Refined Regret for Adversarial MDPs with Linear Function Approximation

Yan Dai [1] **Haipeng Luo** [2] **Chen-Yu Wei** [3] **Julian Zimmert** [4]

Abstract

We consider learning in an adversarial Markov Decision Process (MDP) where the loss functions can change arbitrarily over K episodes and the state space can be arbitrarily large. We assume that the Q-function of any policy is linear in some known features, that is, a linear function approximation exists. The best existing regret upper bound for this setting (Luo et al., 2021b) is of order $\widetilde{\mathcal{O}}(K^{2/3})$ (omitting all other dependencies), given access to a simulator. This paper provides two algorithms that improve the regret to $\widetilde{\mathcal{O}}(\sqrt{K})$ in the same setting. Our first algorithm makes use of a refined analysis of the Follow-the-Regularized-Leader (FTRL) algorithm with the log-barrier regularizer. This analysis allows the loss estimators to be arbitrarily negative and might be of independent interest. Our second algorithm develops a magnitude-reduced loss estimator, further removing the polynomial dependency on the number of actions in the first algorithm and leading to the optimal regret bound (up to logarithmic terms and dependency on the horizon). Moreover, we also extend the first algorithm to simulator-free linear MDPs, which achieves $\widetilde{\mathcal{O}}(K^{8/9})$ regret and greatly improves over the best existing bound $\widetilde{\mathcal{O}}(K^{14/15})$. This algorithm relies on a better alternative to the Matrix Geometric Resampling procedure by Neu & Olkhovskaya (2020), which could again be of independent interest.

1. Introduction

Markov Decision Processes (MDPs) have been widely used to model reinforcement learning problems, where an agent needs to make decisions sequentially and to learn from the feedback received from the environment. In this paper, we focus on adversarial MDPs where the loss functions can vary with time and the state space can also be arbitrarily large, capturing the fact that in real-world applications such as robotics, the environment can be non-stationary and the number of states can be prohibitively large.

To handle large state spaces, one of the most common methods in the literature is to assume a linear-function approximation (Yang & Wang, 2020; Jin et al., 2020b; Wei et al., 2021; Zanette et al., 2021; Neu & Olkhovskaya, 2021; Luo et al., 2021b), where the expected loss suffered by any policy from any state-action pair, commonly known as the Q-function, is linear in a set of known features.

While such assumptions are commonly used in the literature, the minimax optimal regret attainable by the agent in adversarial environments is still poorly understood. [1] Specifically, for adversarial linear-Q MDPs, the best existing bound is of order $K^{2/3}$ where K is the number of episodes (Luo et al., 2021b).[2] In that paper, the authors assume a transition simulator (i.e., the agent is allowed to draw a trajectory starting from any state-action pair, sampled from the actual transition and a given policy, without any cost) and achieve $\widetilde{\mathcal{O}}(d^{2/3}H^2K^{2/3})$ regret where H is the length of each episode and d is the dimension of the feature space. On the other hand, the best lower bound for this setting (induced from the special case of adversarial linear bandits) is of order $\Omega(\sqrt{K})$ (Dani et al., 2008). Therefore, a natural question arises:

Is it possible to design an algorithm in adversarial linear-Q MDPs that attains $\widetilde{\mathcal{O}}(\sqrt{K})$ regret bound?

In this work, we answer this question in the affirmative by developing two algorithms that both attain $\widetilde{\mathcal{O}}(\sqrt{K})$ regret in adversarial linear-Q MDPs when a simulator is granted, closing the gap with the lower bound. Both of our algorithms follow the same framework of the policy optimization algo-

[1]IIIS, Tsinghua University [2]University of Southern California [3]IDSS, MIT [4]Google Research. Correspondence to: Yan Dai <yan-dai20@mails.tsinghua.edu.cn>, Haipeng Luo <haipengl@usc.edu>, Chen-Yu Wei <chenyuw@mit.edu>, Julian Zimmert <zimmert@google.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]To be more specific, we only consider bandit feedback in this paper, where the agent can only observe her experienced losses. For the easier full-information setting, $\sqrt{K}$-style near-optimal regret has already been achieved (He et al., 2022) (*cf.* Table 1).

[2]Meanwhile, if a "good" exploratory policy π_0 (formalized in Footnote 4) is granted, $\sqrt{K}$-style bounds are also achievable, though with some additional dependencies on the quality of π_0 (Luo et al., 2021b; Neu & Olkhovskaya, 2021); see Table 1.

Table 1. Overview of Our Results and Comparisons with the Most Related Works

Algorithm	Setting[a]	Transition	Assumption[b]	Regret
DILATED BONUS (Luo et al., 2021b)	Linear-Q MDP (Definition 2.2)	Simulator (Definition 2.3)	None	$\widetilde{\mathcal{O}}(d^{2/3}H^2\boldsymbol{K^{2/3}})$
			Exploratory Policy	$\widetilde{\mathcal{O}}\left(\text{poly}(d,H)(\boldsymbol{K/\lambda_0})^{\boldsymbol{1/2}}\right)$
Algorithm 1 (**This work**)			**None**	$\widetilde{\mathcal{O}}(A^{1/2}d^{1/2}H^3\boldsymbol{K^{1/2}})$
Algorithm 2 (**This work**)				$\widetilde{\mathcal{O}}(d^{1/2}H^3\boldsymbol{K^{1/2}})$
POWERS (He et al., 2022)	Linear Mixture MDP	Unknown	Full Information	$\widetilde{\mathcal{O}}(dH\boldsymbol{K^{1/2}})$
ONLINE Q-REPS (Neu & Olkhovskaya, 2021)	Linear MDP (Definition 2.4)	Known	Exploratory Policy	$\widetilde{\mathcal{O}}\left(\text{poly}(d,H)(\boldsymbol{K/\lambda_0})^{\boldsymbol{1/2}}\right)$[c]
DILATED BONUS (Luo et al., 2021a;b)		Unknown	None	$\widetilde{\mathcal{O}}(d^2H^4\boldsymbol{K^{14/15}})$
			Exploratory Policy	$\widetilde{\mathcal{O}}\left(\text{poly}(d,H)\left(\boldsymbol{K/\lambda_0^{2/3}}\right)^{\boldsymbol{6/7}}\right)$
Algorithm 6 (**This work**)			**None**	$\widetilde{\mathcal{O}}(H^{20/9}A^{1/9}d^{2/3}\boldsymbol{K^{8/9}})$

[a]Linear MDP is a special case of linear-Q MDP, while linear mixture MDP is generally incomparable to these two.

[b]The definitions of "exploratory policy" and λ_0 are stated in Footnote 4, while "full information" means the entire loss function is revealed at the end of each episode (which is easier than our bandit-feedback setting).

[c]This is a refined version of the original $\mathcal{O}(\sqrt{K\log K})$ bound presented in their Theorem 1, which contains no explicit dependency on λ_0 by assuming $K=\Omega(\exp(\lambda_0^{-1}))$. See Review hYYK at https://openreview.net/forum?id=gviX23L1bqw.

rithm with *dilated exploration bonuses* of (Luo et al., 2021b), but with important modifications. Specifically, our first algorithm applies Follow-the-Regularized-Leader (FTRL) with the log-barrier regularizer (instead of the negative entropy regularizer used by Luo et al. (2021b)) at each state, and we develop a new analysis inspired by Zimmert & Lattimore (2022) that allows the loss estimators to be arbitrarily negative (in contrast, in the usual analyses of FTRL, the loss estimator cannot be too negative). This new analysis is the key to improving the regret to $\mathcal{O}(\sqrt{K})$ and might be of independent interest even for multi-armed bandits.

Although our first algorithm is simple in design and analysis, the log-barrier regularizer causes an $\mathcal{O}(\sqrt{A})$ factor in the regret bound $\widetilde{\mathcal{O}}(H^3\sqrt{AdK})$ where A is the number of actions. As a rescue, we develop another algorithm that uses the negative entropy regularizer with a *magnitude-reduced* loss estimator. This algorithm attains an $\widetilde{\mathcal{O}}(H^3\sqrt{dK})$ regret bound, which is optimal in d and K up to logarithmic factors and gets rid of the poly(A) dependency. We note that our algorithm is of interest even for the special case of adversarial linear bandits, as it removes the need of explicit John's exploration introduced by Bubeck et al. (2012).

At last, we also apply our method to *simulator-free* linear MDPs (formally defined in Definition 2.4), yielding an efficient algorithm with $\widetilde{\mathcal{O}}(K^{8/9})$ regret and greatly outperforming the best existing bound $\widetilde{\mathcal{O}}(K^{14/15})$ (Luo et al., 2021a).[3] Remarkably, in this application, we not only use our refined analysis for FTRL with the log-barrier regularizer, but also develop a more sample-efficient alternative to the Matrix Geometric Resampling (MGR) method introduced by Neu & Olkhovskaya (2020) and later adopted by Neu & Olkhovskaya (2021) and Luo et al. (2021a), which could also be of independent interest.

[3](Luo et al., 2021a) is a refined version of (Luo et al., 2021b).

1.1. Related Work

MDPs with Linear-Function Approximation. Linear function approximation has been a standard technique for handling large state spaces in RL, but only recently have researchers provided strong regret guarantees for these algorithms under precise conditions. Yang & Wang (2020) introduced a linear function approximation scheme called embedded linear transition MDPs where the transition kernels are bilinear, i.e., the probability of reaching state s' in the h-th step of an episode after taking action a at state s is $\mathbb{P}(s' \mid s,a)=\phi(s,a)^\top M\psi(s')$ for some known feature mappings ϕ and ψ and an unknown M. Jin et al. (2020b) loosen the assumption to linear MDPs (Definition 2.4), i.e., $\mathbb{P}(s' \mid s,a)=\phi(s,a)^\top\nu(s')$ where ν is unknown. Zhou et al. (2021) study the linear mixture MDP with $\mathbb{P}(s' \mid s,a)=\psi(s' \mid s,a)^\top\theta$ where θ is unknown. This generalizes embedded linear transition MDPs but is incomparable with linear MDPs. Another common model is linear-Q MDPs (Abbasi-Yadkori et al., 2019) where the Q-function with respect to any policy π can be written as $Q_h^\pi(s,a)=\phi^\top(s,a)\theta_h^\pi$ for some unknown θ_h^π (Definition 2.2). Linear MDPs are special cases of linear-Q MDPs.

Adversarial MDPs. MDPs with adversarial losses were first studied in the tabular cases where the state space has a small size $S \ll \infty$. Zimin & Neu (2013) first assume known transitions and achieve $\widetilde{\mathcal{O}}(H\sqrt{K})$ regret when full information is available and $\widetilde{\mathcal{O}}(\sqrt{HSAK})$ regret when only bandit feedback is available. Rosenberg & Mansour (2019) then study the unknown-transition case and get $\widetilde{\mathcal{O}}(HS\sqrt{AK})$ regret with full information. Finally, Jin et al. (2020a) tackle the hardest case with unknown transitions and bandit feedback and achieve $\widetilde{\mathcal{O}}(HS\sqrt{AK})$ regret as well.

Results for adversarial MDPs with linear-function approximations are summarized in Table 1. Specifically, Cai et al. (2020) study unknown-transition, full-information linear mixture MDPs and get $\widetilde{\mathcal{O}}(dH^{3/2}\sqrt{K})$ regret; this result is further improved to $\widetilde{\mathcal{O}}(dH\sqrt{K})$ by He et al. (2022). Neu & Olkhovskaya (2021) then study known-transition, bandit-feedback linear MDPs. Provided with a "good" exploratory policy,[4] their algorithm achieves an $\widetilde{\mathcal{O}}(\text{poly}(d,H)\sqrt{K/\lambda_0})$ regret guarantee. Later, Luo et al. (2021b) study bandit-feedback linear-Q MDPs with a simulator, providing an $\widetilde{\mathcal{O}}(d^{2/3}H^2K^{2/3})$ regret bound. A refined version (Luo et al., 2021a) considers simulator-free bandit-feedback linear MDPs, giving an $\widetilde{\mathcal{O}}(d^2H^4K^{14/15})$ bound. Meanwhile, provided with a good exploratory policy (see Footnote 4), these bounds improve to $\widetilde{\mathcal{O}}(\text{poly}(d,H)\sqrt{K/\lambda_0})$ and $\widetilde{\mathcal{O}}(\text{poly}(d,H)\lambda_0^{-4/7}K^{6/7})$, respectively. As a final remark, we point out that exploration in MDPs with huge state spaces is challenging and assuming an exploratory policy is unrealistic — as far as we know, there are no results ensuring even the *existence* of such a "good" exploratory policy, let alone finding it efficiently. We emphasize that our results *do not* require such unrealistic exploratory assumptions.

Policy Optimization Algorithms. Policy optimization algorithms for RL directly optimize the learner's policy. They are more resilient to model misspecification or even adversarial manipulation. But due to their local search nature, they suffer from the notorious *distribution mismatch* issue. Recent theoretical works address this issue by exploration bonuses; see (Agarwal et al., 2020; Shani et al., 2020; Zanette et al., 2021) for stochastic settings and (Luo et al., 2021b) for adversarial settings. While the latter work achieves near-optimal regret in tabular settings, there is a huge room for improvement when considering function approximation. Our work builds on top of their framework (especially the dilated bonus idea) and significantly improves their results in linear-function approximation settings.

Concurrent Works. Aside from this paper, there are several concurrent submissions also studying linear-Q or linear MDPs with adversarial losses, bandit feedback, and unknown transitions. Sherman et al. (2023) propose computationally efficient algorithms for linear MDPs: without simulators, their algorithm achieves $\widetilde{\mathcal{O}}(K^{6/7})$ regret and outperforms our $\widetilde{\mathcal{O}}(K^{8/9})$; however, when simulators are made available, their $\widetilde{\mathcal{O}}(K^{2/3})$ result becomes worse than our $\widetilde{\mathcal{O}}(\sqrt{K})$ bound for linear-Q MDPs (recall that linear MDPs are special linear-Q MDPs), albeit being more computationally efficient. Kong et al. (2023) propose an inefficient algorithm for linear MDPs based on recent ideas of Wagenmaker & Jamieson (2022) and get $\widetilde{\mathcal{O}}(K^{4/5})$ regret. Lancewicki et al. (2023) consider linear-Q MDPs with delayed feedback, which recovers the $\widetilde{\mathcal{O}}(K^{2/3})$ bound by Luo et al. (2021b) when there are no feedback delays.

[4]Formally, a "good" exploratory policy π_0 ensures that $\lambda_{\min}(\Sigma_h^{\pi_0}) \geq \lambda_0$ for all $h \in [H]$ and a positive constant λ_0, where $\Sigma_h^{\pi_0}$ means the covariance of π_0 in layer h (see Eq. (2)).

2. Preliminaries

Notations. For $N \in \mathbb{N}$, $[N]$ denotes the set $\{1, 2, \ldots, N\}$. For a (possibly infinite) set X, we denote the probability simplex over X by $\triangle(X)$. For a random event $\mathcal{E}$, denote its indicator by $\mathbb{1}[\mathcal{E}]$. For two square matrices A, B of the same size, $\langle A, B\rangle$ stands for $\text{Tr}(A^\top B)$. For $x \in \mathbb{R}$, define $(x)_-$ as $\min\{x, 0\}$. We use $\widetilde{\mathcal{O}}$ to hide all logarithmic factors.

No-Regret Learning in MDPs. An (episodic) adversarial MDP is specified by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, \ell)$ where $\mathcal{S}$ is the state space (possibly infinite), $\mathcal{A}$ is the action space (assumed to be finite with size $A = |\mathcal{A}|$), $\mathbb{P}\colon \mathcal{S}\times\mathcal{A} \to \triangle(\mathcal{S})$ is the transition, and $\ell\colon [K] \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the loss function chosen arbitrarily by an adversary. Following Luo et al. (2021b), the state space is assumed to be *layered*, i.e., $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_H$ where $\mathcal{S}_h \cap \mathcal{S}_{h'} = \varnothing$ for any $1 \leq h < h' \leq H$, and transition is only possible from one layer to the next one, that is, $\mathbb{P}(s' \mid s, a) \neq 0$ only when $s \in \mathcal{S}_h$ and $s' \in \mathcal{S}_{h+1}$ for some $h < H$. We also assume that there is an initial state s_1 such that $\mathcal{S}_1 = \{s_1\}$. Note that as the regret does not contain any dependency on the size of $\mathcal{S}$, this assumption is made without loss of generality.

The game lasts for K episodes, each with length H. For each episode k, the agent is initialized at state s_1. For each step $h \in [H]$, she chooses an action $a_h \in \mathcal{A}$, suffers and observes the loss $\ell_k(s_h, a_h)$, and transits to a new state s_{h+1} independently sampled from the transition $\mathbb{P}(\cdot \mid s_h, a_h)$.

A policy π of the agent is a mapping from $\mathcal{S}$ to $\triangle(\mathcal{A})$. Let Π be the set of all policies. For each episode $k \in [K]$, let $\pi_k \in \Pi$ be the policy deployed by the agent. Its expected loss is then indicated by $V_k^{\pi_k}(s_1)$, where the *state-value function* (or V-function in short) $V_k^{\pi}(s_1)$ is defined as follows for any episode k and policy π:

$$V_k^{\pi}(s_1) \triangleq \mathbb{E}\left[\sum_{h=1}^{H} \ell_k(s_h, a_h) \middle| (s_h, a_h) \sim \pi, \forall h \in [H]\right],$$

with $(s_h, a_h) \sim \pi, \forall h \in [H]$ denoting a trajectory sampled

from π. The agent aims to minimize the cumulative total loss collected in all episodes, or equivalently, the *regret*:

Definition 2.1 (Regret)**.** The regret of the agent is

$$\mathcal{R}_K \triangleq \mathbb{E}\left[\sum_{k=1}^{K} V_k^{\pi_1}(s_1)\right] - \sum_{k=1}^{K} V_k^{\pi^*}(s_1),$$

where the expectation is taken over the randomness of both the agent and the transition, and π^* is the optimal policy in hindsight (i.e., $\pi^* \in \operatorname{argmin}_{\pi\in\Pi} \sum_{k=1}^{K} V_k^{\pi}(s_1)$).

2.1. Linear-Q MDP and Linear MDP

A concept closely related to the V-function is the *action-value function* (*a.k.a.* Q-function), which denotes the expected loss suffered by a policy π starting from a given state-action pair (s, a). Formally, we define for all $(s, a) \in \mathcal{S}\times\mathcal{A}$:

$$Q_k^{\pi}(s,a) = \ell_k(s,a) + \mathbb{1}[s \notin \mathcal{S}_H] \mathop{\mathbb{E}}_{\substack{s'\sim\mathbb{P}(\cdot|s,a)\\ a'\sim\pi(\cdot|s')}} \left[Q_k^{\pi}(s',a')\right]. \quad (1)$$

We can then define linear-Q MDPs as follows.

Definition 2.2 ((Luo et al., 2021b, Assumption 1))**.** In a *linear-Q MDP*, each state-action pair (s, a) is associated with a known feature $\phi(s, a) \in \mathbb{R}^d$ with $\|\phi(s, a)\|_2 \le 1$. Moreover, for any policy $\pi \in \Pi$, episode $k \in [K]$, and layer $h \in [H]$, there exists a (hidden) vector $\theta_{k,h}^{\pi} \in \mathbb{R}^d$ such that

$$Q_k^{\pi}(s_h, a_h) = \phi(s_h, a_h)^{\mathsf{T}} \theta_{k,h}^{\pi}, \quad \forall s_h \in \mathcal{S}_h, a_h \in \mathcal{A}.$$

We assume that $\|\theta_{k,h}^{\pi}\|_2 \le \sqrt{d}H$ for all k, h, π.

Definition 2.2 does not specify any specific structure on the transition, making learning extremely difficult. Consequently, prior works all assume the availability of a simulator that allows us to sample from the hidden transition:

Definition 2.3 (Simulator)**.** A *simulator* accepts a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and generates a next-state output sampled from the true transition, i.e., $s' \sim \mathbb{P}(\cdot \mid s, a)$.

When a simulator is unavailable, we consider a special case called linear MDPs which further impose a linear structure on the transition, enabling the agent to learn:

Definition 2.4 ((Luo et al., 2021b, Assumption 3))**.** A *linear MDP* is a linear-Q MDP that additionally satisfies the following property: for any $h \in [H]$, the transition from layer h to layer $h + 1$ can be written as follows:

$$\mathbb{P}(s_{h+1} \mid s_h, a_h) = \langle \phi(s_h, a_h), \nu(s_{h+1})\rangle,$$
$$\forall s_{h+1} \in \mathcal{S}_{h+1}, s_h \in \mathcal{S}_h, a_h \in \mathcal{A}.$$

Here, the mapping $\nu\colon \mathcal{S} \to \mathbb{R}^d$ is also unrevealed to the agent. We also assume that $\|\nu(s)\|_2 \le \sqrt{d}$ for all $s \in \mathcal{S}$.

In MDPs with linear function approximation, another important quantity associated with each policy π is its *covariance matrix* at each layer $h \in [H]$, defined as follows:

$$\Sigma_h^{\pi} = \mathop{\mathbb{E}}_{(s_h,a_h)\sim\pi}[\phi(s_h, a_h)\phi(s_h, a_h)^{\mathsf{T}}], \quad \forall h \in [H]. \quad (2)$$

2.2. Dilated Bonuses for Policy Optimization

We now briefly introduce the policy optimization method and the key dilated bonus idea of Luo et al. (2021b), upon which our algorithms are built. The foundation of policy optimization is the performance difference lemma (Kakade & Langford, 2002), which asserts that the regret of the agent can be viewed as the weighted average of the regret of some local bandit problem over each state. Formally, we have

$$\mathcal{R}_K = \sum_{h=1}^{H} \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\sum_{k=1}^{K}\sum_{a} \left(\pi_k(a|s_h) - \pi^*(a|s_h)\right) Q_k^{\pi_k}(s_h, a)\right],$$

where the part inside the expectation is exactly the regret of a multi-armed bandit (MAB) problem at state s_h with "loss" $Q_k^{\pi_k}(s_h, a)$ (instead of $\ell_k(s_h, a)$) for action a. Policy optimization algorithms then naturally run a bandit algorithm at each state with an appropriate Q-function estimator to learn the best policy directly. For example, in linear-Q MDPs, since the "loss" $Q_k^{\pi_k}(s_h, a)$ is linear in some feature, it suggests running an adversarial linear bandit algorithm such as Exp2 (Bubeck et al., 2012) at each state.

However, as discussed in detail by Luo et al. (2021b), the bias/variance of the Q-function estimator often leads to a regret term of the form $\sum_{k,h} \mathbb{E}_{(s_h,a_h)\sim\pi^*}[b_k(s_h, a_h)]$ for some non-negative functions $b_1, \ldots, b_K\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\ge 0}$, where $b_k(s, a)$ is often prohibitively large if (s, a) is rarely visited by the agent. Hence, the expectation over $(s_h, a_h) \sim \pi^*$ could be potentially large as well, while the expectation over $(s_h, a_h) \sim \pi_k$ is relatively small. This is the well-known *distribution mismatch* issue: for example, when applying a linear bandit algorithm at each state, $b_k(s_h, a_h)$ is roughly $\beta\|\phi(s_h, a_h)\|^2_{(\Sigma_h^{\pi_k})^{-1}}$ for some $\beta > 0$. Thus, $\mathbb{E}_{(s_h,a_h)\sim\pi_k}[b_k(s_h, a_h)] = \beta\big\langle(\Sigma_h^{\pi_k})^{-1}, \mathbb{E}_{(s_h,a_h)\sim\pi_k}[\phi(s_h, a_h)\phi(s_h, a_h)^{\mathsf{T}}]\big\rangle = \beta d$; on the other hand, its counterpart with (s_h, a_h) drawn from π^* (i.e., $\mathbb{E}_{(s_h,a_h)\sim\pi^*}[b_k(s_h, a_h)]$) could be arbitrarily large.

To address this distribution mismatch issue and "convert" the measure from π_k to π_*, Luo et al. (2021b) consider treating these functions as exploration bonuses and further propose the so-called "*dilated bonus*" functions $B_k(s, a)$:

$$B_k(s,a) = b_k(s,a) + \mathbb{1}[s \notin \mathcal{S}_H]\big(1 + \tfrac{1}{H}\big) \mathop{\mathbb{E}}_{\substack{s'\sim\mathbb{P}(\cdot|s,a)\\ a'\sim\pi_k(\cdot|s')}} \left[B_k(s',a')\right]. \quad (3)$$

Compared to Eq. (1), B_k can be viewed the Q-function of b_k, except that it assigns slightly more weight to deeper layers

via the extra weighting $1+\frac{1}{H}$ to encourage more exploration to those layers. Their algorithms then try to minimize the regret with respect to the "optimistic" loss $Q_k^{\pi_k}-B_k$, instead of just $Q_k^{\pi_k}$, at each state. Their analysis relies on the following key lemma, which shows that if the regret w.r.t. $Q_k^{\pi_k}-B_k$ at each state is in some particular form, then the distribution mismatch issue can be resolved: that is, the final regret $\mathcal{R}_K$ is in terms of $\sum_{k,h}\mathbb{E}_{(s_h,a_h)\sim\pi_k}[b_k(s_h,a_h)]$ instead of $\sum_{k,h}\mathbb{E}_{(s_h,a_h)\sim\pi^*}[b_k(s_h,a_h)]$.

Lemma 2.5 (Lemma 3.1 by Luo et al. (2021a)). *If $\{b_k\}_{k=1}^K$ are non-negative, $B_k(s,a)$ is defined as in Eq. (3), and the following holds for any state s:*

$$\mathbb{E}\left[\sum_{k=1}^K\sum_a\left(\pi_k(a|s)-\pi^*(a|s)\right)\left(Q_k^{\pi_k}(s,a)-B_k(s,a)\right)\right]$$
$$\le X(s)+\mathbb{E}\left[\sum_{k=1}^K\sum_{a\in\mathcal{A}}\left(\pi^*(a|s)b_k(s,a)+\frac{1}{H}\pi_k(a|s)B_k(s,a)\right)\right],$$

where $X(s)$ is an arbitrary function, then

$$\mathcal{R}_K\le\sum_{h=1}^H\mathop{\mathbb{E}}_{s_h\sim\pi^*}[X(s_h)]+3\sum_{k=1}^K\sum_{h=1}^H\mathop{\mathbb{E}}_{(s_h,a_h)\sim\pi_k}[b_k(s_h,a_h)].$$

Luo et al. (2021b) show that the per-state regret bound in the condition of Lemma 2.5 indeed holds when deploying Follow-the-Regularized-Leader (FTRL) with the *negative entropy* regularizer at each state. However, for technical issues (which we will discuss and resolve in Sections 3 and 4), they only achieve $\widetilde{\mathcal{O}}(K^{2/3})$ regret in linear-Q MDPs.

3. $\widetilde{\mathcal{O}}(\sqrt{K})$ Regret for Linear-Q MDPs via Refined Log-Barrier Analysis

As mentioned, our first algorithm replaces the negative entropy regularizer in FTRL used by Luo et al. (2021b) with another regularizer called the *log-barrier*. Specifically, FTRL applied to the action space $\mathcal{A}$ maintains a sequence of distributions $x_1,x_2,\ldots,x_T\in\triangle([A])$ via the FTRL update

$$x_t=\mathop{\operatorname{argmin}}_{x\in\triangle([A])}\left\{\eta\left\langle x,\sum_{\tau<t}c_\tau\right\rangle+\Psi(x)\right\},$$

where $c_1,\ldots,c_T\in\mathbb{R}^A$ is the loss sequence, $\Psi:\triangle([A])\to\mathbb{R}$ is the regularizer, and $\eta>0$ is the learning rate. The classical MAB algorithm Exp3 (Auer et al., 2002) and its linear-bandit variant Exp2 (Bubeck et al., 2012) both use the negative entropy regularizer $\Psi(x)=\sum_{i=1}^A p_i\ln p_i$.

Starting from (Foster et al., 2016), a sequence of works discover many nice properties of a different regularizer called log-barrier, defined as $\Psi(p)=\sum_{i=1}^A\ln\frac{1}{p_i}$. Here, we present yet another new and useful property of log-barrier, summarized in the following lemma.[5] Our proof is inspired by the analysis of the log-determinant regularizer by Zimmert & Lattimore (2022) and is deferred to Appendix B.1.

Lemma 3.1. *Let $x_1,\ldots,x_T\in\triangle([A])$ be defined as*

$$x_t=\mathop{\operatorname{argmin}}_{x\in\triangle([A])}\left\{\eta\left\langle x,\sum_{\tau<t}c_\tau\right\rangle+\Psi(x)\right\},\quad\forall t=1,\ldots,T,$$

where $c_t\in\mathbb{R}^A$ is an arbitrary loss vector corresponding to the t-th iteration and $\Psi(p)=\sum_{i=1}^A\ln\frac{1}{p_i}$ is the log-barrier regularizer. Then the regret against any distribution $y\in\triangle([A])$ with respect to $\{c_t\}_{t=1}^T$ is bounded as:

$$\sum_{t=1}^T\langle x_t-y,c_t\rangle\le\frac{\Psi(y)-\Psi(x_1)}{\eta}+\eta\sum_{t=1}^T\sum_{i=1}^A x_{t,i}c_{t,i}^2.$$

Readers familiar with the Exp2/Exp3 analysis would immediately recognize this regret bound since it is also the same bound that FTRL with negative entropy (also known as Hedge) enjoys. However, the key distinction is that for log-barrier, this holds without *any* requirement on the magnitude of $c_{t,i}$, while for negative entropy, one must require $\eta c_{t,i}\ge-1$ for all $t\in[T]$ and $i\in[A]$ (i.e., losses cannot be too negative; see Lemma C.1 in the appendix for more details). This turns out to be critical for improving the regret when applying it to linear-Q MDPs, as discussed later.

Remark 3.2. Our Lemma 3.1 also answers the open question raised by Zheng et al. (2019) (see their remark after Lemma 14): it is indeed possible for FTRL with log-barrier to attain a $\sum_i x_{t,i}c_{t,i}^2$-style bound without any restrictions on the losses. As log-barrier regularizers are widely used in the literature for its better data-adaptivity (Wei & Luo, 2018; Ito, 2021), we expect this result to be of independent interest.

Linear-Q Algorithm. Our final algorithm for linear-Q MDPs is shown in Algorithm 1, which is nearly the same as (Luo et al., 2021b, Algorithm 2) except for the part marked in blue where we use the log-barrier regularizer, as mentioned. Specifically, based on the discussions in Section 2.2, the loss fed to FTRL is $\widehat{Q}_k-B_k$. Here, $\widehat{Q}_k$ (defined in Eq. (5)) is a standard estimator for $Q_k^{\pi_k}$ involving an estimate $\widehat{\Sigma}_{k,h}^\dagger$ for the inverse of $\Sigma_h^{\pi_k}$, which is constructed via the Matrix Geometric Resampling procedure (Neu & Olkhovskaya, 2020) with the help of the simulator. Meanwhile, B_k is the dilated bonus following the idea of Eq. (3) with $b_k(s,a)$ defined as follows for all $s\in\mathcal{S}_h$ and $a\in\mathcal{A}$:

$$b_k(s,a)=\beta\left(\|\phi(s,a)\|_{\widehat{\Sigma}_{k,h}^\dagger}^2+\mathop{\mathbb{E}}_{\widetilde{a}\sim\pi_k(\cdot|s)}\left[\|\phi(s,\widetilde{a})\|_{\widehat{\Sigma}_{k,h}^\dagger}^2\right]\right).$$

[5] When revising this manuscript, we found that a key property we use when proving this lemma (see our Eq. (11)) was also independently developed by Putta & Agrawal (2022, Corollary 7) under the name "new local-norm lower-bounds for Bregman divergences". They used this property to handle scale-free adversarial Multi-Armed Bandits (MABs) where the losses are not always constantly bounded but can be arbitrarily positive or negative.

Algorithm 1 Improved Linear-Q Algorithm (Using Log-Barrier Regularizers)

Input: Learning rate η, bonus parameter β, MGR parameters γ and ϵ, FTRL regularizer $\Psi(p) = \sum_{i=1}^{|\mathcal{A}|} \ln \frac{1}{p_i}$.

1: **for** $k = 1, 2, \ldots, K$ **do**

2: Let $\pi_k \in \Pi$ be defined as follows for all $s \in \mathcal{S}$:

$$\pi_k(s) = \operatorname*{argmin}_{p \in \triangle(\mathcal{A})} \Big\{ \Psi(p) + \eta \sum_{k'<k} \langle p(\cdot), \widehat{Q}_{k'}(s,\cdot) - B_{k'}(s,\cdot) \rangle \Big\}, \tag{4}$$

where $B_k(s,a)$ is calculated in Algorithm 4.

3: Execute π_k, observing the trajectory $(s_{k,h}, a_{k,h})$ and losses $\ell_k(s_{k,h}, a_{k,h})$ for all $h \in [H]$.

4: Construct covariance matrix inverse estimate $\widehat{\Sigma}^{\dagger}_{k,h}$ using Matrix Geometric Resampling (Algorithm 3 in the appendix) s.t. $\|\widehat{\Sigma}^{\dagger}_{k,h}\|_2 \le \frac{1}{\gamma}$,

$$\left\| \mathbb{E}[\widehat{\Sigma}^{\dagger}_{k,h}] - (\gamma I + \Sigma^{\pi_k}_h)^{-1} \right\|_2 \le \epsilon,$$

and the following holds w.p. $1 - K^{-3}$:

$$\left\| (\widehat{\Sigma}^{\dagger}_{k,h})^{1/2} \Sigma^{\pi_k}_h (\widehat{\Sigma}^{\dagger}_{k,h})^{1/2} \right\|_2 \le 2.$$

5: Let $L_{k,h} = \sum_{h' \ge h} \ell_k(s_{k,h'}, a_{k,h'})$. Estimate the Q-function $Q^{\pi_k}_k(s,a)$ as follows (for $s \in \mathcal{S}_h$):

$$\widehat{Q}_k(s,a) = \phi(s,a)^{\mathsf{T}} \widehat{\Sigma}^{\dagger}_{k,h} \phi(s_{k,h}, a_{k,h}) L_{k,h}. \tag{5}$$

6: **end for**

The calculation of B_k again requires the simulator – see Algorithm 4 for the calculation procedure and (Luo et al., 2021b) for more detailed discussions.

With the help of the property of the log-barrier regularizer stated in Lemma 3.1, we are able to show that our algorithm achieves $\widetilde{\mathcal{O}}(\sqrt{K})$ regret, formally stated as follows.

Theorem 3.3. *Algorithm 1 when applied to a linear-Q MDP (Definition 2.2) with a simulator ensures the following when $12\eta\beta H^2 \le \gamma$, $8\eta H^2 \le \beta$, and $\epsilon \le (H^2K)^{-1}$:*

$$\mathcal{R}_K = \widetilde{\mathcal{O}}\left(\beta dHK + \frac{\gamma}{\beta} dH^3 K + \frac{AH}{\eta} + \frac{\beta}{\gamma} AH^2 + A\sqrt{d}H^2 + \frac{\eta H^3}{\gamma^2 K^2} \right).$$

Picking $\eta = \sqrt{\frac{A}{dH^4K}}$, $\beta = 8\sqrt{\frac{A}{dK}}$, $\gamma = \frac{96A}{dK}$, and $\epsilon = \frac{1}{H^2K}$, we conclude that $\mathcal{R}_K = \widetilde{\mathcal{O}}(H^3\sqrt{AdK})$.

We defer the full proof to Appendix B and provide a sketch below highlighting why removing any constraints on the losses fed to FTRL is critical to improving the regret.

Proof Sketch. To bound $\mathcal{R}_K$ by the dilated bonus lemma (Lemma 2.5), we focus on a single (k,s)-pair and bound

$$\sum_a (\pi_k(a \mid s) - \pi^*(a \mid s))(Q^{\pi_k}_k(s,a) - B_k(s,a)). \tag{6}$$

As the FTRL lemma only applies to the losses fed into Eq. (4) (namely $\widehat{Q}_k(s,a) - B_k(s,a)$), we add and substract $\widehat{Q}_k(s,a)$ in Eq. (6). Hence, after summing over $k \in [K]$ and $s \sim \pi^*$, we need to consider the following three terms to figure out the term $X(s)$ in Lemma 2.5:

- BIAS-1 $= \sum_{k,h} \mathbb{E}_{s_h \sim \pi^*} \Big[\mathbb{E}_{a_h \sim \pi_k} \Big[Q^{\pi_k}_k(s_h,a_h) - \widehat{Q}_k(s_h,a_h) \Big] \Big]$, measuring the under-estimation of $\widehat{Q}_k$ w.r.t. π_k.
- BIAS-2 $= \sum_{k,h} \mathbb{E}_{s_h \sim \pi^*} \Big[\mathbb{E}_{a_h \sim \pi^*} \Big[\widehat{Q}_k(s_h,a_h) - Q^{\pi_k}_k(s_h,a_h) \Big] \Big]$, measuring the over-estimation of $\widehat{Q}_k$ w.r.t. π^*.
- REG-TERM $= \sum_{k,h} \mathbb{E}_{s_h \sim \pi^*} \Big[\sum_a (\pi_k(a|s) - \pi^*(a|s))(\widehat{Q}_k(s,a) - B_k(s,a)) \Big]$, which can be tackled by Lemma 3.1.

As BIAS-1 and BIAS-2 are independent of the regularizer, they are handled similarly to the original analysis and both contribute $\widetilde{\mathcal{O}}(\frac{\gamma}{\beta} dH^3K)$ to $\sum_h \mathbb{E}_{s_h \sim \pi^*}[X(s_h)]$ (see Lemma B.1 in the appendix). To handle REG-TERM, we apply Lemma 3.1. As in standard log-barrier analyses, since $\Psi(\pi^*(\cdot \mid s)) - \Psi(\pi_1(\cdot \mid s))$ is potentially infinity, we introduce a smooth version $\widetilde{\pi}^*$ defined via

$$\widetilde{\pi}^*(a \mid s) = (1 - AK^{-1})\pi^*(a \mid s) + K^{-1}.$$

Applying Lemma 3.1 to a given state $s \in \mathcal{S}$, we derive

$$\Psi(\widetilde{\pi}^*(\cdot \mid s)) - \Psi(\pi_1(\cdot \mid s)) \le A \log K.$$

Hence, fixing the state s, we can write

$$\begin{aligned} &\sum_{k=1}^{K} \sum_{a \in \mathcal{A}} (\pi_k(a \mid s) - \pi^*(a \mid s))(\widehat{Q}_k(s,a) - B_k(s,a)) \\ \le &\sum_{k=1}^{K} \sum_{a \in \mathcal{A}} (\widetilde{\pi}^*(a \mid s) - \pi^*(a \mid s))(\widehat{Q}_k(s,a) - B_k(s,a)) + \\ &\frac{A \log K}{\eta} + \eta \sum_{k=1}^{K} \sum_{a \in \mathcal{A}} \pi_k(a \mid s)^2 (\widehat{Q}_k(s,a) - B_k(s,a))^2. \end{aligned}$$

After summing over h and taking expectation over s, we show that the first term is of order $\widetilde{\mathcal{O}}(AH^2(\sqrt{d} + \beta/\gamma))$, while the last term's contribution to $\sum_h \mathbb{E}_{s_h \sim \pi^*}[X(s_h)]$ is of order $\widetilde{\mathcal{O}}(\frac{\eta H^3}{\gamma^2 K^2})$ by the same analysis of (Luo et al., 2021b). Finally, we combine everything, apply Lemma 2.5,

and show that the term $\sum_{k,h} \mathbb{E}_{(s_h,a_h)\sim\pi_k}[b_k(s_h,a_h)]$ in this case is $\widetilde{\mathcal{O}}(\beta dHK)$, finishing the proof for the first bound.

While this bound looks almost identical to that of (Luo et al., 2021b), their analysis requires $2H\eta \le \gamma$ because of the use of the negative-entropy regularizer. This prevents them from picking a very small γ and eventually leads to sub-optimal $\widetilde{\mathcal{O}}(K^{2/3})$ regret. On the other hand, our log-barrier analysis allows us to drop this requirement and pick a γ as small as K^{-1}, though bearing some extra factors polynomial in A (the number of actions). Indeed, optimizing the bound over the parameters (see the second claim), we achieve $\mathcal{R}_K = \widetilde{\mathcal{O}}(H^3\sqrt{AdK})$ which is of order $\sqrt{K}$. □

4. Removing Polynomial Dependency on A via a New Magnitude-Reduced Estimator

One drawback of using log-barrier is that the term $\Psi(y) - \Psi(x_1)$ from Lemma 3.1 leads to poly(A) dependency (while for negative entropy, this is only $\log A$). To get around this issue, we go back to using negative entropy. Recall that the issue of this regularizer is that to ensure the same bound as Lemma 3.1, we require $\eta c_{t,i} \ge -1$, which translates to the requirement $\eta(\widehat{Q}_k(s,a) - B_k(s,a)) \ge -1$ in the context of linear-Q MDPs. The challenging part here is that $\widehat{Q}_k(s,a)$, as defined in Eq. (5), could be as negative as $-H/\gamma$, which then restricts η to be of order $\mathcal{O}(\gamma/H)$, as mentioned.

To resolve this, we propose a new Q-function estimator that has a smaller magnitude in the negative direction. Our high-level idea is the following: for a possibly negative random variable Z, define another *magnitude-reduced* random variable as (recall the notation that $(Z)_- = \min\{Z, 0\}$):

$$\widehat{Z} = Z - (Z)_- + \mathbb{E}[(Z)_-].$$

The magnitude-reduced $\widehat{Z}$ then has the same expectation as Z. They also share the same order of second moments:

$$\mathbb{E}[\widehat{Z}^2] \le 2\,\mathbb{E}[Z^2] + 2(\mathbb{E}[(Z)_-])^2 = \mathcal{O}(\mathbb{E}[Z^2]).$$

More importantly, we have $\widehat{Z} \ge \mathbb{E}[(Z)_-]$ and thus the smallest possible value of $\widehat{Z}$ is $\mathbb{E}[(Z)_-]$: no less than (and often much larger than) the smallest possible value of Z. Thus, it becomes much easier to ensure $\eta c_{t,i} \ge -1$.

Applying this idea to our context, we propose a new Q-function estimator as in Eq. (9), where $m_k(s,a)$ exactly takes the role of $\mathbb{E}[(Z)_-]$, except that we have no access to the real expectation but have to approximate it with samples; see Line 7 of Algorithm 2 for more details.

To make sure that these samples are "consistent" with those used in constructing $\widehat{\Sigma}^{\dagger}_{k,h}$, we perform a check in Line 6 and repeat the sampling until the check passes. Since both Eq. (8) and $\|\widetilde{\Sigma}_{k,h} - \Sigma_{k,h}\|_2 \le \gamma$ hold with probability at least $1 - K^{-3}$, the expected number of trials is only $1 + o(1)$. Our algorithm is shown in Algorithm 2, where the parts different from (Luo et al., 2021b) are again highlighted in blue.

Algorithm 2 Improved Linear-Q Algorithm (Using Magnitude-Reduced Loss Estimators)

Input: Learning rate η, bonus parameter β, MGR parameters γ and ϵ, covariance estimation parameter M.

1: **for** $k = 1, 2, \ldots, K$ **do**

2: Let $\pi_k \in \Pi$ be defined as follows for all $s \in \mathcal{S}$:

$$\pi_k(a \mid s) \propto \exp\Big(-\eta \sum_{k'<k} (\widehat{Q}_{k'}(s,a) - B_{k'}(s,a))\Big),$$

where $B_k(s,a)$ is calculated in Algorithm 4.

3: Execute π_k, observing the trajectory $(s_{k,h}, a_{k,h})$ and losses $\ell_k(s_{k,h}, a_{k,h})$ for all $h \in [H]$.

4: Construct $\widehat{\Sigma}^{\dagger}_{k,h}$ using Matrix Geometric Resampling (Algorithm 3 in the appendix) s.t. $\|\widehat{\Sigma}^{\dagger}_{k,h}\|_2 \le \frac{1}{\gamma}$,

$$\Big\|\mathbb{E}[\widehat{\Sigma}^{\dagger}_{k,h}] - (\gamma I + \Sigma^{\pi_k}_h)^{-1}\Big\|_2 \le \epsilon, \tag{7}$$

and the following holds w.p. $1 - K^{-3}$:

$$\Big\|(\widehat{\Sigma}^{\dagger}_{k,h})^{1/2}\Sigma^{\pi_k}_h(\widehat{\Sigma}^{\dagger}_{k,h})^{1/2}\Big\|_2 \le 2. \tag{8}$$

5: Simulate π_k for M times, giving $\{(s_{m,h}, a_{m,h})\}_{m,h}$. Estimate the covariance matrix $\Sigma^{\pi_k}_h$ as

$$\widetilde{\Sigma}_{k,h} = \frac{1}{M}\sum_{m=1}^{M} \phi(s_{m,h}, a_{m,h})\phi(s_{m,h}, a_{m,h})^{\mathsf{T}}.$$

6: If $\|(\widehat{\Sigma}^{\dagger}_{k,h})^{1/2}\widetilde{\Sigma}_{k,h}(\widehat{\Sigma}^{\dagger}_{k,h})^{1/2}\|_2 \ge 3$, **goto** Line 4.

7: For each $s \in \mathcal{S}_h$ and $a \in \mathcal{A}$, define $m_k(s,a)$ as

$$m_k(s,a) = \frac{1}{M}\sum_{m=1}^{M}\Big(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{m,h}, a_{m,h})\Big)_-.$$

8: Let $L_{k,h} = \sum_{h' \ge h} \ell_k(s_{k,h'}, a_{k,h'})$. Estimate the Q-function $Q^{\pi_k}_k(s,a)$ as follows (for $s \in \mathcal{S}_h$):

$$\begin{aligned}\widehat{Q}_k(s,a) = {} & \phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{k,h}, a_{k,h})L_{k,h} \\ & - H\Big(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{k,h}, a_{k,h})\Big)_- \\ & + Hm_k(s,a).\end{aligned} \tag{9}$$

9: **end for**

Our analysis shows that the new Q-function estimator indeed has a significantly smaller magnitude: it can only be as negative as $-H/\sqrt{\gamma}$ (as opposed to the previous $-H/\gamma$

bound). This only restricts η to be of order $\mathcal{O}(\sqrt{\gamma}/H)$, enabling us to pick $\gamma \approx 1/K$ as in Theorem 3.3 and achieve $\widetilde{\mathcal{O}}(\sqrt{K})$ regret again, as formalized in the next theorem. Note that our final regret bound not only has no poly(A) dependency, but is also optimal in both d and K as one cannot do better even in the special case of linear bandits.

Theorem 4.1. *When applied to a linear-Q MDP (Definition 2.2) with a simulator, Algorithm 2 ensures the following when $\eta\beta\gamma^{-1} \le \frac{1}{12H^2}$, $\eta^2\gamma^{-1} \le \frac{1}{12H^2}$, $8\eta H^2 \le \beta$, $\epsilon \le (H^2K)^{-1}$, and $M = 32\gamma^{-2}\log K$:*

$$\mathcal{R}_K = \widetilde{\mathcal{O}}\left(\beta dHK + \frac{\gamma}{\beta}dH^3K + \frac{H}{\eta} + \eta dH^3K + \frac{\eta}{\gamma^2K^2}H^3\right).$$

Picking $\eta = \frac{1}{\sqrt{dH^4K}}$, $\beta = \frac{8}{\sqrt{dK}}$, $\gamma = \frac{96}{dK}$, and $\epsilon = \frac{1}{H^2K}$, we conclude that $\mathcal{R}_K = \widetilde{\mathcal{O}}(H^3\sqrt{dK})$.

Proof Sketch. Due to space limitations, we only sketch why $\widehat{Q}_k(s,a)$ is now at least $-\mathcal{O}(H/\sqrt{\gamma})$: since $L_{k,h} \in [0,H]$, we have $\widehat{Q}_k(s,a) \ge Hm_k(s,a)$. By Jensen's inequality, we can bound $m_k(s,a)^2$ for all (s,a) as:

$$\begin{aligned} m_k(s,a)^2 &\le \frac{1}{M}\sum_{m=1}^{M}\left(\phi(s,a)^\top\widehat{\Sigma}_{k,h}^\dagger\phi(s_{m,h},a_{m,h})\right)^2 \\ &= \phi(s,a)^\top\widehat{\Sigma}_{k,h}^\dagger\widetilde{\Sigma}_h^{\pi_k}\widehat{\Sigma}_{k,h}^\dagger\phi(s,a) \\ &\le 3\left\|\phi(s,a)\right\|_{\widehat{\Sigma}_{k,h}^\dagger}^2 \le \frac{3}{\gamma}, \end{aligned}$$

where the second inequality is by Line 6. Therefore, we have $\widehat{Q}_k(s,a) \ge -\frac{\sqrt{3}H}{\sqrt{\gamma}}$, a reduced magnitude compared to that of a standard estimator (which is defined in Eq. (5)). □

5. Generalizing to Linear MDPs

When a simulator is unavailable, we consider a special case of linear-Q MDPs: linear MDPs (Definition 2.4), where the transition is also linear in the features. As Luo et al. (2021a) show, this makes the dilated bonuses linear as well, and thus we can efficiently estimate them without using a simulator. Due to various technical obstacles, they only achieve $\widetilde{\mathcal{O}}(K^{14/15})$ regret. We improve it to $\widetilde{\mathcal{O}}(K^{8/9})$ via two key modifications to their algorithm. As the algorithm is fairly lengthy due to the lack of simulators and the estimation of the dilated bonus function, we defer it (Algorithm 6) to the appendix. We refer the reader to (Luo et al., 2021a) for a detail description of their algorithm, and only focus on our two modifications (highlighted in blue) described below.

The first obvious modification is to apply one of the techniques introduced in the last two sections. However, since the magnitude-reduced estimator we developed in Algorithm 2 requires extra samples (see Line 5), which, without a simulator, can only be done via real episodes and in turn introduces more regret, we go with the log-barrier approach instead, even though it leads to extra poly(A) factors.

But it turns out that this only leads to a mild improvement in the regret, since the constraint on η is not the bottleneck of their analysis. Instead, one important bottleneck comes from using the Matrix Geometric Resampling (MGR) procedure (Neu & Olkhovskaya, 2020) to estimate the covariance matrix inverse, which again, without a simulator, can only be done via running the same policy for multiple real episodes (called an *epoch*) to collect samples. It is thus critical to make this step as sample efficient as possible.

To resolve this issue, our key observation is that MGR uses too many samples to produce an estimate that is in a sense more accurate than required: specifically, it needs $\mathcal{O}(\epsilon^{-2}\gamma^{-3})$ samples to ensure a bound like Eq. (7). Instead, we find that a weaker multiplicative approximation guarantee is enough, which only requires $\mathcal{O}(\gamma^{-2})$ samples. Moreover, this is achieved by simply taking the (regularized) inverse of the empirical average of $\mathcal{O}(\gamma^{-2})$ samples.

More concretely, for a policy $\widetilde{\pi}_j$ (which mixes π_j with some exploration policy; see Algorithm 6 for more details) in epoch j of the algorithm, we collect roughly $\mathcal{O}(\gamma^{-2})$ samples using this policy and construct an empirical average covariance matrix $\widetilde{\Sigma}_{j,h}$ for layer h. Then we let $\widehat{\Sigma}_{j,h}^\dagger = (\gamma I + \widetilde{\Sigma}_{j,h})^{-1}$ be the estimation for $(\gamma I + \Sigma_h^{\widetilde{\pi}_j})^{-1}$ (see Line 17 of Algorithm 6). We then prove the following:

Lemma 5.1. *If the epoch length W in Algorithm 6 is set to $(4d\log\frac{d}{\delta})\gamma^{-2}$, then*

$$(1-\sqrt{\gamma})(\gamma I+\Sigma_h^{\widetilde{\pi}_j}) \preceq (\gamma I+\widetilde{\Sigma}_{j,h}) \preceq (1+\sqrt{\gamma})(\gamma I+\Sigma_h^{\widetilde{\pi}_j})$$

holds with probability at least $1-2\delta$.

The proof relies on a new matrix concentration bound (Lemma A.4 in the appendix), which can be of independent interest. A direct corollary of this lemma is:

Corollary 5.2. *If $W = (4d\log\frac{d}{\delta})\gamma^{-2}$ and $\gamma \le \frac{1}{4}$, then with probability $1-2\delta$, we have the following:*

$$(1-2\sqrt{\gamma})I \preceq (\widehat{\Sigma}_{j,h}^\dagger)^{1/2}(\gamma I+\Sigma_h^{\widetilde{\pi}_j})(\widehat{\Sigma}_{j,h}^\dagger)^{1/2} \preceq (1+2\sqrt{\gamma})I.$$

Proof. Simply left and right multiply $(\widehat{\Sigma}_{j,h}^\dagger)^{1/2} = (\gamma I + \widetilde{\Sigma}_{j,h})^{-1/2}$ in the bound of Lemma 5.1 and use the fact $[\frac{1}{1+\sqrt{\gamma}}, \frac{1}{1-\sqrt{\gamma}}] \subseteq [1-2\sqrt{\gamma}, 1+2\sqrt{\gamma}]$ when $\gamma \le \frac{1}{4}$. □

Together with a new analysis to bound bias terms in the regret, such approximations turn out to be enough: the bias terms enjoy almost the same bound (up to constants) as we

had in previous sections. Hence, we finally get the following theorem; see Appendix D for a full proof.

Theorem 5.3 (Informal version of Theorem D.1)**.** *When applied to linear MDPs, Algorithm 6 with proper tuning ensures $\mathcal{R}_K = \widetilde{\mathcal{O}}(K^{8/9})$ (omitting all other dependencies).*

Proof Sketch. Ignoring less important parts, at a high level we still decompose the regret into several bias terms and a REG-TERM. In this sketch, we only provide key ideas in bounding the bias terms using Corollary 5.2. Specifically, we illustrate how to bound $\mathbb{E}[\text{BIAS-1}]$, defined as follows:

$$\text{BIAS-1} \triangleq W \sum_{j,h} \mathop{\mathbb{E}}_{s_h \sim \pi^*} \left[\mathop{\mathbb{E}}_{a_h \sim \pi_j} \left[\overline{Q}_j^{\pi_j}(s_h, a_h) - \widehat{Q}_j(s_h, a_h) \right] \right],$$

where $\overline{Q}_j^{\pi_j}(s,a) = \phi(s,a)^\mathsf{T} \overline{\theta}_{j,h}^{\pi_j}$ is the average Q-functions in epoch j induced by policy π_j, and $\widehat{Q}_j(s,a)$ is its estimator. By direct calculation, one may check that

$$\begin{aligned}
&\mathbb{E}\left[\overline{Q}_j^{\pi_j}(s,a) - \widehat{Q}_j(s,a)\right] \\
&= \phi(s,a)^\mathsf{T}(I - \widehat{\Sigma}_{j,h}^\dagger \Sigma_h^{\widetilde{\pi}_j})\overline{\theta}_{j,h}^{\pi_j}, \\
&= \phi(s,a)^\mathsf{T}\widehat{\Sigma}_{j,h}^\dagger(\gamma I + \widetilde{\Sigma}_{j,h} - \Sigma_h^{\widetilde{\pi}_j})\overline{\theta}_{j,h}^{\pi_j} \\
&\le \|\phi(s,a)\|_{\widehat{\Sigma}_{j,h}^\dagger} \times \left\|(\gamma I + \widetilde{\Sigma}_{j,h} - \Sigma_h^{\widetilde{\pi}_j})\overline{\theta}_{j,h}^{\pi_j}\right\|_{\widehat{\Sigma}_{j,h}^\dagger} \\
&\le \|\phi(s,a)\|_{\widehat{\Sigma}_{j,h}^\dagger} \times \Bigg(\left\|\gamma \overline{\theta}_{j,h}^{\pi_j}(s,a)\right\|_{\widehat{\Sigma}_{j,h}^\dagger} + \\
&\qquad\qquad \left\|(\widetilde{\Sigma}_{j,h} - \Sigma_h^{\widetilde{\pi}_j})\overline{\theta}_{j,h}^{\pi_j}\right\|_{\widehat{\Sigma}_{j,h}^\dagger} \Bigg) \\
&\le \frac{\beta}{4}\|\phi(s,a)\|^2_{\widehat{\Sigma}_{j,h}^\dagger} + \frac{2}{\beta}\left\|\gamma \overline{\theta}_{j,h}^{\pi_j}(s,a)\right\|^2_{\widehat{\Sigma}_{j,h}^\dagger} + \\
&\quad \frac{2}{\beta}\left\|(\widetilde{\Sigma}_{j,h} - \Sigma_h^{\widetilde{\pi}_j})\overline{\theta}_{j,h}^{\pi_j}\right\|^2_{\widehat{\Sigma}_{j,h}^\dagger},
\end{aligned}$$

where we apply Cauchy-Schwarz inequality, triangle inequality, and AM-GM inequality (twice).

The first term contributes to $\mathcal{O}(\beta dHK)$ after summing over k and h and applying Lemma 2.5, while the second term is bounded by $\mathcal{O}(\frac{\gamma}{\beta} dH^2)$ for any j because of the fact that $\|\widehat{\Sigma}_{j,h}^\dagger\|_2 \le \gamma^{-1}$. By definition of $\overline{\theta}$, the last term becomes

$$\frac{2}{\beta}\left\|(\widetilde{\Sigma}_{j,h} - \Sigma_h^{\widetilde{\pi}_j})\widehat{\Sigma}_{j,h}^\dagger(\widetilde{\Sigma}_{j,h} - \Sigma_h^{\widetilde{\pi}_j})\right\|_2 dH^2.$$

By some algebraic manipulations, we write

$$\begin{aligned}
&(\widetilde{\Sigma}_{j,h} - \Sigma_h^{\widetilde{\pi}_j})\widehat{\Sigma}_{j,h}^\dagger(\widetilde{\Sigma}_{j,h} - \Sigma_h^{\widetilde{\pi}_j}) \\
&= (\widehat{\Sigma}_{j,h}^\dagger)^{-1/2}\big(I - (\widehat{\Sigma}_{j,h}^\dagger)^{1/2}(\gamma I + \Sigma_h^{\widetilde{\pi}_j})(\widehat{\Sigma}_{j,h}^\dagger)^{1/2}\big)^2(\widehat{\Sigma}_{j,h}^\dagger)^{-1/2}.
\end{aligned}$$

Thanks to Corollary 5.2, we know that

$$-2\sqrt{\gamma}I \preceq I - (\widehat{\Sigma}_{j,h}^\dagger)^{1/2}(\gamma I + \Sigma_h^{\widetilde{\pi}_j})(\widehat{\Sigma}_{j,h}^\dagger)^{1/2} \preceq 2\sqrt{\gamma}I.$$

Hence, the last term is also of order $\mathcal{O}(\frac{\gamma}{\beta} dH^2)$ – same as the second term. All other bias terms are bounded analogously.

It only remains to bound the REG-TERM, whose calculation is the same as we did in the proof of Theorem 3.3. Property tuning all the parameters then gives $\widetilde{\mathcal{O}}(K^{8/9})$ regret. □

6. Conclusion

In this paper, we study policy optimization algorithms in adversarial MDPs with linear function approximation.

Building on top of the dilated bonus approach introduced by Luo et al. (2021b), we derive two algorithms which both achieve *the first* $\widetilde{\mathcal{O}}(\sqrt{K})$-style regret bounds in linear-Q adversarial MDPs when a simulator is accessible. Technically speaking, the first algorithm uses a refined analysis for FTRL with the log-barrier regularizer, while the second one relies on a new magnitude-reduced loss estimator.

We further generalize the first approach to simulator-free linear MDPs and get $\widetilde{\mathcal{O}}(K^{8/9})$ regret, greatly improving over the best-known $\widetilde{\mathcal{O}}(T^{14/15})$ bound (Luo et al., 2021a). This generalization also contains an alternative to the Matrix Geometric Resampling procedure (Neu & Olkhovskaya, 2020) using a new matrix concentration bound (Lemma A.4).

We expect all these techniques to be of independent interest and potentially useful for other problems. In light of various concurrent works on linear MDPs (Sherman et al., 2023; Kong et al., 2023; Lancewicki et al., 2023), further improving the result for linear MDPs is a key future direction — either in terms of K or A.

Acknowledgements

We thank the anonymous reviewers for their insightful comments. HL is supported by NSF Award IIS-1943607 and a Google Research Scholar Award.

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pp. 3692–3702. PMLR, 2019.

Agarwal, A., Henaff, M., Kakade, S., and Sun, W. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 33:13399–13412, 2020.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pp. 41–1. JMLR Workshop and Conference Proceedings, 2012.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Dani, V., Kakade, S. M., and Hayes, T. The price of bandit information for online optimization. *Advances in Neural Information Processing Systems*, 20, 2008.

Foster, D. J., Li, Z., Lykouris, T., Sridharan, K., and Tardos, E. Learning in games: Robustness of fast convergence. *Advances in Neural Information Processing Systems*, 29, 2016.

He, J., Zhou, D., and Gu, Q. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 4259–4280. PMLR, 2022.

Ito, S. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory*, pp. 2552–2583. PMLR, 2021.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pp. 4860–4869. PMLR, 2020a.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020b.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Kong, F., Zhang, X., Wang, B., and Li, S. Improved regret bounds for linear adversarial mdps via linear optimization. *arXiv preprint arXiv:2302.06834*, 2023.

Lancewicki, T., Rosenberg, A., and Sotnikov, D. Delay-adapted policy optimization and improved regret for adversarial mdp with delayed bandit feedback. *arXiv preprint arXiv:2305.07911*, 2023.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Luo, H., Wei, C.-Y., and Lee, C.-W. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *arXiv preprint arXiv:2107.08346*, 2021a.

Luo, H., Wei, C.-Y., and Lee, C.-W. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34:22931–22942, 2021b.

Meng, L. and Zheng, B. The optimal perturbation bounds of the moore–penrose inverse under the frobenius norm. *Linear algebra and its applications*, 432(4):956–963, 2010.

Neu, G. and Olkhovskaya, J. Efficient and robust algorithms for adversarial linear contextual bandits. In *Conference on Learning Theory*, pp. 3049–3068. PMLR, 2020.

Neu, G. and Olkhovskaya, J. Online learning in mdps with linear function approximation and bandit feedback. *Advances in Neural Information Processing Systems*, 34: 10407–10417, 2021.

Putta, S. R. and Agrawal, S. Scale-free adversarial multi armed bandits. In *International Conference on Algorithmic Learning Theory*, pp. 910–930. PMLR, 2022.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pp. 5478–5486. PMLR, 2019.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613. PMLR, 2020.

Sherman, U., Koren, T., and Mansour, Y. Improved regret for efficient online reinforcement learning with linear function approximation. *arXiv preprint arXiv:2301.13087*, 2023.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Wagenmaker, A. and Jamieson, K. G. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *Advances in Neural Information Processing Systems*, 35:5968–5981, 2022.

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

Wei, C.-Y. and Luo, H. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pp. 1263–1291. PMLR, 2018.

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

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

Zanette, A., Cheng, C.-A., and Agarwal, A. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pp. 4473–4525. PMLR, 2021.

Zheng, K., Luo, H., Diakonikolas, I., and Wang, L. Equipping experts/bandits with long-term memory. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021.

Zimin, A. and Neu, G. Online learning in episodic markovian decision processes by relative entropy policy search. *Advances in neural information processing systems*, 26, 2013.

Zimmert, J. and Lattimore, T. Return of the bias: Almost minimax optimal high probability bounds for adversarial linear bandits. In *Conference on Learning Theory*, pp. 3285–3312. PMLR, 2022.

Supplementary Materials

A Auxiliary Lemmas and Omitted Algorithms **13**

A.1 Matrix Geometric Resampling Procedure . . . 13

A.2 Dilated Bonus Calculation in Linear-Q MDP Algorithms . . . 13

A.3 POLICYCOVER Algorithm in Linear MDP Algorithms . . . 13

A.4 Stochastic Matrix Concentration . . . 14

B Omitted Proofs in Section 3 (Linear-Q Algorithm Using Log-Barrier Regularizers) **17**

B.1 Property of FTRL with Log-Barrier Regularizers . . . 17

B.2 Proof of Main Theorem . . . 17

B.3 Bounding BIAS-1 and BIAS-2 . . . 19

B.4 Bounding REG-TERM . . . 20

C Omitted Proofs in Section 4 (Linear-Q Algorithm Using Magnitude-Reduced Estimators) **22**

C.1 Property of FTRL with Negative-Entropy Regularizers (a.k.a. Hedge) . . . 22

C.2 Proof of Main Theorem . . . 22

C.3 Bounding BIAS-1 and BIAS-2 . . . 24

C.4 Bounding REG-TERM . . . 24

D Omitted Proofs in Section 5 (Linear MDP Algorithm) **28**

D.1 Pseudocode of the Improved Linear MDP Algorithm . . . 28

D.2 Alternative to the Matrix Geometric Resampling Procedure . . . 28

D.3 Proof of Main Theorem . . . 29

D.4 Bounding the Bias Terms . . . 31

D.5 Bounding REG-TERM . . . 32

D.6 Bounding the Magnitudes of Bonuses . . . 34

A. Auxiliary Lemmas and Omitted Algorithms

A.1. Matrix Geometric Resampling Procedure

Matrix Geometric Resampling algorithm, introduced by Neu & Olkhovskaya (2020) and improved by Luo et al. (2021b), generates a close estimation of $(\gamma I + \Sigma_h^\pi)^{-1}$ for given $\pi \in \Pi$ and $h \in [H]$. Formally, we present it in Algorithm 3 and state its performance guarantees in Lemma A.1.

Algorithm 3 Matrix Geometric Resampling Algorithm

Input: Policy π, regularization parameter γ, desired precision ϵ.
Output: A set of matrices $\{\widehat{\Sigma}_h^\dagger\}_{h=1}^H$.

Set $M \geq \frac{24\ln(dHT)}{\epsilon^2\gamma^2}$ and $N \geq \frac{2}{\gamma}\ln\frac{1}{\epsilon\gamma}$. Set $c = \frac{1}{2}$.
Draw MN trajectories of π using the simulator; denote them by $\{(s_{m,n,h}, a_{m,n,h})\}_{m\in[M],n\in[N],h\in[H]}$.
for $m = 1, 2, \ldots, M$ **do**
  for $n = 1, 2, \ldots, N$ **do**
    Set $Y_{n,h} = \gamma I + \phi(s_{m,n,h}, a_{m,n,h})\phi(s_{m,n,h}, a_{m,n,h})^\top$ and $Z_{n,h} = \prod_{n'=1}^n (I - cY_{n',h})$ for all $h \in [H]$.
  end for
  Calculate $\widehat{\Sigma}_{m,h}^\dagger = cI + c\sum_{n=1}^N Z_{n,h}$ for all $h \in [H]$.
end for
Set $\widehat{\Sigma}_h^\dagger = \frac{1}{M}\sum_{m=1}^M \widehat{\Sigma}_{m,h}^\dagger$ for all $h \in [H]$.

Lemma A.1 (Lemma D.1 of Luo et al. (2021b))**.** *With the configurations of M and N stated in Algorithm 3, for any policy π, we have $\|\widehat{\Sigma}_h^\dagger\|_2 \leq \gamma^{-1}$, $\|\mathbb{E}[\widehat{\Sigma}_h^\dagger] - (\gamma I + \Sigma_h^\pi)^{-1}\|_2 \leq \epsilon$. Moreover, there exists a good event that happens with probability $1 - \frac{1}{K^3}$, under which the following two properties hold:*

$$\|\mathbb{E}[\widehat{\Sigma}_h^\dagger] - (\gamma I + \Sigma_h^\pi)^{-1}\|_2 \leq 2\epsilon, \quad \|\widehat{\Sigma}_h^\dagger \Sigma_h^\pi\|_2 \leq 1 + 2\epsilon.$$

A.2. Dilated Bonus Calculation in Linear-Q MDP Algorithms

The following algorithm (which is Algorithm 3 of Luo et al. (2021b)) indicates how we calculate the dilated bonus function $B_k(s, a)$ with the help of the simulator in Algorithms 1 and 2. In both algorithms, we define the bonus as

$$b_k(s,a) = \beta\left(\|\phi(s,a)\|^2_{\widehat{\Sigma}^\dagger_{j,h}} + \mathop{\mathbb{E}}_{\widetilde{a}\sim\pi_k(\cdot|s)}\left[\|\phi(s,\widetilde{a})\|^2_{\widehat{\Sigma}^\dagger_{j,h}}\right]\right).$$

Algorithm 4 Dilated Bonus Calculation in Linear-Q Algorithms

Input: Episode $k \in [K]$, state $s \in \mathcal{S}$, action $a \in \mathcal{A}$.
Output: The dilated bonus function $B_k(s, a)$.

if BONUS(k, s, a) is called before **then return** the value calculated at that time.
Let h be the layer that s lies in, i.e., $s \in \mathcal{S}_h$. **if** $h = H + 1$ **then return** 0.
Call the simulator for a next state $s' \sim \mathbb{P}(\cdot \mid s, a)$. Calculate $\pi_k(\cdot \mid s)$ and $\pi_k(\cdot \mid s')$ according to the algorithm.
return $\beta\left(\|\phi(s,a)\|^2_{\widehat{\Sigma}^\dagger_{k,h}} + \mathbb{E}_{\widetilde{a}\sim\pi_k(\cdot|s)}\left[\|\phi(s,\widetilde{a})\|^2_{\widehat{\Sigma}^\dagger_{k,h}}\right]\right) + (1 + \frac{1}{H})B_k(s', a')$ where a' is a sample from $\pi_k(\cdot \mid s')$.

Note that, during the calculation of π_k and $B_k(s', a')$, we are essentially recursively calling this algorithm.

A.3. POLICYCOVER Algorithm in Linear MDP Algorithms

The following algorithm is Algorithm 6 of Luo et al. (2021a) (which itself builds upon Algorithm 1 of Wang et al. (2020); we refer the readers to the original paper for more details). This algorithm generates a mixture of policies, which we call the policy cover π_{cov}, together with an estimate of the (regularized) inverse of its covariance matrix, namely $\{\widehat{\Sigma}_h^{\text{cov}}\}_{h=1}^H$. It ensures the following property:

Lemma A.2 (Lemma D.4 by Luo et al. (2021a)). *With probability $1-\delta$, we have the following for all policies π and any $h \in [H]$:*

$$\Pr_{s_h \sim \pi}[s_h \notin \mathcal{K}] = \widetilde{\mathcal{O}}\left(\frac{dH}{\alpha}\right),$$

where $\mathcal{K}$ is set of known states defined as follows:

$$\mathcal{K} = \left\{ s \in \mathcal{S} \middle| \forall a \in \mathcal{A}, \|\phi(s,a)\|^2_{(\widehat{\Sigma}_h^{cov})^{-1}} \le \alpha \right\},$$

where h is the layer that s lies in.

Algorithm 5 POLICYCOVER for Linear MDPs

Input: Configurations M_0, N_0, α, failure probability δ.

Set $\Gamma_{1,h} \leftarrow I$ for all $h \in [H]$. Set $\widetilde{\beta} = 60dH\sqrt{\ln \frac{K}{\delta}}$.
for $m = 1, 2, \ldots, M_0$ **do**
  Set $\widehat{V}_m(x_{h+1}) \leftarrow 0$.
  for $h = H, H-1, \ldots, 1$ **do**
    For all $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, let

$$\begin{aligned}
\widehat{Q}_m(s,a) &= \min\{r_m(s,a) + \widetilde{\beta}\|\phi(s,a)\|_{\Gamma_{m,h}^{-1}} + \phi(s,a)^\top \widehat{\theta}_{m,h}, H\}, \\
\widehat{V}_m(s) &= \max_{a \in \mathcal{A}} \widehat{Q}_m(s,a), \\
\pi_m(a \mid s) &= \mathbb{1}[a = \operatorname*{argmax}_{a' \in \mathcal{A}} \widehat{Q}_m(s,a')],
\end{aligned}$$

    where

$$\begin{aligned}
r_m(s,a) &= \text{ramp}_{\frac{1}{K}}(\|\phi(s,a)\|_{\Gamma_{m,h}^{-1}} - \frac{\alpha}{M_0}), \\
\widehat{\theta}_{m,h} &= \Gamma_{m,h}^{-1}(\frac{1}{N_0}\sum_{m'=1}^{m-1}\sum_{n'=1}^{N_0} \phi(s_{m',n',h}, a_{m',n',h})\widehat{V}_m(s_{m',n',h+1})).
\end{aligned}$$

    Here, $\text{ramp}_z(y)$ is 0 if $y \le -z$, 1 if $y \ge 0$, and $\frac{y}{z}+1$ otherwise.
  end for
  for $n = 1, 2, \ldots, N_0$ **do**
    Execute π_m and get trajectory $\{(s_{m,n,h}, a_{m,n,h})\}_{h=1}^H$.
  end for
  Compute $\Gamma_{m+1,h}$ as

$$\Gamma_{m+1,h} \leftarrow \Gamma_{m,h} + \frac{1}{N_0}\sum_{n=1}^{N_0} \phi(s_{m,n,h}, a_{m,n,h})\phi(s_{m,n,h}, a_{m,n,h})^\top.$$

  return π_{cov} defined as the uniform mixture of $\pi_1, \pi_2, \ldots, \pi_{M_0}$ and $\widehat{\Sigma}_h^{\text{cov}}$ defined as $\frac{1}{M_0}\Gamma_{M_0+1,h}$ (for all h).
end for

A.4. Stochastic Matrix Concentration

We first state the well-known Matrix Azuma inequality.

Lemma A.3 (Matrix Azuma; see (Tropp, 2012, Theorem 7.1)). *Let $\{X_k\}_{k=1}^n$ be a adapted sequence of $d \times d$ self-adjoint matrices. Let $\{A_k\}_{k=1}^n$ be a fixed sequence of self-adjoint matrices such that*

$$\mathbb{E}_k[X_k] = 0, \quad X_k^2 \preceq A_k^2 \text{ almost surely}.$$

Let $\sigma^2 = \|\frac{1}{n}\sum_{k=1}^n A_k^2\|_2$. *Then for all* $\epsilon > 0$*:*

$$\Pr\left\{\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_2 \geq \epsilon\right\} \leq d\exp\left(-\frac{n\epsilon^2}{8\sigma^2}\right).$$

Then, we state the following lemma, which we use to replace the MGR procedure in Linear MDPs.

Lemma A.4. *Let* $H_1, H_2, \ldots, H_n$ *be i.i.d. PSD matrices s.t.* $\mathbb{E}[H_i] = H$*,* $H_i \preceq I$ *a.s., and* $H \succeq \frac{1}{dn}\log\frac{d}{\delta} I$*, then with probability* $1-2\delta$*,*

$$-\sqrt{\frac{d}{n}\log\frac{d}{\delta}}H^{1/2} \preceq \frac{1}{n}\sum_{i=1}^n H_i - H \preceq \sqrt{\frac{d}{n}\log\frac{d}{\delta}}H^{1/2}.$$

Proof. Let $G = \sqrt{\frac{1}{dn}\log\frac{d}{\delta}}H^{-1/2}$. We first show that

$$-\frac{2\log\frac{d}{\delta}}{n}G^{-1} \preceq \frac{1}{n}\sum_{i=1}^n H_i - H$$

holds with probability $1-\delta$. We have

$$\Pr\left\{-\frac{2\log(\frac{d}{\delta})}{n}G^{-1} \not\preceq \frac{1}{n}\sum_{i=1}^n H_i - H\right\} = \Pr\left\{\sum_{i=1}^n G^{1/2}\left(H - H_i\right)G^{1/2} - \log(\frac{d}{\delta})I \not\preceq \log(\frac{d}{\delta})I\right\}.$$

Since $\log$ is operator monotone, we have for PSD matrices A, B: $\exp(A) \preceq \exp(B) \Rightarrow \log(\exp(A)) \preceq \log(\exp(B))$ (note that the reverse does not generally hold). We have $\Pr\{\exp(A) \preceq \exp(B)\} \leq \Pr\{A \preceq B\}$ and hence $\Pr\{A \not\preceq B\} \leq \Pr\{\exp(A) \not\preceq \exp(B)\}$. Hence

$$\begin{aligned}
&\Pr\left\{\sum_{i=1}^n G^{1/2}\left(H - H_i\right)G^{1/2} - \log(\frac{d}{\delta})I \not\preceq \log(\frac{d}{\delta})I\right\} \\
&\leq \Pr\left\{\exp\left(\sum_{i=1}^n G^{1/2}\left(H - H_i\right)G^{1/2} - \log(\frac{d}{\delta})I\right) \not\preceq \frac{d}{\delta}I\right\} \\
&\leq \Pr\left\{\mathrm{Tr}\left(\exp\left(\sum_{i=1}^n G^{1/2}\left(H - H_i\right)G^{1/2} - \log(\frac{d}{\delta})I\right)\right) > \frac{d}{\delta}\right\} \\
&\leq \frac{\mathbb{E}\left[\mathrm{Tr}\left(\exp\left(\sum_{i=1}^n G^{1/2}\left(H - H_i\right)G^{1/2} - \log(\frac{d}{\delta})I\right)\right)\right]}{d}\delta \qquad \text{(Markov's inequality)}
\end{aligned}$$

Using the Golden-Thompson inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Tr}\left(\exp\left(\sum_{i=1}^n G^{1/2}\left(H - H_i\right)G^{1/2} - \log(\frac{d}{\delta})I\right)\right)\right] \\
&\leq \mathbb{E}\left[\mathrm{Tr}\left(\exp\left(\sum_{i=1}^{n-1} G^{1/2}\left(H - H_i\right)G^{1/2} - \frac{\log(\frac{d}{\delta})}{n}I\right)\exp(G^{1/2}\left(H - H_n\right)G^{1/2} - \frac{\log(\frac{d}{\delta})}{n}I)\right)\right] \\
&= \mathrm{Tr}\left(\mathbb{E}\left[\exp\left(\sum_{i=1}^{n-1} G^{1/2}\left(H - H_i\right)G^{1/2} - \frac{\log(\frac{d}{\delta})}{n}I\right)\right]\mathbb{E}\left[\exp(G^{1/2}\left(H - H_n\right)G^{1/2}\right]\exp\left(-\frac{\log(\frac{d}{\delta})}{n}I)\right)\right)
\end{aligned}$$

We have due to $G^{1/2}(H - H_n)G^{1/2} \preceq I$:

$$\mathbb{E}\left[\exp(G^{1/2}\left(H - H_n\right)G^{1/2})\right] \preceq \mathbb{E}[I + (G^{1/2}\left(H - H_n\right)G^{1/2}) + (G^{1/2}\left(H - H_n\right)G^{1/2})^2] \preceq I + \mathbb{E}\left[(G^{1/2}H_nG^{1/2})^2\right].$$

By the Araki–Lieb–Thirring inequality, we have $\mathrm{Tr}((ABA)^2) \le \mathrm{Tr}(A^2B^2A^2)$, hence

$$\mathrm{Tr}\left(\mathbb{E}[(G^{1/2}H_nG^{1/2})^2]\right) \le \mathbb{E}[\mathrm{Tr}(GH_n^2G)] \le \mathbb{E}[\mathrm{Tr}(GH_nG)] \qquad (H_n \preceq I)$$
$$= \mathrm{Tr}(GHG) \le \frac{\log(\frac{d}{\delta})}{n}.$$

Combining this with the previous result yields

$$\mathbb{E}\left[\exp(G^{1/2}\,(H-H_n)\,G^{1/2})\right]\exp\left(-\frac{\log(\frac{d}{\delta})}{n}I)\right) \preceq \left(1+\frac{\log(\frac{d}{\delta})}{n}\right)I\exp\left(-\frac{\log(\frac{d}{\delta})}{n}I\right) \preceq I\,.$$

Applying this recursively yields

$$\mathbb{E}\left[\mathrm{Tr}\left(\exp\left(\sum_{i=1}^{n}G^{1/2}\,(H-H_i)\,G^{1/2}-\log(\frac{d}{\delta})I\right)\right)\right] \le \mathrm{Tr}\,(I) = d\,,$$

which concludes the proof of the claim. By symmetry and union bound, we have the following with probability $1-2\delta$:

$$-\frac{2\log\frac{d}{\delta}}{n}G^{-1} \preceq \frac{1}{n}\sum_{i=1}^{n}H_i - H \preceq \frac{2\log\frac{d}{\delta}}{n}G^{-1}.$$

This then directly simplifies to our conclusion given $H \succeq \frac{1}{dn}\log\frac{d}{\delta}I$. □

B. Omitted Proofs in Section 3 (Linear-Q Algorithm Using Log-Barrier Regularizers)

B.1. Property of FTRL with Log-Barrier Regularizers

Proof of Lemma 3.1. We first introduce the notation of Bregman divergences $D_\Psi(y,x) = \Psi(y) - \Psi(x) - \langle \nabla\Psi(x), y-x\rangle$, which is heavily used in FTRL analyses. By standard FTRL analysis (see, e.g., Theorem 28.5 by Lattimore & Szepesvári (2020)), we know

$$\sum_{t=1}^{T}\langle x_t - y, c_t\rangle \le \frac{\Psi(y)-\Psi(x_1)}{\eta} + \sum_{t=1}^{T}\left(\langle x_t - x_{t+1}, c_t\rangle - \eta^{-1}D_\Psi(x_{t+1},x_t)\right). \tag{10}$$

Consider $D_\Psi(x_{t+1},x_t)$. By definition of D_Ψ and our choice that $\Psi(p) = \sum_{i=1}^{A}\ln\frac{1}{p_i}$, we have

$$\begin{aligned} D_\Psi(x_{t+1},x_t) &= \Psi(x_{t+1}) - \Psi(x_t) - \langle \nabla\Psi(x_t), x_{t+1}-x_t\rangle \\ &= \sum_{i=1}^{A}\left(\ln\frac{x_t}{x_{t+1}} + \frac{x_{t+1}-x_t}{x_t}\right). \end{aligned}$$

Observe that the following holds for all $x\in(0,1)$ and $x+\Delta\in(0,1)$:

$$\ln\frac{x}{x+\Delta} + \frac{\Delta}{x} = \int_0^\Delta \frac{\alpha}{x(x+\alpha)}\mathrm{d}\alpha \ge \int_0^\Delta \frac{\alpha}{x}\mathrm{d}\alpha = \frac{\Delta^2}{2x},$$

For all i, we apply the inequality about with $x = x_{t,i}$ and $\Delta = x_{t+1,i} - x_{t,i}$ (as x_t and x_{t+1} both belong to the simplex, the conditions $x\in(0,1)$ and $(x+\Delta)\in(0,1)$ indeed hold). We then get the following lower bound on $D_\Psi(x_{t+1},x_t)$:

$$D_\Psi(x_{t+1},x_t) \ge \sum_{i=1}^{A}\frac{(x_{t+1}-x_t)^2}{2x_t}. \tag{11}$$

We further have

$$\langle x_t - x_{t+1}, c_t\rangle - \eta^{-1}D_\Psi(x_{t+1},x_t) \le \sum_{i=1}^{A}\left((x_{t,i}-x_{t+1,i})c_{t,i} - \frac{(x_{t+1}-x_t)^2}{2x_t}\right) \le \sum_{i=1}^{A}\frac{1}{2}x_{t,i}\eta c_{t,i}^2,$$

where the last step uses AM-GM inequality $-a^2+2ab\le b^2$. Plugging this back to Eq. (10) gives our conclusion. □

B.2. Proof of Main Theorem

Proof of Theorem 3.3. As sketched in the main text, we consider the following expression in order to apply the dilated bonus lemma (Lemma 2.5):

$$\begin{aligned} &\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\sum_{a_h\in\mathcal{A}}(\pi_k(a_h\mid s_h)-\pi^*(a_h\mid s_h))(Q_k^{\pi_k}(s_h,a_h)-B_k(s_h,a_h))\right] \\ &= \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}\left[Q_k^{\pi_k}(s_h,a_h)-\widehat{Q}_k(s_h,a_h)\right]\right]}_{\text{BIAS-1}} + \\ &\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi^*(\cdot|s_h)}\left[\widehat{Q}_k(s_h,a_h)-Q_k^{\pi_k}(s_h,a_h)\right]\right]}_{\text{BIAS-2}} + \\ &\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\left\langle \pi_k(\cdot\mid s_h)-\pi^*(\cdot\mid s_h), \widehat{Q}_k(s_h,\cdot)-B_k(s_h,\cdot)\right\rangle\right]}_{\text{REG-TERM}}. \end{aligned} \tag{12}$$

According to Lemma B.1, we know that

$$\mathbb{E}[\textsc{Bias-1}] + \mathbb{E}[\textsc{Bias-2}] \leq \frac{\beta}{4}\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}]\right]\right] + \frac{\beta}{4}\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi^*(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}]\right]\right] + \mathcal{O}\left(\frac{\gamma}{\beta}dH^3K + \epsilon(H+\beta)HK\right).$$

Moreover, by Lemma B.2, we can see that

$$\mathbb{E}[\textsc{Reg-Term}] \leq H\eta^{-1}A\ln K + \mathcal{O}\left(AH^2\left(\epsilon+\sqrt{d}+\frac{\beta}{\gamma}\right)\right) + 2\eta H^2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}\left[\|\phi(s_h,a_h)\|_{\widehat{\Sigma}^{\dagger}_{k,h}}\right]\right] + \mathcal{O}\left(\frac{\eta H^3}{\gamma^2K^2}\right) + \frac{1}{H}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[B_k(s_h,a_h)]\right].$$

Plugging back into Eq. (12), we know that

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\sum_{a_h\in\mathcal{A}}(\pi_k(a_h\mid s_h)-\pi^*(a_h\mid s_h))(Q_k^{\pi_k}(s_h,a_h)-B_k(s_h,a_h))\right]\\
\leq\ &\frac{\beta}{4}\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}]\right]\right] +\\
&\frac{\beta}{4}\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi^*(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}]\right]\right] +\\
&2\eta H^2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}\left[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}\right]\right] +\\
&\frac{1}{H}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[B_k(s_h,a_h)]\right] +\\
&\widetilde{\mathcal{O}}\left(\frac{\gamma}{\beta}dH^3K+\epsilon(H+\beta)HK+\frac{H}{\eta}A+AH^2\left(\epsilon+\sqrt{d}+\frac{\beta}{\gamma}\right)+\frac{\eta H^3}{\gamma^2K^2}\right).
\end{aligned}$$

Using $(6\eta H^2+\frac{\beta}{4})\leq\beta$, we apply Lemma 2.5 with $b_k(s,a)=\|\phi(s,a)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}+\mathbb{E}_{a'\sim\pi_k(\cdot|s)}[\|\phi(s,a')\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}]$, giving

$$\mathcal{R}_K \leq \beta\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi_k}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}]\right]\right] + \widetilde{\mathcal{O}}\left(\frac{\gamma}{\beta}dH^3K+\epsilon(H+\beta)HK+\frac{H}{\eta}A+AH^2\left(\epsilon+\sqrt{d}+\frac{\beta}{\gamma}\right)+\frac{\eta H^3}{\gamma^2K^2}\right).$$

Fixing an episode $k\in[K]$ and $h\in[H]$, we have

$$\mathop{\mathbb{E}}_{s_h\sim\pi_k}\left[\sum_a \pi_k(a\mid s)\|\phi(s,a)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}\right] \leq \mathop{\mathbb{E}}_{s_h\sim\pi_k}\left[\sum_a \pi_k(a\mid s)\|\phi(s,a)\|^2_{(\gamma I+\Sigma_h^{\pi_k})^{-1}}\right] + \mathcal{O}(\epsilon)$$

$$\leq \mathop{\mathbb{E}}_{s_h\sim\pi_k}\left[\sum_a \pi_k(a\mid s)\|\phi(s,a)\|^2_{(\Sigma_h^{\pi_k})^{-1}}\right] + \mathcal{O}(\epsilon)$$
$$= \left\langle (\Sigma_h^{\pi_k})^{-1}, \mathop{\mathbb{E}}_{(s_h,a_h)\sim\pi_k}[\phi(s_h,a_h)\phi(s_h,a_h)^\top]\right\rangle = d. \tag{13}$$

Therefore, we can conclude the following if $12\eta\beta H^2 \leq \gamma$ and $8\eta H^2 \leq \beta$:

$$\mathcal{R}_K = \widetilde{\mathcal{O}}\left(\beta dHK + \frac{\gamma}{\beta}dH^3K + \epsilon(H+\beta)HK + \frac{H}{\eta}A + AH^2\left(\epsilon + \sqrt{d} + \frac{\beta}{\gamma}\right) + \frac{\eta H^3}{\gamma^2K^2}\right).$$

It remains to tune the parameters. We first pick $\epsilon = \frac{1}{H^2K}$, which makes all terms related to ϵ constantly-bounded. Setting $\eta \approx K^{-1/2}$ and $\gamma \approx K^{-1}$, the last term is also $o(1)$. Hence, removing all constantly-bounded terms give

$$\mathcal{R}_K = \widetilde{\mathcal{O}}\left(AH\frac{1}{\eta} + \beta dHK + \frac{\gamma}{\beta}dH^3K + AH^2\frac{\beta}{\gamma}\right).$$

We then get $\mathcal{R}_K = \widetilde{\mathcal{O}}(\sqrt{AdH^6K}) = \widetilde{\mathcal{O}}(A^{1/2}d^{1/2}H^3K^{1/2})$ by picking:

$$\eta = \sqrt{\frac{A}{dH^4K}}, \beta = 8\sqrt{\frac{A}{dK}}, \gamma = 96\frac{A}{dK}.$$

It's straightforward to verify that they satisfy $12\eta\beta H^2 \leq \gamma$ and $8\eta H^2 \leq \beta$. □

B.3. Bounding BIAS-1 and BIAS-2

Lemma B.1. *In Algorithm 1, we have*

$$\mathbb{E}[\text{BIAS-1}] + \mathbb{E}[\text{BIAS-2}] \leq \frac{\beta}{4}\mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{k,h}}]\right]\right] +$$
$$\frac{\beta}{4}\mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi^*(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{k,h}}]\right]\right] + \mathcal{O}\left(\frac{\gamma}{\beta}dH^3K + \epsilon(H+\beta)HK\right).$$

Proof. As BIAS-1 has nothing to do with the choice of the regularizer, it can be bounded the same as the original algorithm (Luo et al., 2021b, Lemma D.2), which we also include below for completeness: fixing a specific (k,s,a) and suppose that $s \in \mathcal{S}_h$. Then we have the following, where every expectation is taken to the randomness in the k-th episode:

$$\begin{aligned}
\mathbb{E}\left[Q_k^{\pi_k}(s,a) - \widehat{Q}_k(s,a)\right] &= \phi(s,a)^\top\theta_{k,h}^{\pi_k} - \phi(s,a)^\top\mathbb{E}\left[\widehat{\Sigma}^\dagger_{k,h}\phi(s_{k,h},a_{k,h})L_k\right] \\
&= \phi(s,a)^\top\theta_{k,h}^{\pi_k} - \phi(s,a)^\top\mathbb{E}\left[\widehat{\Sigma}^\dagger_{k,h}\right]\mathbb{E}\left[\phi(s_{k,h},a_{k,h})\phi(s_{k,h},a_{k,h})^\top\theta_{k,h}^{\pi_k}\right] \\
&\overset{(a)}{\leq} \phi(s,a)^\top\left(I - (\gamma I + \Sigma_h^{\pi_k})^{-1}\Sigma_h^{\pi_k}\right)\theta_{k,h}^{\pi_k} + \epsilon H \\
&= \gamma\phi(s,a)^\top(\gamma I + \Sigma_h^{\pi_k})^{-1}\theta_{k,h}^{\pi_k} + \epsilon H \\
&\overset{(b)}{\leq} \gamma\|\phi(s,a)\|_{(\gamma I+\Sigma_h^{\pi_k})^{-1}}\|\theta_{k,h}^{\pi_k}\|_{(\gamma I+\Sigma_h^{\pi_k})^{-1}} + \epsilon H \\
&\overset{(c)}{\leq} \frac{\beta}{4}\|\phi(s,a)\|^2_{(\gamma I+\Sigma_h^{\pi_k})^{-1}} + \frac{\gamma^2}{\beta}\|\theta_{k,h}^{\pi_k}\|^2_{(\gamma I+\Sigma_h^{\pi_k})^{-1}} + \epsilon H \\
&\overset{(d)}{\leq} \frac{\beta}{4}\mathbb{E}\left[\|\phi(s,a)\|^2_{\widehat{\Sigma}^\dagger_{k,h}}\right] + \frac{\gamma}{\beta}dH^2 + \epsilon H + \epsilon\frac{\beta}{4}.
\end{aligned}$$

where (a) used Eq. (7) (which follows from Lemma A.1) and the assumption that $\|\phi(s,a)\| \le 1$ and $L_k \le H$, (b) used Cauchy-Schwartz inequality, (c) used AM-GM inequality, and (d) used the assumption that $\|\theta_{k,h}^{\pi_k}\| \le \sqrt{d}H$ (see Definition 2.2) and again Eq. (7). Hence, we have

$$\mathbb{E}[\textsc{Bias-1}] \le \frac{\beta}{4}\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}]\right]\right] + \mathcal{O}\left(\frac{\gamma}{\beta}dH^3K + \epsilon(H+\beta)HK\right).$$

Similarly, we have

$$\mathbb{E}[\textsc{Bias-2}] \le \frac{\beta}{4}\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi^*(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}]\right]\right] + \mathcal{O}\left(\frac{\gamma}{\beta}dH^3K + \epsilon(H+\beta)HK\right).$$

Combining these two parts together gives our conclusion. □

B.4. Bounding Reg-Term

Lemma B.2. *Under the assumption that $12\eta\beta H^2 \le \gamma$, we have the following in Algorithm 1:*

$$\begin{aligned}\mathbb{E}[\textsc{Reg-Term}] \le{}& H\eta^{-1}A\ln K + \mathcal{O}\left(AH^2\left(\epsilon+\sqrt{d}+\frac{\beta}{\gamma}\right)\right) + \\ & 2\eta H^2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}\left[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}\right]\right] + \mathcal{O}\left(\frac{\eta H^3}{\gamma^2K^2}\right) + \\ & \frac{1}{H}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[B_k(s_h,a_h)]\right].\end{aligned}$$

Proof. Using Lemma 3.1, we get the following for all $k\in[K]$, $s\in\mathcal{S}_h$ (where $h\in[H]$), and $\widetilde{\pi}^*\in\Pi$:

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}(\pi_k(a\mid s)-\pi^*(a\mid s))(\widehat{Q}_k(s,a)-B_k(s,a)) \\ \le{}& \frac{\Psi(\widetilde{\pi}^*(\cdot\mid s))-\Psi(\pi_1(\cdot\mid s))}{\eta}+ \\ &\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}(\widetilde{\pi}^*(a\mid s)-\pi^*(a\mid s))(\widehat{Q}_k(s,a)-B_k(s,a))+ \\ &\eta\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\pi_k(a\mid s)(\widehat{Q}_k(s,a)-B_k(s,a))^2.\end{aligned}$$

By picking $\widetilde{\pi}^*(a\mid s)=(1-AK^{-1})\pi^*(a\mid s)+K^{-1}$, the first term is bounded by

$$\frac{\Psi(\widetilde{\pi}^*(\cdot\mid s))-\Psi(\pi_1(\cdot\mid s))}{\eta} = \frac{1}{\eta}\sum_{a\in\mathcal{A}}\ln\frac{\pi_1(a\mid s)}{\widetilde{\pi}^*(a\mid s)} \le \frac{1}{\eta}\sum_{a\in\mathcal{A}}\ln\frac{\pi_1(a\mid s)}{K^{-1}} \le \eta^{-1}A\ln K.$$

Meanwhile, the second term is bounded by

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}(\widetilde{\pi}^*(a\mid s)-\pi^*(a\mid s))(\widehat{Q}_k(s,a)-B_k(s,a)) \\ =&\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}(-AK^{-1}\pi^*(a\mid s)+K^{-1})(\widehat{Q}_k(s,a)-B_k(s,a)).\end{aligned}$$

Firstly, we have the following as $\|\widehat{\Sigma}_{k,h}^{\dagger}\|_2 \le \gamma^{-1}$:

$$B_k(s,a) \le H\left(1+\frac{1}{H}\right)^H \times 2\beta \sup_{s,a,h} \|\phi(s,a)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}} \le 6\beta H\gamma^{-1}. \tag{14}$$

Moreover, according to the calculation in Lemma B.1, we know that

$$\mathbb{E}[\widehat{Q}_k(s,a) - Q_k^{\pi_k}(s,a)] \le \gamma\phi(s,a)^{\mathsf{T}}(\gamma I + \Sigma_h^{\pi_k})^{-1}\theta_{k,h}^{\pi_k} + \epsilon H \le \mathcal{O}((\epsilon+\sqrt{d})H),$$

while, at the same time, $Q_k^{\pi_k}(s,a) \in [0,H]$.

Hence, after taking expectations on both sides, we know that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^K \sum_{a\in\mathcal{A}} (\widetilde{\pi}^*(a \mid s) - \pi^*(a \mid s))(\widehat{Q}_k(s,a) - B_k(s,a))\right] \\
&\le \sum_{k=1}^K \sum_{a\in\mathcal{A}} (|-AK^{-1}\pi^*(a \mid s)| + |K^{-1}|)|\mathbb{E}[\widehat{Q}_k(s,a) - B_k(s,a)]| \\
&\le K \times 2AK^{-1} \times \mathcal{O}\left((\epsilon+\sqrt{d})H + H + \frac{\beta}{\gamma}H\right) = \mathcal{O}\left(AH\left(\epsilon+\sqrt{d}+\frac{\beta}{\gamma}\right)\right).
\end{aligned}$$

Then consider the last term, which is directly bounded by

$$\begin{aligned}
&\eta\sum_{k=1}^K \sum_{a\in\mathcal{A}} \pi_k(a \mid s)^2(\widehat{Q}_k(s,a) - B_k(s,a))^2 \\
&\le 2\eta\sum_{k=1}^K \sum_{a\in\mathcal{A}} \pi_k(a \mid s)\widehat{Q}_k(s,a)^2 + 2\eta\sum_{k=1}^K \sum_{a\in\mathcal{A}} \pi_k(a \mid s)B_k(s,a)^2.
\end{aligned}$$

The first term can be calculated as follows, following the original proof (Luo et al., 2021b, Lemma D.3):

$$\begin{aligned}
\mathbb{E}[\widehat{Q}_k(s,a)^2] &\le H^2\,\mathbb{E}[\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s_{k,h},a_{k,h})\phi(s_{k,h},a_{k,h})^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s,a)] \\
&= H^2\,\mathbb{E}[\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\Sigma_h^{\pi_k}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s,a)] \\
&\overset{(a)}{\le} 2H^2\,\mathbb{E}[\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s,a)] + \mathcal{O}\left(\frac{H^2}{\gamma^2K^3}\right) = 2H^2\,\mathbb{E}\left[\|\phi(s,a)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}\right] + \mathcal{O}\left(\frac{H^2}{\gamma^2K^3}\right),
\end{aligned}$$

where (a) used Eq. (8), which happens with probability $1-K^{-3}$ for each k (when it does not hold, we simply use the bound $\|\widehat{\Sigma}_{k,h}^{\dagger}\|_2 \le \gamma^{-1}$). Then we can conclude the following by adding back the summation over $h \in [H]$ and $s_h \sim \pi^*$:

$$\begin{aligned}
\mathbb{E}[\textsc{Reg-Term}] \le{}& H\eta^{-1}A\ln K + \mathcal{O}\left(AH^2\left(\epsilon+\sqrt{d}+\frac{\beta}{\gamma}\right)\right) + \\
&2\eta H^2\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}\left[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}\right]\right] + \mathcal{O}\left(\frac{\eta H^3}{\gamma^2K^2}\right) + \\
&\frac{1}{H}\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[B_k(s_h,a_h)]\right],
\end{aligned}$$

where the last term comes from the magnitude of B_k (Eq. (14)) and the assumption that $12\eta\beta H^2 \le \gamma$. □

C. Omitted Proofs in Section 4 (Linear-Q Algorithm Using Magnitude-Reduced Estimators)

C.1. Property of FTRL with Negative-Entropy Regularizers (a.k.a. Hedge)

The following result is a classic result for the Hedge algorithm. For the sake of completeness, we also include a proof here.

Lemma C.1. *Let $x_0, x_1, x_2, \ldots, x_T \in \mathbb{R}^A$ be defined as*

$$x_{t+1,i} = (x_{t,i}\exp(-\eta c_{t,i})) \bigg/ \left(\sum_{i'=1}^{A} x_{t,i'}\exp(-\eta c_{t,i'})\right), \quad \forall 0 \le t < T,$$

where $c_t \in \mathbb{R}^A$ is the loss corresponding to the t-th iteration. Suppose that $\eta c_{t,i} \ge -1$ for all $t \in [T]$ and $i \in [A]$. Then

$$\sum_{t=1}^{T}\langle x_t - y, c_t\rangle \le \frac{\log A}{\eta} + \eta\sum_{t=1}^{T}\sum_{i=1}^{A} x_{t,i}c_{t,i}^2$$

holds for any distribution $y \in \triangle([A])$ when $x_0 = (\frac{1}{A}, \frac{1}{A}, \ldots, \frac{1}{A})$.

Proof. By linearity, it suffices to prove the inequality for all one-hot y's. Without loss of generality, let $y = \mathbf{1}_{i^*}$ where $i^* \in [A]$. Define $C_{t,i} = \sum_{t'=1}^{t} c_{t',i}$ as the prefix sum of $c_{t,i}$. Let

$$\Phi_t = \frac{1}{\eta}\ln\left(\sum_{i=1}^{A}\exp\left(-\eta C_{t,i}\right)\right),$$

then by definition of x_t, we have

$$\begin{aligned}
\Phi_t - \Phi_{t-1} &= \frac{1}{\eta}\ln\left(\frac{\sum_{i=1}^{A}\exp(-\eta C_{t,i})}{\sum_{i=1}^{A}\exp(-\eta C_{t-1,i})}\right) = \frac{1}{\eta}\ln\left(\sum_{i=1}^{A} x_{t,i}\exp(-\eta c_{t,i})\right) \\
&\overset{(a)}{\le} \frac{1}{\eta}\ln\left(\sum_{i=1}^{A} x_{t,i}(1 - \eta c_{t,i} + \eta^2 c_{t,i}^2)\right) = \frac{1}{\eta}\ln\left(1 - \eta\langle x_t, c_t\rangle + \eta^2\sum_{i=1}^{A} x_{t,i}c_{t,i}^2\right) \\
&\overset{(b)}{\le} -\langle x_t, c_t\rangle + \eta\sum_{i=1}^{A} x_{t,i}c_{t,i}^2,
\end{aligned}$$

where (a) used $\exp(-x) \le 1 - x + x^2$ for all $x \ge -1$ and (b) used $\ln(1+x) \le x$ (again for all $x \ge -1$). Therefore, summing over $t = 1, 2, \ldots, T$ gives

$$\begin{aligned}
\sum_{t=1}^{T}\langle x_t, c_t\rangle &\le \Phi_0 - \Phi_T + \eta\sum_{t=1}^{T}\sum_{i=1}^{A} x_{t,i}c_{t,i}^2 \\
&\le \frac{\ln N}{\eta} - \frac{1}{\eta}\ln\left(\exp(-\eta C_{T,i^*})\right) + \eta\sum_{t=1}^{T}\sum_{i=1}^{N} p_t(i)\ell_t^2(i) \\
&\le \frac{\ln A}{\eta} + L_T(i^*) + \eta\sum_{t=1}^{T}\sum_{i=1}^{A} x_{t,i}c_{t,i}^2.
\end{aligned}$$

Moving C_{t,i^*} to the LHS then shows the inequality for $y = \mathbf{1}_{i^*}$. The result then extends to all $y \in \triangle([A])$ by linearity. □

C.2. Proof of Main Theorem

Proof of Theorem 4.1. We first consider Line 6 in the algorithm. As sketched in the main text, we shall expect such an operation to be repeated for $1 + o(1)$ times because Eq. (8) and $\|\widetilde{\Sigma}_{k,h} - \Sigma_h^{\pi_k}\|_2 \le \gamma$ both happens with probability $1 - K^{-3}$ — the first claim follows from Lemma A.1 and the second one comes from Lemma A.3 (where we set $X_m =$

$\phi(s_{m,h}, a_{m,h})\phi(s_{m,h}, a_{m,h})^\top - \Sigma_h^{\pi_k}$ and $A_{m,h} = I \succeq X_{m,h}$). With these two conditions and the fact that $\|\widehat{\Sigma}_{k,h}^\dagger\|_2 \le \gamma^{-1}$, the desired condition trivially holds. Therefore, such an operation brings neither extra regret nor extra computational complexity, and we focus on the regret analysis from now on.

We define BIAS-1, BIAS-2, and REG-TERM exactly the same as Theorem 3.3 (i.e., Eq. (12)). The BIAS-1 and BIAS-2 terms are bounded by Lemma C.2, as follows:

$$\mathbb{E}[\text{BIAS-1}] + \mathbb{E}[\text{BIAS-2}] \le \frac{\beta}{4}\mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{k,h}}]\right]\right] + \frac{\beta}{4}\mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi^*(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{k,h}}]\right]\right] + \mathcal{O}\left(\frac{\gamma}{\beta}dH^3K + \epsilon(H+\beta)HK\right).$$

Assuming $12\eta^2H^2 \le \gamma$ and $12\eta\beta H^2 \le \gamma$, REG-TERM is bounded by Lemma C.3 as

$$\mathbb{E}[\text{REG-TERM}] \le H\eta^{-1}\ln A + 6\eta H^2\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}\left[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{k,h}}\right]\right] + \frac{1}{H}\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[B_k(s_h,a_h)]\right].$$

Plugging into the regret decomposition, we get

$$\begin{aligned}
&\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\sum_{a_h\in\mathcal{A}}(\pi_k(a_h\mid s_h) - \pi^*(a_h\mid s_h))(Q_k^{\pi_k}(s_h,a_h) - B_k(s_h,a_h))\right] \\
\le\ & \frac{\beta}{4}\mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{k,h}}]\right]\right] + \\
& \frac{\beta}{4}\mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi^*(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{k,h}}]\right]\right] + \\
& 6\eta H^2\,\mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}\left[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{k,h}}\right]\right]\right] + \\
& \frac{1}{H}\mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[B_k(s_h,a_h)]\right]\right] + \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\gamma}{\beta}dH^3K + \epsilon(H+\beta)HK\right).
\end{aligned}$$

Using the condition that $(6\eta H^2 + \frac{\beta}{4}) \le \beta$, we can apply Lemma 2.5 to conclude that

$$\mathcal{R}_K = \widetilde{\mathcal{O}}\left(\mathbb{E}\left[\beta\sum_{h=1}^H \mathop{\mathbb{E}}_{s_h\sim\pi_k}\left[\sum_{k=1}^K\sum_a \pi_k(a\mid s)\|\phi(s,a)\|^2_{\widehat{\Sigma}^\dagger_{k,h}}\right]\right] + \frac{H}{\eta} + \frac{\gamma}{\beta}dH^3K + \epsilon(H+\beta)HK\right).$$

By Eq. (13), we can conclude the following when assuming $8\eta H^2 \le \beta$:

$$\mathcal{R}_K = \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\gamma}{\beta}dH^3K + \epsilon(H+\beta)HK + \beta dHK\right).$$

Plugging in the configurations that (again, one can see that this configuration satisfies all the conditions)

$$\eta = \frac{1}{\sqrt{dK}H^2}, \beta = \frac{8}{\sqrt{dK}}, \gamma = \frac{96}{dK}, \epsilon = \frac{1}{H^2K},$$

we then conclude that $\mathcal{R}_K = \widetilde{\mathcal{O}}(\sqrt{dH^6K})$. □

C.3. Bounding BIAS-1 and BIAS-2

Lemma C.2. *In Algorithm 2, we have*

$$\mathbb{E}[\text{BIAS-1}] + \mathbb{E}[\text{BIAS-2}] \le \frac{\beta}{4}\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}]\right]\right] + \frac{\beta}{4}\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi^*(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}]\right]\right] + \mathcal{O}\left(\frac{\gamma}{\beta}dH^3K + \epsilon(H+\beta)HK\right).$$

Proof. Fixing a specific (k,s,a) and assume $s\in\mathcal{S}_h$. Then we have (again, all expectations are taken w.r.t. randomness in the k-th episode)

$$\begin{aligned}\mathbb{E}\left[\widehat{Q}_k(s,a)\right] &= \mathbb{E}\left[\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{k,h},a_{k,h})L_{k,h}\right] - H\,\mathbb{E}\left[\left(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{k,h},a_{k,h})\right)_-\right] + \\ &\quad H\,\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\left(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{m,h},a_{m,h})\right)_-\right] \\ &= \mathbb{E}\left[\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{k,h},a_{k,h})L_{k,h}\right],\end{aligned}$$

where the last equality is because $(s_{k,h},a_{k,h})$ and $(s_{m,h},a_{m,h})$ are both sampled from π_k. The rest of the proof is then identical to Lemma B.1. □

C.4. Bounding REG-TERM

Lemma C.3. *Suppose that* $12H^2\eta^2\le\gamma$, $12\eta\beta H^2\le\gamma$, $8\eta H^2\le\gamma$, *and* $M=32\gamma^{-2}\log K$. *Then in Algorithm 2, we have*

$$\mathbb{E}[\text{REG-TERM}] \le H\eta^{-1}\ln A + 6\eta H^2\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}\left[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}\right]\right]\right] + \frac{1}{H}\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}[B_k(s_h,a_h)]\right]\right].$$

Proof. To apply the Hedge lemma (Lemma C.1), we need to ensure that

$$\eta(\widehat{Q}_k(s,a) - B_k(s,a)) \ge -1, \quad \forall (s,a)\in\mathcal{S}\times\mathcal{A}, k\in[K].$$

Fix an (s,a,k) tuple and assume that $s\in\mathcal{S}_h$. We have

$$\begin{aligned}\widehat{Q}_k(s,a) - Hm_k(s,a) &= \phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{k,h},a_{k,h})L_{k,h} - H(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{k,h},a_{k,h}))_- \\ &= \begin{cases}\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{k,h},a_{k,h})L_{k,h}, & \phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{k,h},a_{k,h})\ge 0\\ -\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{k,h},a_{k,h})(H-L_{k,h}), & \phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{k,h},a_{k,h})<0\end{cases} \\ &\ge 0,\end{aligned}$$

as $L_{k,h}\in[0,H]$. Hence, $\widehat{Q}_k(s,a)\ge Hm_k(s,a)$ always holds. We consider $\widehat{Q}_k(s,a)\ge Hm_k(s,a)$ and $B_k(s,a)$ separately.

We first claim that $\eta Hm_k(s,a)\ge -\frac{1}{2}$, which ensures $\eta\widehat{Q}_k(s,a)\ge -\frac{1}{2}$. To see this, we only need to show $\eta^2H^2m_k(s,a)^2\le\frac{1}{4}$. By definition, $m_k(s,a)^2$ can be written as the following:

$$m_k(s,a)^2 = \left(\frac{1}{M}\sum_{m=1}^{M}\left(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}^{\dagger}_{k,h}\phi(s_{m,h},a_{m,h})\right)_-\right)^2$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \frac{1}{M}\sum_{m=1}^{M}\left(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s_{m,h},a_{m,h})\right)_{-}^{2} \\
&= \phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\left(\frac{1}{M}\sum_{m=1}^{M}\phi(s_{m,h},a_{m,h})\phi(s_{m,h},a_{m,h})^{\mathsf{T}}\right)\widehat{\Sigma}_{k,h}^{\dagger}\phi(s,a) \\
&= \phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\widetilde{\Sigma}_{k,h}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s,a) \\
&\overset{(b)}{\leq} 3\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s,a) \leq 3\gamma^{-1}.
\end{aligned}
$$

where (a) used Jensen inequality, (b) used Line 6 of Algorithm 2, and the last step uses the fact that $\|\widehat{\Sigma}_{k,h}^{\dagger}\|_2 \leq \gamma^{-1}$. Hence, it only remains to ensure that $12H^2\eta^2\gamma^{-1} \leq 1$, which is guaranteed by the assumption.

Meanwhile, we claim that $\eta B_k(s,a) \leq \frac{1}{2}$. By definition of $B_k(s,a)$ and the fact that $\|\widehat{\Sigma}_{k,h}^{\dagger}\|_2 \leq \gamma^{-1}$, we have

$$
\eta B_k(s,a) \leq \eta H\left(1+\frac{1}{H}\right)^{H} \times 2\beta \sup_{s,a,h}\|\phi(s,a)\|_{\widehat{\Sigma}_{k,h}^{\dagger}}^{2} \leq 6\eta\beta H\gamma^{-1}, \tag{15}
$$

which is bounded by $\frac{1}{2H}$ according to the condition that $12\eta\beta H^2 \leq \gamma$.

Therefore, fixing $h \in [H]$ and $s \in \mathcal{S}_h$, we can apply the Hedge lemma (Lemma C.1):

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}(\pi_k(a \mid s) - \pi^*(a \mid s))(\widehat{Q}_k(s,a) - B_k(s,a))\right] \\
&\leq \frac{\ln A}{\eta} + 2\eta\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\pi_k(a \mid s)\widehat{Q}_k(s,a)^2\right] + 2\eta\,\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\pi_k(a \mid s)B_k(s,a)^2\right].
\end{aligned}
$$

For the second term, we can write

$$
\begin{aligned}
\widehat{Q}_k(s,a)^2 &= (\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s_{k,h},a_{k,h})L_{k,h} - H(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s_{k,h},a_{k,h}))_{-} + Hm_k(s,a))^2 \\
&\leq 2H^2(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s_{k,h},a_{k,h}))^2 + 2H^2(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s_{k,h},a_{k,h}))_{-}^2 + 2H^2m_k^2(s,a).
\end{aligned}
$$

After taking expectations on both sides, we have

$$
\begin{aligned}
\mathbb{E}[\widehat{Q}_k^2(s,a)] &\leq 2H^2\,\mathbb{E}[(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s_{k,h},a_{k,h}))^2] + 2H^2\,\mathbb{E}[(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s_{k,h},a_{k,h}))_{-}^2] + \\
&\quad 2H^2\,\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s_{m,h},a_{m,h}))_{-}\right]^2 \\
&\leq 6H^2\,\mathbb{E}[(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s_{k,h},a_{k,h}))^2],
\end{aligned}
$$

where we used $(X)_{-}^2 \leq X^2$, Jensen's inequality, and the fact that $(s_{k,h},a_{k,h})$ and $(s_{m,h},a_{m,h})$ are both sampled from π_k. Meanwhile, we can also calculate that

$$
\begin{aligned}
&\mathbb{E}[(\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s_{k,h},a_{k,h}))^2] \\
&= \mathbb{E}\left[\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s_{k,h},a_{k,h})\phi(s_{k,h},a_{k,h})^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s,a)\right] \\
&= \mathbb{E}\left[\phi(s,a)^{\mathsf{T}}\widehat{\Sigma}_{k,h}^{\dagger}\Sigma_k^{\pi_k}\widehat{\Sigma}_{k,h}^{\dagger}\phi(s,a)\right] \leq \mathbb{E}\left[\|\phi(s,a)\|_{\widehat{\Sigma}_{k,h}^{\dagger}}^{2}\right],
\end{aligned}
$$

where the last inequality again uses Line 6 of Algorithm 2. Hence, after summing up over the expectations when $s_h \sim \pi^*$ for all $h \in [H]$, we can conclude that

$$
\mathbb{E}[\textsc{Reg-Term}] \leq H\eta^{-1}\ln A + 6\eta H^2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_k(\cdot|s_h)}\left[\|\phi(s_h,a_h)\|_{\widehat{\Sigma}_{k,h}^{\dagger}}^{2}\right]\right]
$$

$$+ \frac{1}{H} \sum_{k=1}^{K} \sum_{h=1}^{H} \mathop{\mathbb{E}}_{s_h \sim \pi^*} \left[\mathop{\mathbb{E}}_{a_h \sim \pi_k(\cdot|s_h)} [B_k(s_h, a_h)] \right],$$

where the last term comes from Eq. (15) and the condition that $12\eta\beta H^2 \leq \gamma$. □

Algorithm 6 Improved Linear MDP Algorithm

Input: Learning rate η, PolicyCover parameters α and T_0, bonus parameter β, covariance estimation parameter $\gamma \in (0, \frac{1}{4})$, epoch length W, FTRL regularizer $\Psi(p) = \sum_{i=1}^{A} \ln \frac{1}{p_i}$, exploration probability δ_e

Let π_{cov} and $\widehat{\Sigma}_h^{\text{cov}}$ be the outputs of the PolicyCover algorithm (see Algorithm 5).

Let $\mathcal{K} = \{s \in \mathcal{S} \mid \|\phi(s,a)\|^2_{(\widehat{\Sigma}_h^{\text{cov}})^{-1}} \le \alpha, \forall a \in \mathcal{A}\}$ be all the "known" states (defined in Lemma A.2).

for $j = 1, 2, \ldots, J = (T - T_0)/W$ **do**

  Calculate $\pi_j \in \Pi$ as follows for all $s \in \mathcal{S}_h$:

$$\pi_j(s) = \operatorname*{argmin}_{p \in \triangle(\mathcal{A})} \left\{ \Psi(p) + \eta \sum_{\tau < j} \sum_{a \in \mathcal{A}} p(a) \big(\widehat{Q}_\tau(s,a) - \widehat{B}_\tau(s,a) \big) \right\},$$

  where $\widehat{Q}_\tau(s,a) = \phi(s,a)^\mathsf{T} \widehat{\theta}_{\tau,h}$, $\widehat{B}_\tau(s,a) = b_\tau(s,a) + \phi(s,a)^\mathsf{T} \widehat{\Lambda}_{\tau,h}$, and the bonus function is defined as

$$b_\tau(s,a) = \mathbb{1}[s \in \mathcal{K}] \times \beta \left(\|\phi(s,a)\|^2_{\widehat{\Sigma}^\dagger_{\tau,h}} + \mathop{\mathbb{E}}_{a' \sim \pi_\tau(\cdot|s)} \left[\|\phi(s,a')\|^2_{\widehat{\Sigma}^\dagger_{\tau,h}} \right] \right).$$

  Randomly partition the episodes in the current epoch, i.e., $\{K_0 + (j-1)W + 1, \ldots, K_0 + jW\}$, into two halves $\mathcal{T}_j$ and $\mathcal{T}'_j$, such that where $|\mathcal{T}_j| = |\mathcal{T}'_j| = \frac{W}{2}$.

  for $k = K_0 + (j-1)W + 1, \ldots, K_0 + jW$ **do**

    Let Y_k be a sample from a Bernoulli distribution $\text{Ber}(\delta_e)$.

    if $Y_k = 0$ **then**

      Execute π_j for this episode and observe $\{(s_{k,h}, a_{k,h})\}_{h=1}^H$ (together with $\ell_k(s_{k,h}, a_{k,h})$).

    else if $Y_k = 1$ and $k \in \mathcal{T}_j$ **then**

      Execute π_{cov} and observe $\{(s_{k,h}, a_{k,h})\}_{h=1}^H$ together with $\ell_k(s_{k,h}, a_{k,h})$.

    else

      Let h_k be uniformly sampled from $[H]$. Execute π_{cov} for steps $1, 2, \ldots, h-1$ and π_j for the remaining ones. Again, observe the trajectory $\{(s_{k,h}, a_{k,h})\}_{h=1}^H$ and the losses $\ell_k(s_{k,h}, a_{k,h})$.

    end if

  end for

  Define $\widetilde{\pi}_j$ as the mixture of $(1 - \delta_e)$ times π_j and δ_e times π_{cov} (i.e., expected policy played in epoch j).

  Estimate the covariance matrix $\Sigma_h^{\widetilde{\pi}_j}$ as follows, and estimate $(\gamma I + \Sigma_h^{\widetilde{\pi}_j})^{-1}$ by $\widehat{\Sigma}^\dagger_{j,h} = (\gamma I + \widetilde{\Sigma}_{j,h})^{-1}$.

$$\widetilde{\Sigma}_{j,h} = \frac{1}{|\mathcal{T}_j|} \sum_{k \in \mathcal{T}_j} \phi(s_{k,h}, a_{k,h}) \phi(s_{k,h}, a_{k,h})^\mathsf{T},$$

  Estimate the average Q-function kernel $\overline{\theta}^{\pi_j}_{j,h} \triangleq \frac{1}{|\mathcal{T}'_j|} \sum_{k \in \mathcal{T}'_j} \theta^{\pi_j}_{k,h}$ as follows:

$$\widehat{\theta}_{j,h} = \widehat{\Sigma}^\dagger_{j,h} \left(\frac{1}{|\mathcal{T}'_j|} \sum_{k \in \mathcal{T}'_j} \left((1 - Y_k) + Y_k H \mathbb{1}[h = h_k] \right) \phi(s_{k,h}, a_{k,h}) L_{k,h} \right),$$

  where $L_{k,h} = \sum_{h'=h}^H \ell_k(s_{k,h'}, a_{k,h'})$.

  Estimate the average dilated bonus kernel $\overline{\Lambda}^{\pi_j}_{j,h} \triangleq \frac{1}{|\mathcal{T}'_j|} \sum_{k \in \mathcal{T}'_j} \Lambda^{\pi_j}_{k,h}$ as follows:

$$\widehat{\Lambda}_{j,h} = \widehat{\Sigma}^\dagger_{j,h} \left(\frac{1}{|\mathcal{T}'_j|} \sum_{k \in \mathcal{T}'_j} \left((1 - Y_k) + Y_k H \mathbb{1}[h = h_k] \right) \phi(s_{k,h}, a_{k,h}) D_{k,h} \right),$$

  where $D_{k,h} = \sum_{h'=h+1}^H (1 + \frac{1}{H})^{i-h} b_j(s_{k,h}, a_{k,h})$.

end for

D. Omitted Proofs in Section 5 (Linear MDP Algorithm)

D.1. Pseudocode of the Improved Linear MDP Algorithm

This section briefly discusses the linear MDP algorithm (presented in Algorithm 6). Apart from the new covariance estimation technique introduced in the main text, it is also different from Algorithm 1 in some other aspects due to the distinct nature of linear-Q MDPs and (simulator-free) linear MDPs, listed as follows:

Firstly, as there are no simulators, we cannot calculate the dilated bonus function recursively like Algorithm 4. Fortunately, in linear MDPs, as observed by Luo et al. (2021a), for any k associated with some policy π_k and bonus function b_k, we can write $B_k(s,a)$ defined in Eq. (3) as $B_k(s,a) = b_k(s,a) + \phi(s,a)^\top \Lambda_{k,h}^{\pi_k}$, where ($\nu$ is defined in Definition 2.4)

$$\Lambda_{k,h}^{\pi} \triangleq \left(1+\frac{1}{H}\right)\int_{s_{h+1}} \mathop{\mathbb{E}}_{a_{h+1}\sim\pi(\cdot|s_{h+1})} \left[B_k(s_{h+1},a_{h+1})\right] \nu(s_{h+1}) \mathrm{d}s_{h+1}.$$

Hence, $B_k(s,a)$ is also linear in $\phi(s,a)$. This allows us to estimate the B_k just like $Q_k^{\pi_k}$, as we see in Line 19.

Notice that, as there are no more simulators, we cannot directly "assume" a good covariance estimation like in Algorithm 1 (which ensures Eqs. (7) and (8)). Instead, we should divide the time horizon into several *epochs* and execute (nearly) the same policy during each epoch to ensure a good estimation. See Algorithm 6 for more details.

D.2. Alternative to the Matrix Geometric Resampling Procedure

Proof of Lemma 5.1. By definition, we know the following holds for all $k \in \mathcal{T}_j$:

$$\mathbb{E}[\phi(s_{k,h},a_{k,h})\phi(s_{k,h},a_{k,h})^\top] = \Sigma_h^{\widetilde{\pi}_j}, \quad \phi(s_{k,h},a_{k,h})\phi(s_{k,h},a_{k,h})^\top \preceq I$$

Moreover, each $(s_{k,h},a_{k,h})$ is i.i.d. Thus, we can apply Lemma A.4 with

$$H_i = \frac{1}{2}\left(\gamma I + \phi(s_{k_i,h},a_{k_i,h})\phi(s_{k_i,h},a_{k_i,h})^\top\right), H = \frac{1}{2}\left(\gamma I + \Sigma_h^{\widetilde{\pi}_j}\right), \quad i = 1,2,\ldots,|\mathcal{T}_j|,$$

where k_i stands for the i-th element in $\mathcal{T}_j$. We then have the following according to Lemma A.4:

$$-\sqrt{\frac{d}{|\mathcal{T}_j|}\log\frac{d}{\delta}}H^{1/2} \preceq \frac{1}{|\mathcal{T}_j|}\sum_{i=1}^{|\mathcal{T}_j|} H_i - H \preceq \sqrt{\frac{d}{|\mathcal{T}_j|}\log\frac{d}{\delta}}H^{1/2}, \quad \text{if } \gamma \geq 2\frac{d}{|\mathcal{T}_j|}\log\frac{d}{\delta}. \tag{16}$$

Let the empirical average of all H_i's be $\widehat{H}$, i.e.,

$$\widehat{H} = \frac{1}{|\mathcal{T}_j|}\sum_{i=1}^{|\mathcal{T}_j|} H_i = \frac{1}{2}(\gamma I + \widetilde{\Sigma}_{j,h}).$$

Then we can arrive at the following under the same condition as Eq. (16):

$$I - \sqrt{\frac{d}{|\mathcal{T}_j|}\log\frac{d}{\delta}}H^{-1/2} \preceq \widehat{H}H^{-1} \preceq I + \sqrt{\frac{d}{|\mathcal{T}_j|}\log\frac{d}{\delta}}H^{-1/2}.$$

Moreover, by the definition of H, we know $H^{-1/2} \preceq \sqrt{2}(\gamma I)^{-1/2}$. Setting $W = 4d\log\frac{d}{\delta}\gamma^{-2}$ (which ensures $\gamma \geq 2\frac{d}{|\mathcal{T}_j|}\log\frac{d}{\delta}$ as $|\mathcal{T}_j| = \frac{W}{2}$), the LHS and RHS become $(1-\sqrt{\gamma})I$ and $(1+\sqrt{\gamma})I$, respectively. Hence,

$$(1-\sqrt{\gamma})\tfrac{1}{2}(\gamma I + \Sigma_h^{\widetilde{\pi}_j}) \preceq \tfrac{1}{2}(\gamma I + \widetilde{\Sigma}_{j,h}) \preceq (1+\sqrt{\gamma})\tfrac{1}{2}(\gamma I + \Sigma_h^{\widetilde{\pi}_j}),$$

which gives our conclusion after multiplying 2 on both sides. □

D.3. Proof of Main Theorem

We first state the formal version of Theorem 5.3:

Theorem D.1. *Suppose that* $\alpha = \frac{\delta_e}{6\beta}$, $M_0 \geq \alpha^2 dH^2$, $N_0 \geq 100\frac{M_0^3}{\alpha^2}\log\frac{K}{\delta}$, $W = 4d\log\frac{d}{\delta}\gamma^{-2}$, $\gamma \geq 36\frac{\beta^2}{\delta_e^2}$, *and* $100\eta H^4 \leq \beta$. *Further pick* $\delta = K^{-3}$. *Then Algorithm 6 applied to linear MDPs (Definition 2.4) ensures*

$$\mathcal{R}_K = \widetilde{\mathcal{O}}\left(\frac{\delta_e^6}{\beta^6}d^4H^8 + \delta_e K + \beta dHK + \frac{\gamma}{\beta}dH^3K + \frac{H}{\eta}Ad\gamma^{-2} + Ad^{3/2}H^2\gamma^{-2} + \frac{H^3}{\gamma^2}K\right).$$

With some proper tuning, we can ensure $\mathcal{R}_K = \widetilde{\mathcal{O}}((H^{20}Ad^6)^{1/9}K^{8/9})$.

Proof of Theorem 5.3. The regret decomposition is the same as Theorem 6.1 by Luo et al. (2021a), which we include below. As sketched in the main text, we decompose the episodes into three parts: those executing POLICYCOVER, those using exploratory policies (i.e., the Ber(δ_e) gives 1), and the ones executing π_j.

For the first part, it's trivially bounded by $\mathcal{O}(K_0H)$. For the second part, as we explore with probability δ_e for each episode, the total regret gets bounded by $\mathcal{O}(\delta_e KH)$. For the last part, it suffices to bound the following to apply Lemma 2.5 (where we still consider those exploratory episodes as they only bring extra regret), where $J = \frac{K-K_0}{W}$ denotes the number of epochs for simplicity and $\overline{Q}_j^{\pi}(s,a) = \phi(s,a)^{\mathsf{T}}\overline{\theta}_{j,h}^{\pi}$ denotes the average Q-function in $\mathcal{T}_j'$ (h is such that $s \in \mathcal{S}_h$):

$$\sum_{j=1}^{J}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[W\sum_{a_h\in\mathcal{A}}(\pi_j(a_h\mid s_h) - \pi^*(a_h\mid s_h))(\overline{Q}_j^{\pi_j}(s_h,a_h) - B_j(s_h,a_h))\right].$$

As $B_k(s,a)$ is only bounded for the known states (by definition of $\mathcal{K}$), we first consider the unknown states:

$$\begin{aligned}&\sum_{j=1}^{J}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[W\mathbb{1}[x\notin\mathcal{K}]\sum_{a_h\in\mathcal{A}}(\pi_j(a_h\mid s_h) - \pi^*(a_h\mid s_h))(\overline{Q}_j^{\pi_j}(s,a) - B_j(s_h,a_h))\right]\\ &\leq J\sum_{h=1}^{E}\mathop{\mathbb{E}}_{s_h\sim\pi^*}[W\mathbb{1}[s_h\notin\mathcal{K}]3H] = \widetilde{\mathcal{O}}\left(HK\frac{dH^3}{\alpha}\right)\end{aligned}$$

according to Lemma A.2. For the remaining, we decompose $(\pi_j(a_h\mid s_h) - \pi^*(a_h\mid s_h))(\overline{Q}_j^{\pi_j}(s_h,a_h) - B_j(s_h,a_h))$ into the biases of $\widehat{Q}_j(s_h,a_h)$ and $\widehat{B}_j(s,a_h)$ plus the FTRL regret $(\pi_j(a_h\mid s_h) - \pi^*(a_h\mid s_h))(\widehat{Q}_j(s_h,a_h) - \widehat{B}_j(s_h,a_h))$, i.e.,

$$\begin{aligned}&\sum_{j=1}^{J}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[W\mathbb{1}[s_h\in\mathcal{K}]\sum_{a_h\in\mathcal{A}}(\pi_j(a_h\mid s_h) - \pi^*(a_h\mid s_h))(\overline{Q}_j^{\pi_j}(s_h,a_h) - B_j(s_h,a_h))\right]\\ &= W\underbrace{\sum_{j=1}^{J}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathbb{1}[s_h\in\mathcal{K}]\mathop{\mathbb{E}}_{a_h\sim\pi_j(\cdot|s_h)}\left[\overline{Q}_j^{\pi_j}(s_h,a_h) - \widehat{Q}_j(s_h,a_h))\right]\right]}_{\text{BIAS-1}} +\\ &W\underbrace{\sum_{j=1}^{J}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathbb{1}[s_h\in\mathcal{K}]\mathop{\mathbb{E}}_{a_h\sim\pi^*(\cdot|s_h)}\left[\widehat{Q}_j(s_h,a_h)) - \overline{Q}_j^{\pi_j}(s_h,a_h)\right]\right]}_{\text{BIAS-2}} +\\ &W\underbrace{\sum_{j=1}^{J}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathbb{1}[s_h\in\mathcal{K}]\mathop{\mathbb{E}}_{a_h\sim\pi_j(\cdot|s_h)}\left[\widehat{B}_j(s_h,a_h) - B_j(s_h,a_h)\right]\right]}_{\text{BIAS-3}} +\\ &W\underbrace{\sum_{j=1}^{J}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathbb{1}[s_h\in\mathcal{K}]\mathop{\mathbb{E}}_{a_h\sim\pi^*(\cdot|s_h)}\left[B_j(s_h,a_h) - \widehat{B}_j(s_h,a_h)\right]\right]}_{\text{BIAS-4}} +\end{aligned}$$

$$W \underbrace{\sum_{j=1}^{J}\sum_{h=1}^{H} \mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathbb{1}[s_h\in\mathcal{K}]\sum_{a_h\in\mathcal{A}}(\pi_j(a_h\mid s_h)-\pi^*(a_h\mid s_h))(\widehat{Q}_j(s_h,a_h)-\widehat{B}_j(s_h,a_h))\right]}_{\text{REG-TERM}}.$$

All the bias terms can be bounded similarly, as we will show in Lemmas D.2 and D.3, we can bound them as

$$\begin{aligned}
&\mathbb{E}[\text{BIAS-1}]+\mathbb{E}[\text{BIAS-2}]+\mathbb{E}[\text{BIAS-3}]+\mathbb{E}[\text{BIAS-4}]\\
&\le \frac{\beta}{2}\mathbb{E}\left[\sum_{j=1}^{J}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_j(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{j,h}}]\right]\right]+\\
&\frac{\beta}{2}\mathbb{E}\left[\sum_{j=1}^{J}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\mathop{\mathbb{E}}_{a_h\sim\pi^*(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{j,h}}]\right]\right]+\mathcal{O}\left(\frac{\gamma}{\beta}dH^3J\right).
\end{aligned}$$

Different from Lemmas B.1 and C.2, in that proof, we need to handle the estimation error $\beta^{-1}\|(\widehat{\Sigma}_{j,h}-\Sigma_h^{\pi_j})\overline{\theta}^{\pi_j}_{j,h}(s,a)\|^2_{\widehat{\Sigma}^\dagger_{j,h}}$ by the multiplicative bound Corollary 5.2 instead of the additive one (e.g., Eq. (7)). See Eq. (17) for more details.

For the REG-TERM, we again apply the new FTRL lemma Lemma 3.1, giving the following expression:

$$\begin{aligned}
\mathbb{E}[\text{REG-TERM}] \le{}& \widetilde{\mathcal{O}}\left(\frac{H}{\eta}A+A\sqrt{d}H^2+\frac{H^3}{\gamma^2}\delta J\right)\\
&+2\eta H^3\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\sum_{j=1}^{J}\mathop{\mathbb{E}}_{a_h\sim\pi_j(\cdot|s_h)}\left[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{j,h}}\right]\right]\\
&+\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\sum_{j=1}^{J}\mathop{\mathbb{E}}_{a_h\sim\pi_j(\cdot|s_h)}[b_j(s,a)]\right],
\end{aligned}$$

whose formal proof is in Lemma D.4. In that proof, we need to bound the magnitudes of the bonuses to write $\widehat{B}_j(s_h,a_h)^2\lesssim\frac{1}{H}\widehat{B}_j(s_h,a_h)$. This is done by applying Lemma D.5 later in this section.

By summing up all terms and multiplying W, we can apply Lemma 2.5 by again using $100\eta H^4\le\beta$. Hence, we get the following by using Eq. (13):

$$\mathcal{R}_K\le\mathcal{O}(K_0)+\mathcal{O}(\delta_eK)+\widetilde{\mathcal{O}}\left(\beta dHK+\frac{\gamma}{\beta}dH^3K+\frac{H}{\eta}AW+A\sqrt{d}H^2W+\frac{H^3}{\gamma^2}\delta K\right),$$

where we conditioned on some good event with probability $1-\mathcal{O}(K^{-2}+K\delta)$. Picking $\delta=K^{-3}$, the total regret when the good event does not happen is of order $\mathcal{O}(K^{-2}HK)=o(1)$.

Moreover, by the conditions $\alpha=\frac{\delta_e}{6\beta}$, $M_0\ge\alpha^2dH^2$, and $N_0\ge100\frac{M_0^3}{\alpha^2}\log\frac{K}{\delta}$, we know that $K_0=M_0N_0=\widetilde{\mathcal{O}}(\frac{\delta_e^6}{\beta^6}d^4H^8)$. Meanwhile, by $W=4d\log\frac{d}{\delta}\gamma^{-2}$, we know that $W=\widetilde{\mathcal{O}}(d\gamma^{-2})$. Thus, we have

$$\mathcal{R}_K=\widetilde{\mathcal{O}}\left(\frac{\delta_e^6}{\beta^6}d^4H^8+\delta_eK+\beta dHK+\frac{\gamma}{\beta}dH^3K+\frac{H}{\eta}Ad\gamma^{-2}+Ad^{3/2}H^2\gamma^{-2}\right).$$

The only conditions are then $\gamma\ge36\frac{\beta^2}{\delta_e}$, $100\eta H^4\le\beta$, which allows us to set

$$\delta_e=C(H^{20}Ad^6)^{1/9}K^{-1/9},\beta=\frac{C^2}{36}(H^{13}A^2d^3K^{-2})^{1/9}K^{-2/9},\gamma=\frac{C^3}{36}(H^2A^1)^{1/3}K^{-1/3},\eta=\frac{C^2}{3600}(H^{-23}A^2d^3K^{-2})^{1/9}K^{-2/9},$$

where C is a constant. This ensures $\mathcal{R}_K=\widetilde{\mathcal{O}}((H^{20}Ad^6)^{1/9}K^{8/9})$ (note that only the 2nd, 4th, and 5th term have a $K^{8/9}$ dependency, which means all other terms can be ignored when stating the bound). □

D.4. Bounding the Bias Terms

Lemma D.2. *When $W = 4d \log \frac{d}{\delta} \gamma^{-2}$ and $\gamma < \frac{1}{4}$, the* BIAS-1 *and* BIAS-2 *terms in Algorithm 6 is bounded by*

$$\mathbb{E}[\text{BIAS-1}] + \mathbb{E}[\text{BIAS-2}] \le \frac{\beta}{4} \mathbb{E}\left[\sum_{j=1}^{J}\sum_{h=1}^{H} \mathop{\mathbb{E}}_{s_h \sim \pi^*}\left[\mathop{\mathbb{E}}_{a_h \sim \pi_j(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{j,h}}]\right]\right] + \frac{\beta}{4} \mathbb{E}\left[\sum_{j=1}^{J}\sum_{h=1}^{H} \mathop{\mathbb{E}}_{s_h \sim \pi^*}\left[\mathop{\mathbb{E}}_{a_h \sim \pi^*(\cdot|s_h)}[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}^\dagger_{j,h}}]\right]\right] + \mathcal{O}\left(\frac{\gamma}{\beta} dH^3 K\right).$$

Proof. By direct calculation like Lemma B.1, we get the following for any $j \in [J]$, $h \in [H]$, $s \in \mathcal{S}_h$ and $a \in \mathcal{A}$ (recall that $\widehat{\Sigma}^\dagger_{j,h}$ only depends on the episodes in $\mathcal{T}_j$; again, all expectations are only taken to the randomness in epoch j)

$$\begin{aligned}
&\mathbb{E}\left[\overline{Q}^{\pi_j}_j(s,a) - \widehat{Q}_j(s,a)\right] \\
&= \phi(s,a)^\mathsf{T}\overline{\theta}^{\pi_j}_{j,h} - \phi(s,a)^\mathsf{T}\,\mathbb{E}\left[\widehat{\Sigma}^\dagger_{j,h}\left(\frac{1}{|\mathcal{T}'_j|}\sum_{k\in\mathcal{T}'_j}((1-Y_k) + Y_k H \mathbb{1}[h=h_k])\phi(s_{k,h},a_{k,h})L_{k,h}\right)\right] \\
&= \phi(s,a)^\mathsf{T}\overline{\theta}^{\pi_j}_{j,h} - \phi(s,a)^\mathsf{T}\,\mathbb{E}\left[\widehat{\Sigma}^\dagger_{j,h}\right]\mathbb{E}\left[\left(\frac{1}{|\mathcal{T}'_j|}\sum_{k\in\mathcal{T}'_j}((1-Y_k) + Y_k H \mathbb{1}[h=h_k])\phi(s_{k,h},a_{k,h})\phi(s_{k,h},a_{k,h})^\mathsf{T}\theta^{\pi_j}_{k,h}\right)\right] \\
&= \phi(s,a)^\mathsf{T}\overline{\theta}^{\pi_j}_{j,h} - \phi(s,a)^\mathsf{T}\,\mathbb{E}\left[\widehat{\Sigma}^\dagger_{j,h}\right]\Sigma^{\widetilde{\pi}_j}_h\overline{\theta}^{\pi_j}_{j,h} \\
&= \phi(s,a)^\mathsf{T}(I - \mathbb{E}[\widehat{\Sigma}^\dagger_{j,h}]\Sigma^{\widetilde{\pi}_j}_h)\overline{\theta}^{\pi_j}_{j,h}.
\end{aligned}$$

By using Cauchy-Schwartz inequality, triangle inequality, and the AM-GM inequality, we get the following:

$$\begin{aligned}
\phi(s,a)^\mathsf{T}(I - \widehat{\Sigma}^\dagger_{j,h}\Sigma^{\widetilde{\pi}_j}_h)\overline{\theta}^{\pi_j}_{j,h} &= \phi(s,a)^\mathsf{T}\widehat{\Sigma}^\dagger_{j,h}(\gamma I + \widetilde{\Sigma}_{j,h} - \Sigma^{\widetilde{\pi}_j}_h)\overline{\theta}^{\pi_j}_{j,h} \\
&\le \|\phi(s,a)\|_{\widehat{\Sigma}^\dagger_{j,h}}\left(\left\|(\gamma I)\overline{\theta}^{\pi_j}_{j,h}\right\|_{\widehat{\Sigma}^\dagger_{j,h}} + \left\|(\widetilde{\Sigma}_{j,h} - \Sigma^{\widetilde{\pi}_j}_h)\overline{\theta}^{\pi_j}_{j,h}\right\|_{\widehat{\Sigma}^\dagger_{j,h}}\right) \\
&\le \frac{\beta}{4}\|\phi(s,a)\|^2_{\widehat{\Sigma}^\dagger_{j,h}} + \frac{2}{\beta}\left\|\gamma\overline{\theta}^{\pi_j}_{j,h}(s,a)\right\|^2_{\widehat{\Sigma}^\dagger_{j,h}} + \frac{2}{\beta}\left\|(\widetilde{\Sigma}_{j,h} - \Sigma^{\widetilde{\pi}_j}_h)\overline{\theta}^{\pi_j}_{j,h}\right\|^2_{\widehat{\Sigma}^\dagger_{j,h}}.
\end{aligned}$$

The first term is the usual bonus term in Lemma 2.5, while the second term easily translates to the following using the assumption that $\|\theta^{\pi_j}_{k,h}\| \le \sqrt{d}H$ for all $k \in \mathcal{T}'_j$:

$$\frac{2}{\beta}\left\|\gamma\overline{\theta}^{\pi_j}_{j,h}\right\|^2_{\widehat{\Sigma}^\dagger_{j,h}} \le \frac{2\gamma^2}{\beta}\left\|(\gamma I + \widetilde{\Sigma}_{j,h})^{-1}\right\|_2 dH^2 \le 2\frac{\gamma}{\beta}dH^2,$$

while the last term translates to the following by algebraic manipulations:

$$\begin{aligned}
\frac{2}{\beta}\left\|(\widetilde{\Sigma}_{j,h} - \Sigma^{\widetilde{\pi}_j}_h)\overline{\theta}^{\pi_j}_{j,h}\right\|^2_{\widehat{\Sigma}^\dagger_{j,h}} &= \frac{2}{\beta}\left\|(\widetilde{\Sigma}_{j,h} - \Sigma^{\widetilde{\pi}_j}_h)\widehat{\Sigma}^\dagger_{j,h}(\widetilde{\Sigma}_{j,h} - \Sigma^{\widetilde{\pi}_j}_h)\right\|_2 dH^2 \\
&= \frac{2}{\beta}\left\|(\widehat{\Sigma}^\dagger_{j,h})^{-1/2}\left(I - (\widehat{\Sigma}^\dagger_{j,h})^{1/2}(\gamma I + \Sigma^{\widetilde{\pi}_j}_h)(\widehat{\Sigma}^\dagger_{j,h})^{1/2}\right)^2(\widehat{\Sigma}^\dagger_{j,h})^{-1/2}\right\|_2 dH^2, \qquad (17)
\end{aligned}$$

where one may expand and check the last step indeed holds.

Using Corollary 5.2, the squared-matrix in the middle has its operator norm bounded by 4γ with high probability (if the good event does not hold, then one can directly bound the last term by matrix Azuma and the operator norm of $\widehat{\Sigma}^\dagger_{k,h}$, giving γ^{-3}; as this only happens with probability K^{-3}, this part contributes $o(K^{-2})$ to the total regret and we thus omit it). Meanwhile, the first and last term both has their operator norms bounded by $\sqrt{2}$. Thus,

$$\frac{2}{\beta}\left\|(\widetilde{\Sigma}_{j,h} - \Sigma^{\pi_j}_h)\overline{\theta}^{\pi_j}_{j,h}\right\|^2_{\widehat{\Sigma}^\dagger_{j,h}} dH^2 \le 16\frac{\gamma}{\beta}dH^2.$$

In other words, the last term gets absorbed by the second term up to constants. Hence, the conclusion follows by the same argument as Lemma B.1. □

Lemma D.3. *When $W = 4d \log \frac{d}{\delta} \gamma^{-2}$ and $\gamma < \frac{1}{4}$, the* BIAS-3 *and* BIAS-4 *terms in Algorithm 6 is bounded by the following when the good event in Lemma D.5 holds:*

$$\mathbb{E}[\text{BIAS-3}] + \mathbb{E}[\text{BIAS-4}] \leq \frac{\beta}{4} \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H} \mathop{\mathbb{E}}_{s_h \sim \pi^*}\left[\mathop{\mathbb{E}}_{a_h \sim \pi_k(\cdot|s_h)}[\|\phi(s_h, a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}]\right]\right] + \frac{\beta}{4} \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H} \mathop{\mathbb{E}}_{s_h \sim \pi^*}\left[\mathop{\mathbb{E}}_{a_h \sim \pi^*(\cdot|s_h)}[\|\phi(s_h, a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{k,h}}]\right]\right] + \mathcal{O}\left(\frac{\gamma}{\beta} dH^3 K\right).$$

Proof. The proof is identical to Lemma D.2 except for $L_{k,h}$ is replaced by $D_{k,h}$ and $\overline{\theta}^{\pi_j}_{j,h}$ is replaced by $\Lambda^{\pi_j}_{j,h}$. As we also have $b_j(s,a) \in [0,1]$ by Lemma D.5, the same bound holds. □

D.5. Bounding REG-TERM

Lemma D.4. *Assuming $100H\eta H^4 \leq \beta$ and the good events in Lemma D.5 hold. Then the* REG-TERM *has its expectation bounded by*

$$\mathbb{E}[\text{REG-TERM}] \leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta}A + A\sqrt{d}H^2 + \frac{H^3}{\gamma^2}\delta J\right) + 2\eta H^3 \sum_{h=1}^{H} \mathop{\mathbb{E}}_{s_h \sim \pi^*}\left[\sum_{j=1}^{J} \mathop{\mathbb{E}}_{a_h \sim \pi_j(\cdot|s_h)}\left[\|\phi(s_h, a_h)\|^2_{\widehat{\Sigma}^{\dagger}_{j,h}}\right]\right] + \sum_{h=1}^{H} \mathop{\mathbb{E}}_{s_h \sim \pi^*}\left[\sum_{j=1}^{J} \mathop{\mathbb{E}}_{a_h \sim \pi_j(\cdot|s_h)}[b_j(s,a)]\right].$$

Proof. The proof generally follows from Lemma B.2, except for some tiny differences due to epoching.

Using Lemma 3.1, we get the following for all $j \in [J]$, $s \in \mathcal{S}_h \cap \mathcal{K}$ (where $h \in [H]$), and $\widetilde{\pi}^* \in \Pi$:

$$\begin{aligned}
&\sum_{j=1}^{J}\sum_{a\in\mathcal{A}}(\pi_j(a \mid s) - \pi^*(a \mid s))(\widehat{Q}_j(s,a) - \widehat{B}_j(s,a)) \\
\leq\ & \frac{\Psi(\widetilde{\pi}^*(\cdot \mid s)) - \Psi(\pi_1(\cdot \mid s))}{\eta} + \\
&\sum_{j=1}^{J}\sum_{a\in\mathcal{A}}(\widetilde{\pi}^*(a \mid s) - \pi^*(a \mid s))(\widehat{Q}_j(s,a) - \widehat{B}_j(s,a)) + \\
&\eta \sum_{j=1}^{J}\sum_{a\in\mathcal{A}} \pi_j(a \mid s)(\widehat{Q}_j(s,a) - \widehat{B}_j(s,a))^2.
\end{aligned}$$

By picking $\widetilde{\pi}^*(a \mid s) = (1 - AJ^{-1})\pi^*(a \mid s) + J^{-1}$ as Lemma B.2, the first term is bounded by $\eta^{-1} A \ln J$. Meanwhile, the second term is bounded by

$$\begin{aligned}
&\sum_{j=1}^{J}\sum_{a\in\mathcal{A}}(\widetilde{\pi}^*(a \mid s) - \pi^*(a \mid s))(\widehat{Q}_j(s,a) - \widehat{B}_j(s,a)) \\
=\ & \sum_{j=1}^{J}\sum_{a\in\mathcal{A}}(-AJ^{-1}\pi^*(a \mid s) + J^{-1})(\widehat{Q}_j(s,a) - \widehat{B}_j(s,a)).
\end{aligned}$$

Like what we did in Lemma B.2, we consider the expected difference between $\widehat{Q}_j$ and $\overline{Q}_j^{\pi_j}$:

$$\mathbb{E}[\widehat{Q}_j(s,a) - \overline{Q}_j^{\pi_j}(s,a)] = \phi(s,a)^\mathsf{T}(I - \mathbb{E}[\widehat{\Sigma}_{j,h}^\dagger \Sigma_h^{\widetilde{\pi}_j}])\overline{\theta}_{j,h}^{\pi_j} \le 2\sqrt{d}H,$$

where the first inequality is due to Lemma D.2 and the second one uses Lemma 5.1. Moreover, according to Lemma D.5, the same bound also holds for $\mathbb{E}[\widehat{B}_j(s,a) - \overline{B}_j^{\pi_j}(s,a)]$.

Hence, after taking expectations on both sides, we know that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{J}\sum_{a\in\mathcal{A}}(\widetilde{\pi}^*(a \mid s) - \pi^*(a \mid s))(\widehat{Q}_j(s,a) - \widehat{B}_j(s,a))\right] \\
&\le \sum_{j=1}^{J}\sum_{a\in\mathcal{A}}(|-AJ^{-1}\pi^*(a \mid s)| + |J^{-1}|)|\mathbb{E}[\widehat{Q}_j(s,a) - \widehat{B}_j(s,a)]| \\
&\le J \times 2AJ^{-1} \times 4\sqrt{d}H = \mathcal{O}\left(A\sqrt{d}H\right).
\end{aligned}$$

Then consider the last term. We still write $(\widehat{Q}_j(s,a) - \widehat{B}_j(s,a))^2 \le 2\widehat{Q}_j(s,a)^2 + 2\widehat{B}_j(s,a)^2$, which can be calculated as follows:

$$\begin{aligned}
\mathbb{E}[\widehat{Q}_j(s,a)^2] &\le H^2\,\mathbb{E}\left[\frac{1}{|\mathcal{T}_j'|}\sum_{k\in\mathcal{T}_j'}\phi(s,a)^\mathsf{T}\widehat{\Sigma}_{j,h}^\dagger\left(((1-Y_k) + Y_k H\mathbb{1}[h=h_k])^2\phi(s_{k,h},a_{k,h})\phi(s_{k,h},a_{k,h})^\mathsf{T}\right)\widehat{\Sigma}_{j,h}^\dagger\phi(s,a)\right] \\
&= H^2\,\mathbb{E}\left[\phi(s,a)^\mathsf{T}\widehat{\Sigma}_{j,h}^\dagger\left((1-\delta_e)\Sigma_h^{\pi_j} + H\delta_e\Sigma_h^{\mathrm{cov}}\right)\widehat{\Sigma}_{j,h}^\dagger\phi(s,a)\right] \\
&\le H^3\,\mathbb{E}\left[\phi(s,a)^\mathsf{T}\widehat{\Sigma}_{j,h}^\dagger\Sigma_h^{\widetilde{\pi}_j}\widehat{\Sigma}_{j,h}^\dagger\phi(s,a)\right].
\end{aligned}$$

Then we use Corollary 5.2. If the good event does not happen, then this term is bounded by $\mathcal{O}(H^3\gamma^{-2}\delta)$. Otherwise, it can be written as $2H^3\,\mathbb{E}\left[\|\phi(s,a)\|^2_{\widehat{\Sigma}_{j,h}^\dagger}\right]$. Then we consider the estimated dilated bonus term:

$$\mathbb{E}[\widehat{B}_j(s,a)^2] \le 2\,\mathbb{E}[b_j(s,a)^2] + 2\,\mathbb{E}[(\phi(s,a)^\mathsf{T}\widehat{\Lambda}_{j,h})^2].$$

According to Lemma D.5 (whose failure only contributes in total $o(1)$ regret as we pick $\delta = K^{-3}$), we know $b_j(s,a) \le 1$, which means $D_{k,h} \le 3H$. Thus, the second term is also bounded by $\mathcal{O}(H^3\,\mathbb{E}[\|\phi(s,a)\|^2_{\widehat{\Sigma}_{j,h}^\dagger}] + H^3\gamma^{-2}\delta)$. Moreover, as $b_j(s,a) \le 1$, the first term is bounded by $\mathbb{E}[b_j(s,a)]$. Thus, by definition of $b_j(s,a)$, we get

$$\mathbb{E}[\widehat{B}_j(s,a)^2] \le \mathcal{O}\left(\frac{H^3}{\beta}\right)b_j(s,a) + \mathcal{O}(H^3\gamma^{-2}\delta)).$$

Putting everything together gives

$$\begin{aligned}
\mathbb{E}[\textsc{Reg-Term}] \le{}& \widetilde{\mathcal{O}}\left(\frac{H}{\eta}A + A\sqrt{d}H^2 + \frac{H^3}{\gamma^2}\delta J\right) \\
&+ 2\eta H^3\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\sum_{j=1}^{J}\mathop{\mathbb{E}}_{a_h\sim\pi_j(\cdot|s_h)}\left[\|\phi(s_h,a_h)\|^2_{\widehat{\Sigma}_{j,h}^\dagger}\right]\right] \\
&+ \frac{100\eta H^3}{\beta}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim\pi^*}\left[\sum_{j=1}^{J}\mathop{\mathbb{E}}_{a_h\sim\pi_j(\cdot|s_h)}[b_j(s,a)]\right].
\end{aligned}$$

This then translates to our conclusion using the condition that $100H^4\eta \le \beta$. □

D.6. Bounding the Magnitudes of Bonuses

Lemma D.5. *Let* $\alpha = \frac{\delta_e}{6\beta}$, $M_0 \geq \alpha^2 dH^2$, $N_0 \geq 100\frac{M_0^3}{\alpha^2}\log\frac{K}{\delta}$, *and* $\gamma \geq 36\frac{\beta^2}{\delta_e}$. *Then with probability* $1-(K^{-2}+K\delta)$, $b_j(s,a) = \mathbb{1}[(s,a)\in\mathcal{K}] \times \beta(\|\phi(s,a)\|^2_{\widehat{\Sigma}^\dagger_{j,h}} + \mathbb{E}_{a'\sim\pi_j(\cdot|s)}[\|\phi(s,a')\|_{\widehat{\Sigma}^\dagger_{j,h}}]) \leq 1$ *for all* $j\in[J]$, $h\in[H]$, $s\in\mathcal{S}_h$ *and* $a\in\mathcal{A}$.

The proof is similar to, but different from Lemma F.1 of Luo et al. (2021a). Here, we use the alternative for the MGR procedure. We also refine the analysis on $M_0^{-1}+\Sigma_h^{\text{cov}} \approx \widehat{\Sigma}_h^{\text{cov}}$ for a smaller N_0, which is critical for our new regret bound.

Proof. It suffices to show that $\beta\|\phi(s,a')\|^2_{\widehat{\Sigma}^\dagger_{k,h}} \leq \frac{1}{2}$ for any $s\in\mathcal{K}$ and $a'\in\mathcal{A}$. Firstly, we have the following with high probability because $\widehat{\Sigma}^\dagger_{j,h}$ is a multiplicative approximation (i.e., Corollary 5.2):

$$\beta\|\phi(s,a')\|_{\widehat{\Sigma}^\dagger_{j,h}} \leq 2\beta\|\phi(s,a')\|^2_{(\gamma I+\Sigma_h^{\tilde{\pi}_j})^{-1}} \leq 2\frac{\beta}{\delta_e}\|\phi(s,a')\|^2_{(\frac{\gamma}{\delta_e}I+\Sigma_h^{\text{cov}})^{-1}} \leq 2\frac{\beta}{\delta_e}\|\phi(s,a')\|^2_{(M_0^{-1}I+\Sigma_h^{\text{cov}})^{-1}},$$

where Σ_h^{cov} is the short-hand notation of $\Sigma_h^{\pi_{\text{cov}}}$ and the last step uses the fact that $\frac{\gamma}{\delta_e}M_0 \geq \frac{\gamma}{\delta_e}\frac{\delta_e^2}{36\beta^2} \geq 1$. Moreover, we argue that $M_0^{-1}+\Sigma_h^{\text{cov}} \approx \widehat{\Sigma}_h^{\text{cov}}$. By definition of $\widehat{\Sigma}_h^{\text{cov}}$ (see Algorithm 5):

$$\widehat{\Sigma}_h^{\text{cov}} = \frac{1}{M_0}I + \frac{1}{M_0N_0}\sum_{m=1}^{M_0}\sum_{n=1}^{N_0}\phi(s_{m,n,h},a_{m,n,h})\phi(s_{m,n,h},a_{m,n,h})^\top,$$

we can apply the Matrix Azuma inequality (Lemma A.3) to ensure that (where we simply pick $X_{mN_0+n} = \phi(s_{m,n,h},a_{m,n,h})\phi(s_{m,n,a},a_{m,n,h})^\top - \Sigma_h^{\pi_m}$ and $A_{mN_0+n} = I \succeq X_{mN_0+n}$; there are in total M_0N_0 matrices):

$$\left\|\widehat{\Sigma}_h^{\text{cov}} - \frac{1}{M_0}I - \Sigma_h^{\text{cov}}\right\|_2 \leq \sqrt{\frac{8}{M_0N_0}\log\frac{d}{\delta}}.$$

The rest follows the proof of the original lemma (Luo et al., 2021a, Lemma F.1). By following the proof of Theorem 2.1 of Meng & Zheng (2010), we can conclude that

$$\left\|\left(\widehat{\Sigma}_h^{\text{cov}}\right)^{-1} - \left(\frac{1}{M_0}I - \Sigma_h^{\text{cov}}\right)^{-1}\right\|_2 \leq M_0^2\left\|\widehat{\Sigma}_h^{\text{cov}} - \frac{1}{M_0}I - \Sigma_h^{\text{cov}}\right\|_2 \leq M_0^2\sqrt{\frac{8}{M_0N_0}\log\frac{d}{\delta}} = \sqrt{\frac{8M_0^3}{N_0}\log\frac{d}{\delta}}.$$

Therefore, we only need $M_0 = \mathcal{O}(N_0^3)$ to make it bounded by $\frac{\alpha}{2}$. Consequently, as $\|\phi(s,a')\|_2 \leq 1$,

$$\begin{aligned}
&\beta\|\phi(s,a')\|_{\widehat{\Sigma}^\dagger_{j,h}} \leq 2\frac{\beta}{\delta_e}\|\phi(s,a')\|_{(M_0^{-1}I+\Sigma_h^{\text{cov}})^{-1}} \\
&\leq 2\frac{\beta}{\delta_e}\left(\|\phi(s,a')\|^2_{(\widehat{\Sigma}_h^{\text{cov}})^{-1}} + \left\|\left(\widehat{\Sigma}_h^{\text{cov}}\right)^{-1} - \left(\frac{1}{M_0}I - \Sigma_h^{\text{cov}}\right)^{-1}\right\|_2\right) \\
&\leq 2\frac{\beta}{\delta_e}\left(\alpha + \frac{\alpha}{2}\right) \leq \frac{1}{2},
\end{aligned}$$

by the condition that $s\in\mathcal{K}$ and our choice of α. □