a}),a\sim\pi_2(\cdot|s)}[g(s,a)]+(1-\gamma)\mathbb{E}_{s\sim d_0,a\sim\pi_2(\cdot|s_0)}[g(s,a)]\,.$$

Proof.

$$\mathbb{E}_{s\sim d^{\pi_1}_P,a\sim\pi_2(\cdot|s)}[g(s,a)]$$
$$=\sum_{t=1}^{\infty}(1-\gamma)\gamma^t\mathbb{E}_{s\sim d^{\pi_1}_{P,t},a\sim\pi_2(\cdot|s)}[g(s,a)]+(1-\gamma)\mathbb{E}_{s\sim d^{\pi_1}_{P,0},a\sim\pi_2(\cdot|s)}[g(s,a)]$$

$$=\gamma\sum_{t=0}^{\infty}(1-\gamma)\gamma^t\mathbb{E}_{s\sim d_{P,t+1}^{\pi_1},a\sim\pi_2(\cdot|s)}[g(s,a)]+(1-\gamma)\mathbb{E}_{s\sim d_0,a\sim\pi_2(\cdot|s_0)}[g(s,a)]$$

$$=\gamma\sum_{t=0}^{\infty}(1-\gamma)\gamma^t\mathbb{E}_{(\tilde{s},\tilde{a})\sim d_{P,t}^{\pi_1},s\sim P(\cdot|\tilde{s},\tilde{a}),a\sim\pi_2(\cdot|s)}[g(s,a)]+(1-\gamma)\mathbb{E}_{s\sim d_0,a\sim\pi_2(\cdot|s_0)}[g(s,a)]$$

$$=\gamma\mathbb{E}_{(\tilde{s},\tilde{a})\sim d_P^{\pi_1},s\sim P(\cdot|\tilde{s},\tilde{a}),a\sim\pi_2(\cdot|s)}[g(s,a)]+(1-\gamma)\mathbb{E}_{s\sim d_0,a\sim\pi_2(\cdot|s_0)}[g(s,a)].$$

$\square$

B Omitted Analysis of The Regret Lower Bound

In this section, we provide the proof of Theorem 4.2. For the remainder of this section, we switch from loss functions to reward functions for convenience since we now consider MDPs with fixed loss functions.

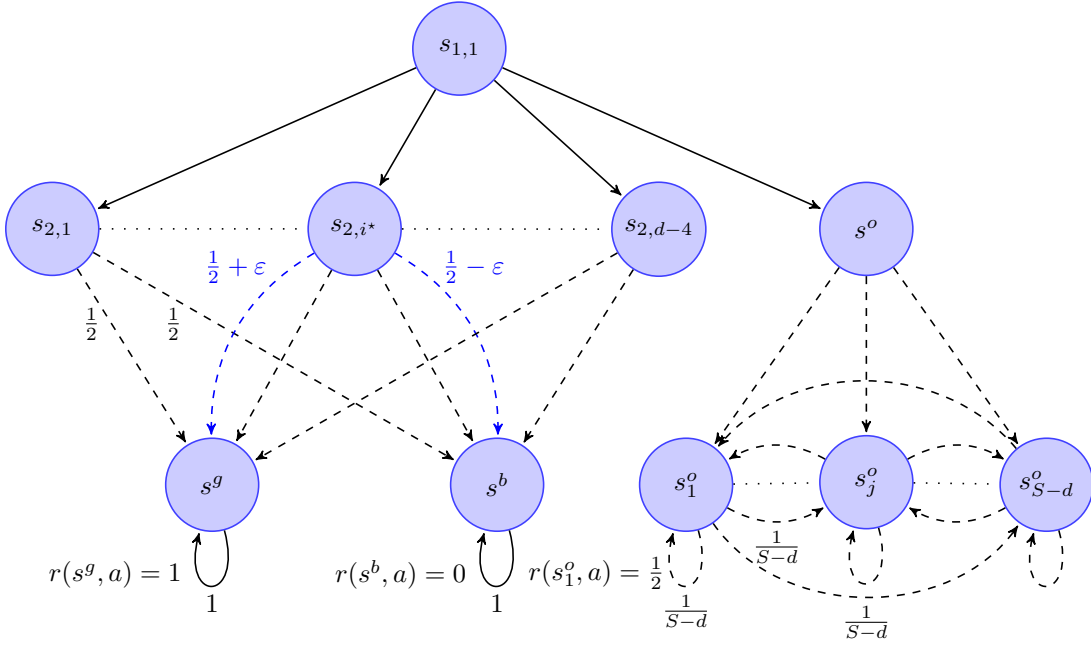### B.1 Construction of Hard-to-learn MDP Instances

Figure 1: The class of the hard-to-learn low-rank MDP instances used in the proof of Theorem 4.2.

To prove our regret lower bound in Theorem 4.2, we construct a class of hard-to-learn low-rank MDP instances, as shown in Figure 1. Note that similar hard MDP instances are first introduced to prove the regret lower bounds for tabular MDPs [Lattimore and Szepesvári, 2020, Domingues et al., 2021] and are recently also used to prove the lower bound of sample complexity for learning low-rank MDPs by Cheng et al. [2023].

To begin with, we first introduce the reference low-rank MDP $\mathcal{M}_0$, with its elements detailed as follows:

- State space: $\mathcal{S}=\{s_{1,1},s_{2,1},s_{2,2},\ldots,s_{2,d-4}\}\cup\{s^g,s^b\}\cup\{s^o\}\cup\mathcal{S}_\mathcal{O}$, where $\mathcal{S}_\mathcal{O}=\{s_i^o\}_{i=1}^{S-d}$ denotes the set of "outlier states", s^g denotes the "good state", and s^b denotes the "bad state".
- Action space: $\mathcal{A}=\{a_1,a_2,\ldots,a_A\}$.
- Reward function: $r(s,a)=\mathbb{I}\{s=s^g\}+\frac{1}{2}\mathbb{I}\{s\in\mathcal{S}_\mathcal{O}\}$.

28

- Transitions:
  - For the initial state $s_{1,1}$, the learner will deterministically transit to state $s_{2,i}$ if taking action a_i, $\forall i \in [d-4]$, and will transit to state s^o otherwise. Formally, $P^\star(s_{2,i} \mid s_{1,1}, a_i) = 1, \forall i \in [d-4]$, and $P^\star(s^o \mid s_{1,1}, a_i) = 1, \forall i \in [A] \setminus [d-4]$.
  - For state $s_{2,i} \in \{s_{2,1}, s_{2,2}, \ldots, s_{2,d-4}\}$, the learner will transit to good state s^g and bad state s^b uniformly at random, no matter what action it takes, *i.e.*, $P^\star(s^g \mid s_{2,i}, a) = P^\star(s^b \mid s_{2,i}, a) = \frac{1}{2}, \forall i \in [d-4]$ and $a \in \mathcal{A}$.
  - For states s^o and $s^o_i \in \mathcal{S}_{\mathcal{O}}$, the learner will uniformly transit to a state $s^o_j \in \mathcal{S}_{\mathcal{O}}$, no matter what action it takes. Formally, $P^\star(s^o_j \mid s^o, a) = P^\star(s^o_j \mid s^o_i, a) = \frac{1}{S-d}$, $\forall s^o_i, s^o_j \in \mathcal{S}_{\mathcal{O}}$ and $a \in \mathcal{A}$.
  - For states s^g and s^b, the learner will stay at the current state no matter what action it takes, which means that $P^\star(s^g \mid s^g, a) = P^\star(s^b \mid s^b, a) = 1$.

Further, the transitions of the above MDP can be realized by $P^\star(s' \mid s, a) = \langle \phi^\star(s,a), \mu^\star(s') \rangle$, with the following features, which thus implies that this MDP is indeed a low-rank MDP:

$$\phi^\star(s_{1,1}, a_i) = \boldsymbol{e}_i, \quad \mu^\star(s_{2,i}) = \boldsymbol{e}_i, \forall i \in [d-4], \quad \mu^\star(s_{1,1}) = \boldsymbol{0}$$

$$\phi^\star(s_{1,1}, a_i) = (0, \ldots, 0, 1, 0), \forall i \in [A] \setminus [d-4], \quad \mu^\star(s^o) = (0, \ldots, 0, 1, 0)$$

$$\phi^\star(s_{2,j}, a) = (0, \ldots, 0, \frac{1}{2}, \frac{1}{2}, 0, 0), \forall a \in \mathcal{A}, \quad \mu^\star(s^g) = (0, \ldots, 0, 1, 0, 0, 0), \quad \mu^\star(s^b) = (0, \ldots, 0, 0, 1, 0, 0)$$

$$\phi^\star(s^o, a) = \phi^\star(s^o_j, a) = (0, \ldots, 0, \frac{1}{S-d}), \forall a \in \mathcal{A}, \quad \mu^\star(s^o_j) = (0, \ldots, 0, 1)$$

$$\phi^\star(s^g, a) = \mu^\star(s^g), \quad \phi^\star(s^b, a) = \mu^\star(s^b), \forall a \in \mathcal{A}.$$

Based on the reference MDP $\mathcal{M}_0$, we define other low-rank MDP instances $\mathcal{M}_{(i^\star, a^\star)}, \forall (i^\star, a^\star) \in [d-4] \times \mathcal{A}$. In specific, the only difference between $\mathcal{M}_{(i^\star, a^\star)}$ and $\mathcal{M}_0$ is that $\phi^\star(s_{2,i^\star}, a^\star) = (0, \ldots, 0, \frac{1}{2} + \varepsilon, \frac{1}{2} - \varepsilon, 0, 0)$, such that $P^\star(s_g \mid s_{2,i^\star}, a^\star) = \frac{1}{2} + \varepsilon$, and $P^\star(s_b \mid s_{2,i^\star}, a^\star) = \frac{1}{2} - \varepsilon$, for some $\varepsilon > 0$ to be defined later.

B.2 Proof of Theorem 4.2

Based on the class of hard-to-learn low-rank MDP instances constructed above, we are now ready to prove the regret lower bound in Theorem 4.2.

Proof of Theorem 4.2. In what follows, we denote by $\mathbb{P}_{(i^\star, a^\star)} := \mathbb{P}_{\mathrm{Alg}, \mathcal{M}_{(i^\star, a^\star)}}$ the probability measure over the outcomes induced by the interaction between Alg and $\mathcal{M}_{(i^\star, a^\star)}$, and by $\mathbb{E}_{(i^\star, a^\star)} := \mathbb{E}_{\mathrm{Alg}, \mathcal{M}_{(i^\star, a^\star)}}$ the expectation with respect to $\mathbb{P}_{(i^\star, a^\star)}$.

Regret of Alg **in** $\mathcal{M}_{(i^\star, a^\star)}$ For some $\mathcal{M}_{(i^\star, a^\star)}$, its optimal policy $\pi^\star_{(i^\star, a^\star)} : \mathcal{S} \to \mathcal{A}$ satisfies that $\pi^\star_{(i^\star, a^\star)}(s_{1,1}) = a_{i^\star}$ and $\pi^\star_{(i^\star, a^\star)}(s_{2,i^\star}) = a^\star$, with the optimal value function

$$V_0^\star(s_{1,1}) = \mathbb{E}\left[\sum_{\tau=0}^{+\infty} \gamma^\tau r(s_\tau, a_\tau) \mid \pi^\star_{(i^\star, a^\star)}, P^\star_{(i^\star, a^\star)}, s_0 = s_{1,1}\right] = \sum_{\tau=2}^{+\infty} \gamma^\tau \left(\frac{1}{2} + \varepsilon\right) = \frac{\gamma^2}{1-\gamma}\left(\frac{1}{2} + \varepsilon\right). \tag{28}$$

For some policy π, it is also clear that its value function satisfies

$$V_0^\pi(s_{1,1}) = \frac{\gamma^2}{1-\gamma}\left(\frac{1}{2} + \varepsilon \mathbb{P}_{(i^\star, a^\star)}((s_2, a_2) = (s_{2,i^\star}, a^\star))\right). \tag{29}$$

Combining Eq. (28) and (29) shows that the regret of Alg in $\mathcal{M}_{(i^\star, a^\star)}$ satisfies

$$\mathcal{R}_K(\mathrm{Alg}, \mathcal{M}_{(i^\star, a^\star)}) = \frac{\gamma^2 \varepsilon}{1-\gamma} K \left(1 - \frac{1}{K} \mathbb{E}_{(i^\star, a^\star)}\left[\sum_{k=1}^K \mathbb{I}\{(s_2^k, a_2^k) = (s_{2,i^\star}, a^\star)\}\right]\right)$$

$$= \frac{\gamma^2 \varepsilon}{1-\gamma} K \left(1 - \frac{1}{K} \mathbb{E}_{(i^\star, a^\star)}\left[N_{(i^\star, a^\star)}^K\right]\right),$$

where we define $N_{(i^\star, a^\star)}^K := \sum_{k=1}^K \mathbb{I}\{(s_2^k, a_2^k) = (s_{2,i^\star}, a^\star)\}$.

Maximum Regret of Alg **over All Possible** $\mathcal{M}_{(i^\star,a^\star)}$ With $\mathcal{R}_K(\text{Alg},\mathcal{M}_{(i^\star,a^\star)})$ in the above equation, we can deduce that

$$\max_{(i^\star,a^\star)} \mathcal{R}_K(\text{Alg},\mathcal{M}_{(i^\star,a^\star)}) \geq \frac{1}{(d-4)A} \sum_{(i^\star,a^\star)} \mathcal{R}_K(\text{Alg},\mathcal{M}_{(i^\star,a^\star)})$$

$$\geq \frac{\gamma^2 \varepsilon}{1-\gamma} K \left(1 - \frac{1}{K(d-4)A} \sum_{(i^\star,a^\star)} \mathbb{E}_{(i^\star,a^\star)} \left[N^K_{(i^\star,a^\star)} \right] \right). \quad (30)$$

To lower bound the above display, it remains to upper bound $\sum_{(i^\star,a^\star)} \mathbb{E}_{(i^\star,a^\star)} \left[N^K_{(i^\star,a^\star)} \right]$. To this end, by Lemma 1 in the work of Garivier et al. [2019] together with the fact that $N^K_{(i^\star,a^\star)}/K \in [0,1]$, it holds that

$$\text{KL}\left(\text{Ber}\left(\frac{1}{K}\mathbb{E}_0\left[N^K_{(i^\star,a^\star)} \right] \right), \text{Ber}\left(\frac{1}{K}\mathbb{E}_{(i^\star,a^\star)}\left[N^K_{(i^\star,a^\star)} \right] \right) \right) \leq \text{KL}\left(\mathbb{P}_0, \mathbb{P}_{(i^\star,a^\star)} \right).$$

This implies that

$$\frac{1}{K}\mathbb{E}_{(i^\star,a^\star)}\left[N^K_{(i^\star,a^\star)} \right] \leq \frac{1}{K}\mathbb{E}_0\left[N^K_{(i^\star,a^\star)} \right] + \sqrt{\frac{1}{2}\text{KL}\left(\mathbb{P}_0, \mathbb{P}_{(i^\star,a^\star)} \right)}$$

$$= \frac{1}{K}\mathbb{E}_0\left[N^K_{(i^\star,a^\star)} \right] + \varepsilon\sqrt{2}\sqrt{\mathbb{E}_0\left[N^K_{(i^\star,a^\star)} \right]},$$

where the inequality is due to Pinsker's inequality that $(p-q)^2 \leq \frac{1}{2}\text{KL}(\text{Ber}(p),\text{Ber}(q))$, for $p,q \in [0,1]$, and the equality comes from Lemma 15.1 of Lattimore and Szepesvári [2020] and Lemma 14 of Domingues et al. [2021] as well as assuming $0 \leq \varepsilon \leq \frac{1}{4}$.

Based on this, one can see that

$$\frac{1}{K}\sum_{(i^\star,a^\star)} \mathbb{E}_{(i^\star,a^\star)}\left[N^K_{(i^\star,a^\star)} \right] \leq \frac{1}{K}\sum_{(i^\star,a^\star)} \mathbb{E}_0\left[N^K_{(i^\star,a^\star)} \right] + \varepsilon\sqrt{2}\sum_{(i^\star,a^\star)} \sqrt{\mathbb{E}_0\left[N^K_{(i^\star,a^\star)} \right]}$$

$$\leq 1 + \varepsilon\sqrt{2}\sqrt{(d-4)AK}, \quad (31)$$

where the second inequality follows from using the Cauchy-Schwartz inequality together with the fact that $N^K_{(i^\star,a^\star)} \leq K$.

Optimizing ε to Lower Bound the Maximum Regret Substituting Eq. (31) into Eq. (30) leads to

$$\max_{(i^\star,a^\star)} \mathcal{R}_K(\text{Alg},\mathcal{M}_{(i^\star,a^\star)}) \geq \frac{\gamma^2 \varepsilon}{1-\gamma} K \left(1 - \frac{1}{(d-4)A} - \varepsilon\sqrt{2}\sqrt{\frac{K}{(d-4)A}} \right)$$

$$\geq \frac{1}{4\sqrt{2}} \cdot \frac{\gamma^2}{1-\gamma} \left(1 - \frac{1}{(d-4)A} \right)^2 \sqrt{(d-4)AK}$$

$$\geq \frac{361}{1600\sqrt{2}} \cdot \frac{\gamma^2}{1-\gamma} \sqrt{(d-4)AK},$$

where the second inequality comes from by choosing $\varepsilon = \frac{1}{2\sqrt{2}}\left(1 - \frac{1}{(d-4)A} \right)\sqrt{\frac{(d-4)A}{K}}$ and the last inequality is due to $d \geq 8$ and $A \geq d-3$. Finally, note that $\varepsilon \leq \frac{1}{4}$ is guaranteed when $K \geq 2(d-4)A$. The proof is thus concluded. $\qquad\square$

C Auxiliary Lemmas

We first introduce the concentration of MLE, the i.i.d. version of which at least dates back to Chapter 7 of Geer [2000] and the non-i.i.d. version of which is first proved by Agarwal et al. [2020] and also appears in the analysis of Uehara et al. [2022].

Lemma C.1 (MLE guarantee). *For some fixed epoch $i \in [N]$, with probability $1 - \delta$, it holds that*

$$\mathbb{E}_{s \sim \left\{ 0.5\rho_{k_i} + 0.5\rho'_{k_i} \right\}, a \sim \bar{\pi}_{k_i}(s)} \left[\left\| \widehat{P}_{k_i}(\cdot \mid s, a) - P^\star(\cdot \mid s, a) \right\|_1^2 \right] \lesssim \zeta, \quad \zeta := \frac{\ln(M/\delta)}{k_i}.$$

Therefore, simultaneously for all epoch $i \in [N]$, with probability $1 - \delta$, it holds that

$$\mathbb{E}_{s \sim \left\{ 0.5\rho_{k_i} + 0.5\rho'_{k_i} \right\}, a \sim \bar{\pi}_{k_i}(s)} \left[\left\| \widehat{P}_{k_i}(\cdot \mid s, a) - P^\star(\cdot \mid s, a) \right\|_1^2 \right] \lesssim \zeta_{k_i}, \quad \zeta_{k_i} := \frac{\ln(MN/\delta)}{k_i}.$$

The following lemma is the canonical elliptical potential lemma.

Lemma C.2 (Lemma 19.4, Lattimore and Szepesvári [2020]). *Let $M_0 = \lambda_0 I \in \mathbb{R}^{d \times d}$ with $\lambda_0 > 0$ and $M_k = M_{k-1} + G_k$, where G_k is positive definite with the maximum eigenvalue $\lambda_{\max}(G_k) \leq 1$ and $\mathrm{Tr}(G_k) \leq B^2$. Then*

$$\sum_{k=1}^K \mathrm{Tr}(G_k M_{k-1}^{-1}) \leq 2 \ln \det(M_K) - 2 \ln \det(M_0) \leq 2d \ln \left(1 + \frac{KB^2}{d\lambda_0} \right).$$

The following lemma guarantees the concentration of the empirical feature covariance matrix and the version for fixed feature mapping $\phi(\cdot)$ is first proved by Zanette et al. [2021]. The proof of this lemma can be readily obtained by taking a union bound over any $\phi \in \Phi$ in the proof of Lemma 39 of Zanette et al. [2021].

Lemma C.3. *Let $\lambda_{k_i} = \Theta(d \ln(k_i |\Phi|/\delta)) = \Theta(d \ln(k_i M/\delta)), \forall i \in [N]$. Then simultaneously for all $i \in [N]$ and all $\phi \in \Phi$, with probability $1 - \delta$, it holds that*

$$\|\phi(s, a)\|_{\widehat{\Sigma}_{k_i, \phi}^{-1}} = \Theta \left(\|\phi(s, a)\|_{\Sigma_{\rho_{k_i} \times \bar{\pi}_{k_i}, \phi}^{-1}} \right).$$

The following is the canonical simulation lemma, which bounds the difference between the performance of the same policy π under two different environments and dates back at least to Abbeel and Ng [2005].

Lemma C.4 (Simulation lemma). *Given two MDP models $(P', \ell - b)$ and (P, ℓ), for any policy π, it holds that*

$$V_{P', \ell-b}^\pi - V_{P, \ell}^\pi = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_{P'}^\pi} \left[-b(s, a) + \gamma \left(P'(\cdot \mid s, a) - P(\cdot \mid s, a) \right)^\top V_{P, \ell}^\pi \right], \quad (32)$$

and

$$V_{P', \ell-b}^\pi - V_{P, \ell}^\pi = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_P^\pi} \left[-b(s, a) + \gamma \left(P'(\cdot \mid s, a) - P(\cdot \mid s, a) \right)^\top V_{P', \ell-b}^\pi \right]. \quad (33)$$