Policy Optimization in Adversarial MDPs: Improved Exploration via Dilated Bonuses

Haipeng Luo* **Chen-Yu Wei*** **Chung-Wei Lee**
{haipengl,chenyu.wei,leechung}@usc.edu
University of Southern California

Abstract

Policy optimization is a widely-used method in reinforcement learning. Due to its local-search nature, however, theoretical guarantees on global optimality often rely on extra assumptions on the Markov Decision Processes (MDPs) that bypass the challenge of global exploration. To eliminate the need of such assumptions, in this work, we develop a general solution that adds *dilated bonuses* to the policy update to facilitate global exploration. To showcase the power and generality of this technique, we apply it to several episodic MDP settings with adversarial losses and bandit feedback, improving and generalizing the state-of-the-art. Specifically, in the tabular case, we obtain $\widetilde{\mathcal{O}}(\sqrt{T})$ regret where T is the number of episodes, improving the $\widetilde{\mathcal{O}}(T^{2/3})$ regret bound by [27]. When the number of states is infinite, under the assumption that the state-action values are linear in some low-dimensional features, we obtain $\widetilde{\mathcal{O}}(T^{2/3})$ regret with the help of a simulator, matching the result of [24] while importantly removing the need of an exploratory policy that their algorithm requires. To our knowledge, this is the first algorithm with sublinear regret for linear function approximation with adversarial losses, bandit feedback, and no exploratory assumptions. Finally, we also discuss how to further improve the regret or remove the need of a simulator using dilated bonuses, when an exploratory policy is available.[1]

1 Introduction

Policy optimization methods are among the most widely-used methods in reinforcement learning. Its empirical success has been demonstrated in various domains such as computer games [26] and robotics [21]. However, due to its local-search nature, global optimality guarantees of policy optimization often rely on unrealistic assumptions to ensure global exploration (see e.g., [1, 3, 24, 30]), making it theoretically less appealing compared to other methods.

Motivated by this issue, a line of recent works [7, 27, 2, 35] equip policy optimization with global exploration by adding exploration bonuses to the update, and prove favorable guarantees even without making extra exploratory assumptions. Moreover, they all demonstrate some robustness aspect of policy optimization (such as being able to handle adversarial losses or a certain degree of model mis-specification). Despite these important progresses, however, many limitations still exist, including worse regret rates comparing to the best value-based or model-based approaches [27, 2, 35], or requiring full-information feedback on the entire loss function (as opposed to the more realistic bandit feedback) [7].

*Equal contribution.

[1] In an improved version of this paper, we show that under the linear MDP assumption, an exploratory policy is not even needed. See https://arxiv.org/abs/2107.08346.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

To address these issues, in this work, we propose a new type of exploration bonuses called *dilated bonuses*, which satisfies a certain *dilated Bellman equation* and provably leads to improved exploration compared to existing works (Section 3). We apply this general idea to advance the state-of-the-art of policy optimization for learning finite-horizon episodic MDPs with *adversarial losses and bandit feedback*. More specifically, our main results are:

- First, in the tabular setting, addressing the main open question left in [27], we improve their $\widetilde{\mathcal{O}}(T^{2/3})$ regret to the optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. This shows that policy optimization, which performs local optimization, is as capable as other occupancy-measure-based global optimization algorithms [15, 20] in terms of global exploration. Moreover, our algorithm is computationally more efficient than those global methods since they require solving some convex optimization in each episode. (Section 4)
- Second, to further deal with large-scale problems, we consider a linear function approximation setting where the state-action values are linear in some known low-dimensional features and also a simulator is available, the same setting considered by [24]. We obtain the same $\widetilde{\mathcal{O}}(T^{2/3})$ regret while importantly removing the need of an exploratory policy that their algorithm requires. Unlike the tabular setting (where we improve existing regret rates of policy optimization), note that researchers have not been able to show *any* sublinear regret for policy optimization without exploratory assumptions for this problem, which shows the critical role of our proposed dilated bonuses. In fact, there are simply no existing algorithms with sublinear regret *at all* for this setting, be it policy-optimization-type or not. This shows the advantage of policy optimization over other approaches, when combined with our dilated bonuses. (Section 5)
- Finally, while the main focus of our work is to show how dilated bonuses are able to provide global exploration, we also discuss their roles in improving the regret rate to $\widetilde{\mathcal{O}}(\sqrt{T})$ in the linear setting above or removing the need of a simulator for the special case of linear MDPs (with $\widetilde{\mathcal{O}}(T^{6/7})$ regret), when an exploratory policy is available. (Section 6)

Related work. In the tabular setting, except for [27], most algorithms apply the occupancy-measure-based framework to handle adversarial losses (e.g., [25, 15, 9, 8]), which as mentioned is computationally expensive. For stochastic losses, there are many more different approaches such as model-based ones [13, 10, 5, 12, 34] and value-based ones [14, 11].

Theoretical studies for linear function approximation have gained increasing interest recently [32, 33, 16]. Most of them study stochastic/stationary losses, with the exception of [24, 7]. Our algorithm for the linear MDP setting bears some similarity to those of [2, 35] which consider stationary losses. However, our algorithm and analysis are arguably simpler than theirs. Specifically, they divide the state space into a known part and an unknown part, with different exploration principle and bonus design for different parts. In contrast, we enjoy a unified bonus design for all states. Besides, in each episode, their algorithms first execute an exploratory policy (from a *policy cover*), and then switch to the policy suggested by the policy optimization algorithm, which inevitably leads to linear regret when facing adversarial losses.

2 Problem Setting

We consider an MDP specified by a state space X (possibly infinite), a finite action space A, and a transition function P with $P(\cdot|x,a)$ specifying the distribution of the next state after taking action a in state x. In particular, we focus on the *finite-horizon episodic setting* in which X admits a layer structure and can be partitioned into $X_0, X_1, \ldots, X_H$ for some fixed parameter H, where X_0 contains only the initial state x_0, X_H contains only the terminal state x_H, and for any $x \in X_h$, $h = 0, \ldots, H-1$, $P(\cdot|x,a)$ is supported on X_{h+1} for all $a \in A$ (that is, transition is only possible from X_h to X_{h+1}). An episode refers to a trajectory that starts from x_0 and ends at x_H following some series of actions and the transition dynamic. The MDP may be assigned with a loss function $\ell : X \times A \to [0,1]$ so that $\ell(x,a)$ specifies the loss suffered when selecting action a in state x.

A policy π for the MDP is a mapping $X \to \Delta(A)$, where $\Delta(A)$ denotes the set of distributions over A and $\pi(a|x)$ is the probability of choosing action a in state x. Given a loss function ℓ

and a policy π, the expected total loss of π is given by $V^\pi(x_0;\ell) = \mathbb{E}\big[\sum_{h=0}^{H-1}\ell(x_h,a_h) \,\big|\, a_h \sim \pi_t(\cdot|x_h), x_{h+1}\sim P(\cdot|x_h,a_h)\big]$. It can also be defined via the Bellman equation involving the *state value function* $V^\pi(x;\ell)$ and the *state-action value function* $Q^\pi(x,a;\ell)$ (a.k.a. Q-function) defined as below: $V(x_H;\ell)=0$,

$$Q^\pi(x,a;\ell) = \ell(x,a) + \mathbb{E}_{x'\sim P(\cdot|x,a)}\left[V^\pi(x';\ell)\right], \text{ and } V^\pi(x;\ell) = \mathbb{E}_{a\sim\pi(\cdot|x)}\left[Q^\pi(x,a;\ell)\right].$$

We study online learning in such a finite-horizon MDP with *unknown transition*, *bandit feedback*, and *adversarial losses*. The learning proceeds through T episodes. Ahead of time, an adversary arbitrarily decides T loss functions $\ell_1,\ldots,\ell_T$, without revealing them to the learner. Then in each episode t, the learner decides a policy π_t based on all information received prior to this episode, executes π_t starting from the initial state x_0, generates and observes a trajectory $\{(x_{t,h},a_{t,h},\ell_t(x_{t,h},a_{t,h}))\}_{h=0}^{H-1}$. Importantly, the learner does not observe any other information about ℓ_t (a.k.a. bandit feedback).[2] The goal of the learner is to minimize the regret, defined as

$$\text{Reg} = \sum_{t=1}^T V_t^{\pi_t}(x_0) - \min_\pi \sum_{t=1}^T V_t^\pi(x_0),$$

where we use $V_t^\pi(x)$ as a shorthand for $V^\pi(x;\ell_t)$ (and similarly $Q_t^\pi(x,a)$ as a shorthand for $Q^\pi(x,a;\ell_t)$). Without further structures, the best existing regret bound is $\widetilde{\mathcal{O}}(H|X|\sqrt{|A|T})$ [15], with an extra $\sqrt{X}$ factor compared to the best existing lower bound [14].

Occupancy measures. For a policy π and a state x, we define $q^\pi(x)$ to be the probability (or probability measure when $|X|$ is infinite) of visiting state x within an episode when following π. When it is necessary to highlight the dependence on the transition, we write it as $q^{P,\pi}(x)$. Further define $q^\pi(x,a)=q^\pi(x)\pi(a|x)$ and $q_t(x,a)=q^{\pi_t}(x,a)$. Finally, we use $q^\star$ as a shorthand for $q^{\pi^\star}$ where $\pi^\star\in\operatorname{argmin}_\pi\sum_{t=1}^T V_t^\pi(x_0)$ is one of the optimal policies.

Note that by definition, we have $V^\pi(x_0;\ell)=\sum_{x,a}q^\pi(x,a)\ell(x,a)$. In fact, we will overload the notation and let $V^\pi(x_0;b)=\sum_{x,a}q^\pi(x,a)b(x,a)$ for any function $b: X\times A\to\mathbb{R}$ (even though it might not correspond to a real loss function).

Other notations. We denote by $\mathbb{E}_t[\cdot]$ and $\text{Var}_t[\cdot]$ the expectation and variance conditioned on everything prior to episode t. For a matrix Σ and a vector z (of appropriate dimension), $\|z\|_\Sigma$ denotes the quadratic norm $\sqrt{z^\top\Sigma z}$. The notation $\widetilde{\mathcal{O}}(\cdot)$ hides all logarithmic factors.

3 Dilated Exploration Bonuses

In this section, we start with a general discussion on designing exploration bonuses (not specific to policy optimization), and then introduce our new dilated bonuses for policy optimization. For simplicity, the exposition in this section assumes a finite state space, but the idea generalizes to an infinite state space.

When analyzing the regret of an algorithm, very often we run into the following form:

$$\text{Reg} = \sum_{t=1}^T V_t^{\pi_t}(x_0) - \sum_{t=1}^T V_t^{\pi^\star}(x_0) \le o(T) + \sum_{t=1}^T\sum_{x,a} q^\star(x,a)b_t(x,a) = o(T) + \sum_{t=1}^T V^{\pi^\star}(x_0;b_t), \tag{1}$$

for some function $b_t(x,a)$ usually related to some estimation error or variance that can be prohibitively large. For example, in policy optimization, the algorithm performs local search in each state essentially using a multi-armed bandit algorithm and treating $Q^{\pi_t}(x,a)$ as the loss of action a in state x. Since $Q^{\pi_t}(x,a)$ is unknown, however, the algorithm has to use some estimator of $Q^{\pi_t}(x,a)$ instead, whose bias and variance both contribute to the b_t function. Usually, $b_t(x,a)$ is large for a rarely-visited state-action pair (x,a) and is inversely related to $q_t(x,a)$, which is exactly why most analysis relies

[2] Full-information feedback, on the other hand, refers to the easier setting where the entire loss function ℓ_t is revealed to the learner at the end of episode t.

on the assumption that some *distribution mismatch coefficient* related to ${q^\star(x,a)}/{q_t(x,a)}$ is bounded (see e.g., [3, 31]).

On the other hand, an important observation is that while $V^{\pi^\star}(x_0; b_t)$ can be prohibitively large, its counterpart with respect to the learner's policy $V^{\pi_t}(x_0; b_t)$ is usually nicely bounded. For example, if $b_t(x,a)$ is inversely related to $q_t(x,a)$ as mentioned, then $V^{\pi_t}(x_0; b_t) = \sum_{x,a} q_t(x,a) b_t(x,a)$ is small no matter how small $q_t(x,a)$ could be for some (x,a). This observation, together with the linearity property $V^\pi(x_0; \ell_t - b_t) = V^\pi(x_0; \ell_t) - V^\pi(x_0; b_t)$, suggests that we treat $\ell_t - b_t$ as the loss function of the problem, or in other words, add a (negative) bonus to each state-action pair, which intuitively encourages exploration due to underestimation. Indeed, assuming for a moment that Eq. (1) still roughly holds even if we treat $\ell_t - b_t$ as the loss function:

$$\sum_{t=1}^T V^{\pi_t}(x_0; \ell_t - b_t) - \sum_{t=1}^T V^{\pi^\star}(x_0; \ell_t - b_t) \lesssim o(T) + \sum_{t=1}^T V^{\pi^\star}(x_0; b_t). \tag{2}$$

Then by linearity and rearranging, we have

$$\text{Reg} = \sum_{t=1}^T V_t^{\pi_t}(x_0) - \sum_{t=1}^T V_t^{\pi^\star}(x_0) \lesssim o(T) + \sum_{t=1}^T V^{\pi_t}(x_0; b_t). \tag{3}$$

Due to the switch from $\pi^\star$ to π_t in the last term compared to Eq. (1), this is usually enough to prove a desirable regret bound without making extra assumptions.

The caveat of this discussion is the assumption of Eq. (2). Indeed, after adding the bonuses, which itself contributes some more bias and variance, one should expect that b_t on the right-hand side of Eq. (2) becomes something larger, breaking the desired cancellation effect to achieve Eq. (3). Indeed, the definition of b_t essentially becomes circular in this sense.

Dilated Bonuses for Policy Optimization To address this issue, we take a closer look at the policy optimization algorithm specifically. As mentioned, policy optimization decomposes the problem into individual multi-armed bandit problems in each state and then performs local optimization. This is based on the well-known performance difference lemma [17]:

$$\text{Reg} = \sum_x q^\star(x) \sum_{t=1}^T \sum_a \Big(\pi_t(a|x) - \pi^\star(a|x)\Big) Q_t^{\pi_t}(x,a),$$

showing that in each state x, the learner is facing a bandit problem with $Q_t^{\pi_t}(x,a)$ being the loss for action a. Correspondingly, incorporating the bonuses b_t for policy optimization means subtracting the bonus $Q^{\pi_t}(x,a;b_t)$ from $Q_t^{\pi_t}(x,a)$ for each action a in each state x. Recall that $Q^{\pi_t}(x,a;b_t)$ satisfies the Bellman equation $Q^{\pi_t}(x,a;b_t) = b_t(x,a) + \mathbb{E}_{x'\sim P(\cdot|x,a)}\mathbb{E}_{a'\sim\pi_t(\cdot|x')}\left[B_t(x',a')\right]$. To resolve the issue mentioned earlier, we propose to replace this bonus function $Q^{\pi_t}(x,a;b_t)$ with its *dilated* version $B_t(s,a)$ satisfying the following *dilated Bellman equation*:

$$B_t(x,a) = b_t(x,a) + \left(1 + \frac{1}{H}\right)\mathbb{E}_{x'\sim P(\cdot|x,a)}\mathbb{E}_{a'\sim\pi_t(\cdot|x')}\left[B_t(x',a')\right] \tag{4}$$

(with $B_t(x_H, a) = 0$ for all a). The only difference compared to the standard Bellman equation is the extra $(1 + \frac{1}{H})$ factor, which slightly increases the weight for deeper layers and thus intuitively induces more exploration for those layers. Due to the extra bonus compared to $Q^{\pi_t}(x,a;b_t)$, the regret bound also increases accordingly. In all our applications, this extra amount of regret turns out to be of the form $\frac{1}{H}\sum_{t=1}^T \sum_{x,a} q^\star(x)\pi_t(a|x)B_t(x,a)$, leading to

$$\begin{aligned}
&\sum_x q^\star(x) \sum_{t=1}^T \sum_a \Big(\pi_t(a|x) - \pi^\star(a|x)\Big)\Big(Q_t^{\pi_t}(x,a) - B_t(x,a)\Big) \\
&\qquad \le o(T) + \sum_{t=1}^T V^{\pi^\star}(x_0; b_t) + \frac{1}{H}\sum_{t=1}^T \sum_{x,a} q^\star(x)\pi_t(a|x)B_t(x,a).
\end{aligned} \tag{5}$$

With some direct calculation, one can show that this is enough to show a regret bound that is only a constant factor larger than the desired bound in Eq. (3)! This is summarized in the following lemma.

Lemma 3.1. *If Eq. (5) holds with B_t defined in Eq. (4), then* $\text{Reg} \le o(T) + 3\sum_{t=1}^{T} V^{\pi_t}(x_0; b_t)$.

The high-level idea of the proof is to show that the bonuses added to a layer h is enough to cancel the large bias/variance term (including those coming from the bonus itself) from layer $h+1$. Therefore, cancellation happens in a layer-by-layer manner except for layer 0, where the total amount of bonus can be shown to be at most $(1+\frac{1}{H})^H \sum_{t=1}^{T} V^{\pi_t}(x_0; b_t) \le 3\sum_{t=1}^{T} V^{\pi_t}(x_0; b_t)$.

Recalling again that $V^{\pi_t}(x_0; b_t)$ is usually nicely bounded, we thus arrive at a favorable regret guarantee without making extra assumptions. Of course, since the transition is unknown, we cannot compute B_t exactly. However, Lemma 3.1 is robust enough to handle either a good approximate version of B_t (see Lemma B.1) or a version where Eq. (4) and Eq. (5) only hold in expectation (see Lemma B.2), which is enough for us to handle unknown transition. In the next three sections, we apply this general idea to different settings, showing what b_t and B_t are concretely in each case.

4 The Tabular Case

In this section, we study the tabular case where the number of states is finite. We propose a policy optimization algorithm with $\widetilde{\mathcal{O}}(\sqrt{T})$ regret, improving the $\widetilde{\mathcal{O}}(T^{2/3})$ regret of [27]. See Algorithm 1 for the complete pseudocode.

Algorithm design. First, to handle unknown transition, we follow the common practice (dating back to [13]) to maintain a confidence set of the transition, which is updated whenever the visitation count of a certain state-action pair is doubled. We call the period between two model updates an epoch, and use $\mathcal{P}_k$ to denote the confidence set for epoch k, formally defined in Eq. (10).

In episode t, the policy π_t is defined via the standard multiplicative weight algorithm (also connected to Natural Policy Gradient [18, 3, 30]), but importantly with the dilated bonuses incorporated such that $\pi_t(a|x) \propto \exp(-\eta \sum_{\tau=1}^{t-1}(\widehat{Q}_\tau(x,a) - B_\tau(x,a)))$. Here, η is a step size parameter, $\widehat{Q}_\tau(x,a)$ is an importance-weighted estimator for $Q_\tau^{\pi_\tau}(x,a)$ defined in Eq. (7), and $B_\tau(x,a)$ is the dilated bonus defined in Eq. (9).

More specifically, for a state x in layer h, $\widehat{Q}_t(x,a)$ is defined as $\frac{L_{t,h}\mathbb{1}_t(x,a)}{\overline{q}_t(x,a)+\gamma}$, where $\mathbb{1}_t(x,a)$ is the indicator of whether (x,a) is visited during episode t; $L_{t,h}$ is the total loss suffered by the learner starting from layer h till the end of the episode; $\overline{q}_t(x,a) = \max_{\widehat{P}\in\mathcal{P}_k} q^{\widehat{P},\pi_t}(x,a)$ is the largest plausible value of $q_t(x,a)$ within the confidence set, which can be computed efficiently using the COMP-UOB procedure of [15] (see also Appendix C.1); and finally γ is a parameter used to control the maximum magnitude of $\widehat{Q}_t(x,a)$, inspired by the work of [23]. To get a sense of this estimator, consider the special case when $\gamma = 0$ and the transition is known so that we can set $\mathcal{P}_k = \{P\}$ and thus $\overline{q}_t = q_t$. Then, since the expectation of $L_{t,h}$ conditioned on (x,a) being visited is $Q_t^{\pi_t}(x,a)$ and the expectation of $\mathbb{1}_t(x,a)$ is $q_t(x,a)$, we know that $\widehat{Q}_t(x,a)$ is an unbiased estimator for $Q_t^{\pi_t}(x,a)$. The extra complication is simply due to the transition being unknown, forcing us to use $\overline{q}_t$ and $\gamma > 0$ to make sure that $\widehat{Q}_t(x,a)$ is an optimistic underestimator, an idea similar to [15].

Next, we explain the design of the dilated bonus B_t. Following the discussions of Section 3, we first figure out what the corresponding b_t function is in Eq. (1), by analyzing the regret bound without using any bonuses. The concrete form of b_t turns out to be Eq. (8), whose value at (x,a) is independent of a and thus written as $b_t(x)$ for simplicity. Note that Eq. (8) depends on the occupancy measure lower bound $\underline{q}_t(s,a) = \min_{\widehat{P}\in\mathcal{P}_k} q^{\widehat{P},\pi_t}(x,a)$, the opposite of $\overline{q}_t(s,a)$, which can also be computed efficiently using a procedure similar to COMP-UOB (see Appendix C.1). Once again, to get a sense of this, consider the special case with a known transition so that we can set $\mathcal{P}_k = \{P\}$ and thus $\overline{q}_t = \underline{q}_t = q_t$. Then, one see that $b_t(x)$ is simply upper bounded by $\mathbb{E}_{a\sim\pi_t(\cdot|x)}\left[{}^{3\gamma H}/_{q_t(x,a)}\right] = {}^{3\gamma H|A|}/_{q_t(x)}$, which is inversely related to the probability of visiting state x, matching the intuition we provided in Section 3 (that $b_t(x)$ is large if x is rarely visited). The extra complication of Eq. (8) is again just due to the unknown transition.

With $b_t(x)$ ready, the final form of the dilated bonus B_t is defined following the dilated Bellman equation of Eq. (4), except that since P is unknown, we once again apply optimism and find the

[3]We use $y \overset{+}{\leftarrow} z$ as a shorthand for the increment operation $y \leftarrow y + z$.

Algorithm 1 Policy Optimization with Dilated Bonuses (Tabular Case)

Parameters: $\delta \in (0,1)$, $\eta = \min\{1/24H^3, 1/\sqrt{|X||A|HT}\}$, $\gamma = 2\eta H$.
Initialization: Set epoch index $k=1$ and confidence set $\mathcal{P}_1$ as the set of all transition functions. For all (x,a,x'), initialize counters $N_0(x,a) = N_1(x,a) = 0$, $N_0(x,a,x') = N_1(x,a,x') = 0$.
for $t = 1, 2, \ldots, T$ **do**

Step 1: Compute and execute policy. Execute π_t for one episode, where

$$\pi_t(a|x) \propto \exp\left(-\eta \sum_{\tau=1}^{t-1}\left(\widehat{Q}_\tau(x,a) - B_\tau(x,a)\right)\right), \tag{6}$$

and obtain trajectory $\{(x_{t,h}, a_{t,h}, \ell_t(x_{t,h}, a_{t,h}))\}_{h=0}^{H-1}$.

Step 2: Construct Q-function estimators. For all $h \in \{0,\ldots,H-1\}$ and $(x,a) \in X_h \times A$,

$$\widehat{Q}_t(x,a) = \frac{L_{t,h}}{\overline{q}_t(x,a) + \gamma}\mathbb{1}_t(x,a), \tag{7}$$

with $L_{t,h} = \sum_{i=h}^{H-1} \ell_t(x_{t,i}, a_{t,i})$, $\overline{q}_t(x,a) = \max_{\widehat{P}\in\mathcal{P}_k} q^{\widehat{P},\pi_t}(x,a)$, $\mathbb{1}_t(x,a) = \mathbb{1}\{x_{t,h} = x, a_{t,h} = a\}$.

Step 3: Construct bonus functions. For all $(x,a) \in X \times A$,

$$b_t(x) = \mathbb{E}_{a\sim\pi_t(\cdot|x)}\left[\frac{3\gamma H + H(\overline{q}_t(x,a) - \underline{q}_t(x,a))}{\overline{q}_t(x,a)+\gamma}\right] \tag{8}$$

$$B_t(x,a) = b_t(x) + \left(1 + \frac{1}{H}\right)\max_{\widehat{P}\in\mathcal{P}_k} \mathbb{E}_{x'\sim\widehat{P}(\cdot|x,a)}\mathbb{E}_{a'\sim\pi_t(\cdot|x')}\left[B_t(x',a')\right] \tag{9}$$

where $\underline{q}_t(x,a) = \min_{\widehat{P}\in\mathcal{P}_k} q^{\widehat{P},\pi_t}(x,a)$ and $B_t(x_H, a) = 0$ for all a.

Step 4: Update model estimation. $\forall h < H$, $N_k(x_{t,h}, a_{t,h}) \xleftarrow{+} 1$, $N_k(x_{t,h}, a_{t,h}, x_{t,h+1}) \xleftarrow{+} 1$.[3]
if $\exists h,\ N_k(x_{t,h}, a_{t,h}) \geq \max\{1, 2N_{k-1}(x_{t,h}, a_{t,h})\}$ **then**

Increment epoch index $k \xleftarrow{+} 1$ and copy counters: $N_k \leftarrow N_{k-1}$, $N_k \leftarrow N_{k-1}$.
Compute empirical transition $\overline{P}_k(x'|x,a) = \frac{N_k(x,a,x')}{\max\{1, N_k(x,a)\}}$ and confidence set:

$$\begin{aligned}\mathcal{P}_k = &\Big\{\widehat{P} : \left|\widehat{P}(x'|x,a) - \overline{P}_k(x'|x,a)\right| \leq \mathit{conf}_k(x'|x,a), \\ &\quad \forall (x,a,x') \in X_h \times A \times X_{h+1}, h = 0, 1, \ldots, H-1\Big\},\end{aligned} \tag{10}$$

where $\quad \mathit{conf}_k(x'|x,a) = 4\sqrt{\frac{\overline{P}_k(x'|x,a)\ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_k(x,a)\}}} + \frac{28\ln\left(\frac{T|X||A|}{\delta}\right)}{3\max\{1, N_k(x,a)\}}$.

largest possible value within the confidence set (see Eq. (9)). This can again be efficiently computed; see Appendix C.1. This concludes the complete algorithm design.

Regret analysis. The regret guarantee of Algorithm 1 is presented below:

Theorem 4.1. *Algorithm 1 ensures that with probability* $1 - \mathcal{O}(\delta)$, $\text{Reg} = \widetilde{\mathcal{O}}\left(H^2|X|\sqrt{AT} + H^4\right)$.

Again, this improves the $\widetilde{\mathcal{O}}(T^{2/3})$ regret of [27]. It almost matches the best existing upper bound for this problem, which is $\widetilde{\mathcal{O}}(H|X|\sqrt{|A|T})$ [15]. While it is unclear to us whether this small gap can be closed using policy optimization, we point out that our algorithm is arguably more efficient than that of [15], which performs global convex optimization over the set of all plausible occupancy measures in each episode.

The complete proof of this theorem is deferred to Appendix C. Here, we only sketch an outline of proving Eq. (5), which, according to the discussions in Section 3, is the most important part of the analysis. Specifically, we decompose the left-hand side of Eq. (5), $\sum_x q^\star(x) \sum_t \langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), Q_t(x,\cdot) - B_t(x,\cdot)\rangle$, as $\textsc{Bias-1} + \textsc{Bias-2} + \textsc{Reg-Term}$, where

- $\textsc{Bias-1} = \sum_x q^\star(x) \sum_t \langle \pi_t(\cdot|x), Q_t(x,\cdot) - \widehat{Q}_t(x,\cdot)\rangle$ measures the amount of underestimation of $\widehat{Q}_t$ related to π_t, which can be bounded by $\sum_t \sum_{x,a} q^\star(x)\pi_t(a|x)\left(\frac{2\gamma H + H(\overline{q}_t(x,a) - \underline{q}_t(x,a))}{\overline{q}_t(x,a)+\gamma}\right) + \widetilde{\mathcal{O}}\left({}^{H}\!/\!{}_{\eta}\right)$ with high probability (Lemma C.1);
- $\textsc{Bias-2} = \sum_x q^\star(x) \sum_t \langle \pi^\star(\cdot|x), \widehat{Q}_t(x,\cdot) - Q_t(x,\cdot)\rangle$ measures the amount of overestimation of $\widehat{Q}_t$ related to $\pi^\star$, which can be bounded by $\widetilde{\mathcal{O}}\left({}^{H}\!/\!{}_{\eta}\right)$ since $\widehat{Q}_t$ is an underestimator (Lemma C.2);
- $\textsc{Reg-Term} = \sum_x q^\star(x) \sum_t \langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), \widehat{Q}_t(x,\cdot) - B_t(x,\cdot)\rangle$ is directly controlled by the multiplicative weight update, and is bounded by $\sum_t \sum_{x,a} q^\star(x)\pi_t(a|x)\left(\frac{\gamma H}{\overline{q}_t(x,a)+\gamma} + \frac{B_t(x,a)}{H}\right) + \widetilde{\mathcal{O}}\left({}^{H}\!/\!{}_{\eta}\right)$ with high probability (Lemma C.3).

Combining all with the definition of b_t proves the key Eq. (5) (with the $o(T)$ term being $\widetilde{\mathcal{O}}({}^{H}\!/\!{}_{\eta})$).

5 The Linear-Q Case

In this section, we move on to the more challenging setting where the number of states might be infinite, and function approximation is used to generalize the learner's experience to unseen states. We consider the most basic linear function approximation scheme where for any π, the Q-function $Q_t^\pi(x,a)$ is linear in some known feature vector $\phi(x,a)$, formally stated below.

Assumption 1 (Linear-Q)**.** *Let $\phi(x,a) \in \mathbb{R}^d$ be a known feature vector of the state-action pair (x,a). We assume that for any episode t, policy π, and layer h, there exists an unknown weight vector $\theta_{t,h}^\pi \in \mathbb{R}^d$ such that for all $(x,a) \in X_h \times A$, $Q_t^\pi(x,a) = \phi(x,a)^\top \theta_{t,h}^\pi$. Without loss of generality, we assume $\|\phi(x,a)\| \le 1$ for all (x,a) and $\|\theta_{t,h}^\pi\| \le \sqrt{d}H$ for all t, h, π.*

For justification on the last condition on norms, see [30, Lemma 8]. This linear-Q assumption has been made in several recent works with stationary losses [1, 30] and also in [24] with the same adversarial losses.[4] It is weaker than the linear MDP assumption (see Section 6) as it does not pose explicit structure requirements on the loss and transition functions. Due to this generality, however, our algorithm also requires access to a *simulator* to obtain samples drawn from the transition, formally stated below.

Assumption 2 (Simulator)**.** *The learner has access to a simulator, which takes a state-action pair $(x,a) \in X \times A$ as input, and generates a random outcome of the next state $x' \sim P(\cdot|x,a)$.*

Note that this assumption is also made by [24] and more earlier works with stationary losses (see e.g., [4, 28]).[5] In this setting, we propose a new policy optimization algorithm with $\widetilde{\mathcal{O}}(T^{2/3})$ regret. See Algorithm 2 for the pseudocode.

Algorithm design. The algorithm still follows the multiplicative weight update Eq. (11) in each state $x \in X_h$ (for some h), but now with $\phi(x,a)^\top \widehat{\theta}_{t,h}$ as an estimator for $Q_t^{\pi_t}(x,a) = \phi(x,a)^\top \theta_{t,h}^{\pi_t}$, and $\textsc{Bonus}(t,x,a)$ as the dilated bonus $B_t(x,a)$. Specifically, the construction of the weight estimator $\widehat{\theta}_{t,h}$ follows the idea of [24] (which itself is based on the linear bandit literature) and is defined in Eq. (12) as $\widehat{\Sigma}_{t,h}^+ \phi(x_{t,h}, a_{t,h}) L_{t,h}$. Here, $\widehat{\Sigma}_{t,h}^+$ is an ϵ-accurate estimator of $(\gamma I + \Sigma_{t,h})^{-1}$, where γ is a small parameter and $\Sigma_{t,h} = \mathbb{E}_t[\phi(x_{t,h}, a_{t,h})\phi(x_{t,h}, a_{t,h})^\top]$ is the covariance matrix for layer h under policy π_t; $L_{t,h} = \sum_{i=h}^{H-1} \ell_t(x_{t,i}, a_{t,i})$ is again the loss suffered by the learner starting from layer h, whose conditional expectation is $Q_t^{\pi_t}(x_{t,h}, a_{t,h}) = \phi(x_{t,h}, a_{t,h})^\top \theta_{t,h}^{\pi_t}$. Therefore,

[4] The assumption in [24] is stated slightly differently (e.g., their feature vectors are independent of the action). However, it is straightforward to verify that the two versions are equivalent.

[5] The simulator required by [24] is in fact slightly weaker than ours and those from earlier works — it only needs to be able to generate a trajectory starting from x_0 for any policy.

Algorithm 2 Policy Optimization with Dilated Bonuses (Linear-Q Case)

parameters: $\gamma, \beta, \eta, \epsilon$, $M = \left\lceil \frac{24 \ln(dHT)}{\epsilon^2\gamma^2} \right\rceil$, $N = \left\lceil \frac{2}{\gamma} \ln \frac{1}{\epsilon\gamma} \right\rceil$.

for $t = 1, 2, \ldots, T$ **do**

Step 1: Interact with the environment. Execute π_t, which is defined such that for each $x \in X_h$,

$$\pi_t(a|x) \propto \exp\left(-\eta \sum_{\tau=1}^{t-1} \left(\phi(x,a)^\top \widehat{\theta}_{\tau,h} - \textsc{Bonus}(\tau, x, a)\right)\right), \tag{11}$$

and obtain trajectory $\{(x_{t,h}, a_{t,h}, \ell_t(x_{t,h}, a_{t,h}))\}_{h=0}^{H-1}$.

Step 2: Construct covariance matrix inverse estimators.

$$\left\{\widehat{\Sigma}^+_{t,h}\right\}_{h=0}^{H-1} = \textsc{GeometricResampling}\left(t, M, N, \gamma\right). \qquad \text{(see Algorithm 7)}$$

Step 3: Construct Q-function weight estimators. For $h = 0, \ldots, H-1$, compute

$$\widehat{\theta}_{t,h} = \widehat{\Sigma}^+_{t,h} \phi(x_{t,h}, a_{t,h}) L_{t,h}, \qquad \text{where } L_{t,h} = \sum_{i=h}^{H-1} \ell_t(x_{t,i}, a_{t,i}). \tag{12}$$

Algorithm 3 $\textsc{Bonus}(t, x, a)$

if $\textsc{Bonus}(t, x, a)$ *has been called before* **then**

return the value of $\textsc{Bonus}(t, x, a)$ calculated last time.

Let h be such that $x \in X_h$. **if** $h = H$ **then** **return** 0.

Compute $\pi_t(\cdot|x)$, defined in Eq. (11) (which involves recursive calls to $\textsc{Bonus}$ for smaller t).

Get a sample of the next state $x' \leftarrow \textsc{Simulator}(x, a)$.

Compute $\pi_t(\cdot|x')$ (again, defined in Eq. (11)), and sample an action $a' \sim \pi_t(\cdot|x')$.

return $\beta\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}} + \mathbb{E}_{j\sim\pi_t(\cdot|x)}\left[\beta\|\phi(x,j)\|^2_{\widehat{\Sigma}^+_{t,h}}\right] + \left(1 + \frac{1}{H}\right) \textsc{Bonus}(t, x', a')$.

when γ and ϵ approach 0, one see that $\widehat{\theta}_{t,h}$ is indeed an unbiased estimator of $\theta^{\pi_t}_{t,h}$. We adopt the $\textsc{GeometricResampling}$ procedure (see Algorithm 7) of [24] to compute $\widehat{\Sigma}^+_{t,h}$, which involves calling the simulator multiple times.

Next, we explain the design of the dilated bonus. Again, following the general principle discussed in Section 3, we identify $b_t(x,a)$ in this case as $\beta\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}} + \mathbb{E}_{j\sim\pi_t(\cdot|x)}\left[\beta\|\phi(x,j)\|^2_{\widehat{\Sigma}^+_{t,h}}\right]$ for some parameter $\beta > 0$. Further following the dilated Bellman equation Eq. (4), we thus define $\textsc{Bonus}(t, x, a)$ recursively as the last line of Algorithm 3, where we replace the expectation $\mathbb{E}_{(x',a')}[\textsc{Bonus}(t, x', a')]$ with one single sample for efficient implementation.

However, even more care is needed to actually implement the algorithm. First, since the state space is potentially infinite, one cannot actually calculate and store the value of $\textsc{Bonus}(t, x, a)$ for all (x, a), but can only calculate them on-the-fly when needed. Moreover, unlike the estimators for $Q^{\pi_t}_t(x, a)$, which can be succinctly represented and stored via the weight estimator $\widehat{\theta}_{t,h}$, this is not possible for $\textsc{Bonus}(t, x, a)$ due to the lack of any structure. Even worse, the definition of $\textsc{Bonus}(t, x, a)$ itself depends on $\pi_t(\cdot|x)$ and also $\pi_t(\cdot|x')$ for the afterstate x', which, according to Eq. (11), further depends on $\textsc{Bonus}(\tau, x, a)$ for $\tau < t$, resulting in a complicated recursive structure. This is also why we present it as a procedure in Algorithm 3 (instead of $B_t(x, a)$). In total, this leads to $(TAH)^{\mathcal{O}(H)}$ number of calls to the simulator. Whether this can be improved is left as a future direction.

Regret guarantee By showing that Eq. (5) holds in expectation for our algorithm, we obtain the following regret guarantee. (See Appendix D for the proof.)

Theorem 5.1. *Under Assumption 1 and Assumption 2, with appropriate choices of the parameters $\gamma, \beta, \eta, \epsilon$, Algorithm 2 ensures $\mathbb{E}[\mathrm{Reg}] = \widetilde{\mathcal{O}}\left(H^2(dT)^{2/3}\right)$ (the dependence on $|A|$ is only logarithmic).*

This matches the $\widetilde{\mathcal{O}}(T^{2/3})$ regret of [24, Theorem 1], without the need of their assumption which essentially says that the learner is given an exploratory policy to start with.[6] To our knowledge, this is the first no-regret algorithm for linear function approximation (with adversarial losses and bandit feedback) when no exploratory assumptions are made.

6 Improvements with an Exploratory Policy

Previous sections have demonstrated the role of dilated bonuses in providing global exploration. In this section, we further discuss what dilated bonuses can achieve when an exploratory policy π_0 is given in linear function approximation settings. Formally, let $\Sigma_h = \mathbb{E}[\phi(x_h, a_h)\phi(x_h, a_h)^\top]$ denote the covariance matrix for features in layer h following π_0 (that is, the expectation is taken over a trajectory $\{(x_h, a_h)\}_{h=0}^{H-1}$ with $a_h \sim \pi_0(\cdot|x_h)$), then we assume the following.

Assumption 3 (An exploratory policy)**.** *An exploratory policy π_0 is given to the learner ahead of time, and guarantees that for any h, the eigenvalues of Σ_h are at least $\lambda_{\min} > 0$.*

The same assumption is made by [24] (where they simply let π_0 be the uniform exploration policy). As mentioned, under this assumption they achieve $\widetilde{\mathcal{O}}(T^{2/3})$ regret. By slightly modifying our Algorithm 2 (specifically, executing π_0 with a small probability in each episode and setting the parameters differently), we achieve the following improved result.

Theorem 6.1. *Under Assumptions 1, 2, and 3, Algorithm 8 ensures* $\mathbb{E}[\mathrm{Reg}] = \widetilde{\mathcal{O}}\big(\sqrt{\frac{H^4 T}{\lambda_{\min}}} + \sqrt{H^5 dT}\big)$.

Removing the simulator One drawback of our algorithm is that it requires exponential in H number of calls to the simulator. To address this issue, and in fact, to also completely remove the need of a simulator, we further consider a special case where the transition function also has a low-rank structure, known as the linear MDP setting.

Assumption 4 (Linear MDP)**.** *The MDP satisfies Assumption 1 and that for any h and $x' \in X_{h+1}$, there exists a weight vector $\nu_h^{x'} \in \mathbb{R}^d$ such that $P(x'|x, a) = \phi(x, a)^\top \nu_h^{x'}$ for all $(x, a) \in X_h \times A$.*

There is a surge of works studying this setting, with [7] being the closest to us. They achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret but require full-information feedback of the loss functions, and there are no existing results for the bandit feedback setting without a simulator. We propose the first algorithm with sublinear regret for this problem, shown in Algorithm 10 of Appendix F due to space limit.

The structure of Algorithm 10 is very similar to that of Algorithm 2, with the same definition of $b_t(x, a)$. However, due to the low-rank transition structure, we are now able to efficiently construct estimators of $B_t(x, a)$ even for unseen state-action pairs using function approximation, bypassing the requirement of a simulator. Specifically, observe that according to Eq. (4), for each $x \in X_h$, under Assumption 4 $B_t(x, a)$ can be written as $b_t(x, a) + \phi(x, a)^\top \Lambda_{t,h}^{\pi_t}$, where $\Lambda_{t,h}^{\pi_t} = (1 + \frac{1}{H}) \int_{x' \in X_{h+1}} \mathbb{E}_{a' \sim \pi_t(\cdot|x')}[B_t(x', a')]\nu_h^{x'} \mathrm{d}x'$ is a vector independent of (x, a). Thus, by the same idea of estimating $\theta_{t,h}^{\pi_t}$, we can estimate $\Lambda_{t,h}^{\pi_t}$ as well, thus succinctly representing $B_t(x, a)$ for all (x, a).

Recall that estimating $\theta_{t,h}^{\pi_t}$ (and thus also $\Lambda_{t,h}^{\pi_t}$) requires constructing the covariance matrix inverse estimate $\widehat{\Sigma}_{t,h}^+$. Due to the lack of a simulator, another important change in the algorithm is to construct $\widehat{\Sigma}_{t,h}^+$ using *online* samples. To do so, we divide the entire horizon into epochs with equal length, and only update the policy optimization algorithm at the beginning of an epoch. Within an epoch, we keep executing the same policy and collect several trajectories, which are then used to construct $\widehat{\Sigma}_{t,h}^+$. With these changes, we successfully remove the need of a simulator, and prove the guarantee below.

Theorem 6.2. *Under Assumption 3 and Assumption 4, Algorithm 10 ensures* $\mathbb{E}[\mathrm{Reg}] = \widetilde{\mathcal{O}}\left(T^{6/7}\right)$ *(see Appendix F for dependence on other parameters).*

One potential direction to further improve our algorithm is to reuse data across different epochs, an idea adopted by several recent works [35, 19] for different problems. We also conjecture that

[6] Under an even strong assumption that every policy is exploratory, they also improve the regret to $\widetilde{\mathcal{O}}(\sqrt{T})$; see [24, Theorem 2].

Assumption 3 can be removed, but we meet some technical difficulty in proving so. We leave these for future investigation.

Acknowledgments and Disclosure of Funding

We thank Gergely Neu and Julia Olkhovskaya for discussions on the technical details of their GEOMETRICRESAMPLING procedure. This work is supported by NSF Award IIS-1943607 and a Google Faculty Research Award.

References

[1] Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702, 2019.

[2] Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. In *Advances in Neural Information Processing Systems*, 2020.

[3] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pages 64–66, 2020.

[4] Mohammad Gheshlaghi Azar, Rémi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model. In *International Conference on Machine Learning*, 2012.

[5] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272, 2017.

[6] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[7] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294, 2020.

[8] Liyu Chen and Haipeng Luo. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. In *International Conference on Machine Learning*, 2021.

[9] Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. In *Conference On Learning Theory*, 2021.

[10] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, 2015.

[11] Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. *International Conference on Learning Representations*, 2020.

[12] Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pages 1578–1586, 2018.

[13] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[14] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in neural information processing systems*, pages 4863–4873, 2018.

[15] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, 2020.

[16] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

[17] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[18] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[19] Nevena Lazic, Dong Yin, Yasin Abbasi-Yadkori, and Csaba Szepesvari. Improved regret bound and experience replay in regularized policy iteration. In *International Conference on Machine Learning*, 2021.

[20] Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in Neural Information Processing Systems*, 2020.

[21] Sergey Levine and Vladlen Koltun. Guided policy search. In *International conference on machine learning*, pages 1–9, 2013.

[22] Haipeng Luo. Lecture 2, introduction to online learning, 2017. Available at `https://haipeng-luo.net/courses/CSCI699/lecture2.pdf`.

[23] Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, 2015.

[24] Gergely Neu and Julia Olkhovskaya. Online learning in mdps with linear function approximation and bandit feedback. *arXiv preprint arXiv:2007.01612*, 2020.

[25] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[26] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[27] Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613, 2020.

[28] Aaron Sidford, Mengdi Wang, Xian Wu, Lin F Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5192–5202, 2018.

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

[30] Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015, 2021.

[31] Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International Conference on Machine Learning*, pages 10170–10180, 2020.

[32] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756, 2020.

[33] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964, 2020.

[34] Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

[35] Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, 2021.

A Auxiliary Lemmas

In this section, we list auxiliary lemmas that are useful in our analysis. First, we show some concentration inequalities.

Lemma A.1 ((A special form of) Freedman's inequality, Theorem 1 of [6]). *Let $\mathcal{F}_0 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ be real random variables such that X_i is $\mathcal{F}_i$-measurable, $\mathbb{E}[X_i|\mathcal{F}_i] = 0$, $|X_i| \le b$, and $\sum_{i=1}^n \mathbb{E}[X_i^2|\mathcal{F}_i] \le V$ for some fixed $b \ge 0$ and $V \ge 0$. Then for any $\delta \in (0,1)$, we have with probability at least $1-\delta$,*

$$\sum_{i=1}^n X_i \le \frac{V}{b} + b\log(1/\delta).$$

Throughout the appendix, we let $\mathcal{F}_t$ be the σ-algebra generated by the observations before episode t.

Lemma A.2 (Adapted from Lemma 11 of [15]; see also [23]). *For all x, a, let $\{z_t(x,a)\}_{t=1}^T$ be a sequence of functions where $z_t(x,a) \in [0,R]$ is $\mathcal{F}_t$-measurable. Let $Z_t(x,a) \in [0,R]$ be a random variable such that $\mathbb{E}_t[Z_t(x,a)] = z_t(x,a)$. Then with probability at least $1-\delta$,*

$$\sum_{t=1}^T \sum_{x,a} \left(\frac{\mathbb{1}_t(x,a) Z_t(x,a)}{\overline{q}_t(x,a)+\gamma} - \frac{q_t(x,a) z_t(x,a)}{\overline{q}_t(x,a)} \right) \le \frac{RH}{2\gamma} \ln \frac{H}{\delta}.$$

Lemma A.3 (Matrix Azuma, Theorem 7.1 of [29]). *Consider an adapted sequence $\{X_k\}_{k=1}^n$ of self-adjoint matrices in dimension d, and a fixed sequence $\{A_k\}_{k=1}^n$ of self-adjoint matrices that satisfy*

$$\mathbb{E}_k X_k = 0 \;\; \text{and} \;\; X_k^2 \preceq A_k^2 \text{ almost surely}$$

Define the variance parameter

$$\sigma^2 = \left\| \frac{1}{n} \sum_{k=1}^n A_k^2 \right\|_{op}.$$

Then, for all $\tau > 0$,

$$\Pr\left\{ \left\| \frac{1}{n} \sum_{k=1}^n X_k \right\|_{op} \ge \tau \right\} \le d e^{-n\tau^2/8\sigma^2}.$$

Next, we show a classic regret bound for the exponential weight algorithm, which can be found, for example, in [22].

Lemma A.4 (Regret bound of exponential weight, extracted from Theorem 1 of [22]). *Let $\eta > 0$, and let $\pi_t \in \Delta(A)$ and $\ell_t \in \mathbb{R}^A$ satisfy the following for all $t \in [T]$ and $a \in A$:*

$$\pi_1(a) = \frac{1}{|A|},$$

$$\pi_{t+1}(a) = \frac{\pi_t(a) e^{-\eta \ell_t(a)}}{\sum_{a' \in A} \pi_t(a') e^{-\eta \ell_t(a')}},$$

$$|\eta \ell_t(a)| \leq 1.$$

Then for any $\pi^\star \in \Delta(A)$,

$$\sum_{t=1}^{T} \sum_{a \in A} (\pi_t(a) - \pi^\star(a))\ell_t(a) \leq \frac{\ln |A|}{\eta} + \eta \sum_{t=1}^{T} \sum_{a \in A} \pi_t(a)\ell_t(a)^2.$$

B Proofs Omitted in Section 3

In this section, we prove Lemma 3.1. In fact, we prove two generalized versions of it. Lemma B.1 states that the lemma holds even when we replace the definition of $B_t(x,a)$ by an upper bound of the right hand side of Eq. (4). (Note that Lemma 3.1 is clearly a special case with $\widehat{P} = P$.)

Lemma B.1. *Let $b_t(x,a)$ be a non-negative loss function, and $\widehat{P}$ be a transition function. Suppose that the following holds for all x, a:*

$$\begin{aligned} B_t(x,a) &= b_t(x,a) + \left(1 + \frac{1}{H}\right) \mathbb{E}_{x' \sim \widehat{P}(\cdot|x,a)} \mathbb{E}_{a' \sim \pi_t(\cdot|x')} \left[B_t(x',a')\right] \\ &\geq b_t(x,a) + \left(1 + \frac{1}{H}\right) \mathbb{E}_{x' \sim P(\cdot|x,a)} \mathbb{E}_{a' \sim \pi_t(\cdot|x')} \left[B_t(x',a')\right] \end{aligned} \tag{13}$$

with $B_t(x_H, a) \triangleq 0$, and suppose that Eq. (5) holds. Then

$$\text{Reg} \leq o(T) + 3 \sum_{t=1}^{T} \widehat{V}^{\pi_t}(x_0; b_t).$$

where $\widehat{V}^\pi$ is the state value function under the transition function $\widehat{P}$ and policy π.

Proof of Lemma B.1. By rearranging Eq. (5), we see that

$$\begin{aligned} \text{Reg} \leq o(T) &+ \underbrace{\sum_{t=1}^{T} \sum_{x,a} q^\star(x)\pi^\star(a|x)b_t(x,a)}_{\text{TERM}_1} \\ &+ \underbrace{\frac{1}{H}\sum_{t=1}^{T} \sum_{x,a} q^\star(x)\pi_t(a|x)B_t(x,a)}_{\text{TERM}_2} + \underbrace{\sum_{t=1}^{T} \sum_{x,a} q^\star(x)\Big(\pi_t(a|x) - \pi^\star(a|x)\Big)B_t(x,a)}_{\text{TERM}_3}. \end{aligned}$$

We first focus on TERM_3, and focus on a single layer $0 \leq h \leq H-1$ and a single t:

$$\begin{aligned} &\sum_{x \in X_h} \sum_{a \in A} q^\star(x) \left(\pi_t(a|x) - \pi^\star(a|x)\right) B_t(x,a) \\ &= \sum_{x \in X_h} \sum_{a \in A} q^\star(x)\pi_t(a|x)B_t(x,a) - \sum_{x \in X_h} \sum_{a \in A} q^\star(x)\pi^\star(a|x)B_t(x,a) \\ &= \sum_{x \in X_h} \sum_{a \in A} q^\star(x)\pi_t(a|x)B_t(x,a) \\ &\qquad - \sum_{x \in X_h} \sum_{a \in A} q^\star(x)\pi^\star(a|x) \left(b_t(x,a) + \left(1 + \frac{1}{H}\right) \mathbb{E}_{x' \sim \widehat{P}(\cdot|x,a)} \mathbb{E}_{a' \sim \pi_t(\cdot|x')} \left[B_t(x',a')\right]\right) \\ &\leq \sum_{x \in X_h} \sum_{a \in A} q^\star(x)\pi_t(a|x)B_t(x,a) \\ &\qquad - \sum_{x \in X_h} \sum_{a \in A} q^\star(x)\pi^\star(a|x) \left(b_t(x,a) + \left(1 + \frac{1}{H}\right) \mathbb{E}_{x' \sim P(\cdot|x,a)} \mathbb{E}_{a' \sim \pi_t(\cdot|x')} \left[B_t(x',a')\right]\right) \\ &= \sum_{x \in X_h} \sum_{a \in A} q^\star(x)\pi_t(a|x)B_t(x,a) - \sum_{x \in X_{h+1}} \sum_{a \in A} q^\star(x)\pi_t(a|x)B_t(x,a) \end{aligned}$$

$$-\sum_{x\in X_h}\sum_{a\in A} q^\star(x)\pi^\star(a|x)b_t(x,a) - \frac{1}{H}\sum_{x\in X_{h+1}}\sum_{a\in A} q^\star(x)\pi_t(a|x)B_t(x,a),$$

where the last step uses the fact $\sum_{x\in X_h}\sum_{a\in A} q^\star(x)\pi^\star(a|x)P(x'|x,a) = q^\star(x')$ (and then changes the notation (x',a') to (x,a)). Now summing this over $h = 0, 1, \ldots, H-1$ and $t = 1, \ldots, T$, and combining with TERM_1 and TERM_2, we get

$$\text{TERM}_1 + \text{TERM}_2 + \text{TERM}_3 = \left(1+\frac{1}{H}\right)\sum_{t=1}^{T}\sum_a \pi_t(a|x_0)B_t(x_0,a).$$

Finally, we relate $\sum_a \pi_t(a|x_0)B_t(x_0,a)$ to $\widehat{V}^{\pi_t}(x_0;b_t)$. Below, we show by induction that for $x \in X_h$ and any a,

$$\sum_{a\in A}\pi_t(a|x)B_t(x,a) \le \left(1+\frac{1}{H}\right)^{H-h-1}\widehat{V}^{\pi_t}(x;b_t).$$

When $h = H-1$, $\sum_a \pi_t(a|x)B_t(x,a) = \sum_a \pi_t(a|x)b_t(x,a) = \widehat{V}^{\pi_t}(x;b_t)$. Suppose that the hypothesis holds for all $x \in X_h$. Then for any $x \in X_{h-1}$,

$$\begin{aligned}
\sum_{a\in A}\pi_t(a|x)B_t(x,a) &= \sum_a \pi_t(a|x)\Big(b_t(x,a) + \left(1+\frac{1}{H}\right)\mathbb{E}_{x'\sim\widehat{P}(\cdot|x,a)}\mathbb{E}_{a'\sim\pi_t(\cdot|x')}\left[B_t(x',a')\right]\Big)\\
&\le \sum_a \pi_t(a|x)\Big(b_t(x,a) + \left(1+\frac{1}{H}\right)^{H-h}\mathbb{E}_{x'\sim\widehat{P}(\cdot|x,a)}\left[\widehat{V}^{\pi_t}(x';b_t)\right]\Big) && \text{(induction hypothesis)}\\
&\le \left(1+\frac{1}{H}\right)^{H-h}\sum_a \pi_t(a|x)\Big(b_t(x,a) + \mathbb{E}_{x'\sim\widehat{P}(\cdot|x,a)}\left[\widehat{V}^{\pi_t}(x';b_t)\right]\Big) && (b_t(x,a)\ge 0)\\
&= \left(1+\frac{1}{H}\right)^{H-h}\widehat{V}^{\pi_t}(x;b_t),
\end{aligned}$$

finishing the induction. Applying the relation on $x = x_0$ and noticing that $\left(1+\frac{1}{H}\right)^H \le e < 3$ finishes the proof. ☐

Besides Lemma B.1, we also show Lemma B.2 below, which guarantees that Lemma 3.1 holds even if Eq. (4) and Eq. (5) only hold in expectation.

Lemma B.2. *Let $b_t(x,a)$ be a non-negative loss function that is fixed at the beginning of episode t, and let π_t be fixed at the beginning of episode t. Let $B_t(x,a)$ be a randomized bonus function that satisfies the following for all x, a:*

$$\mathbb{E}_t\left[B_t(x,a)\right] = b_t(x,a) + \left(1+\frac{1}{H}\right)\mathbb{E}_{x'\sim P(\cdot|x,a)}\mathbb{E}_{a'\sim\pi_t(\cdot|x')}\mathbb{E}_t\Big[B_t(x',a')\Big] \tag{14}$$

with $B_t(x_H,a) \triangleq 0$, and suppose that the following holds (simply taking expectations on Eq. (5)):

$$\begin{aligned}
&\mathbb{E}\left[\sum_x q^\star(x)\sum_{t=1}^{T}\sum_a \Big(\pi_t(a|x)-\pi^\star(a|x)\Big)\Big(Q_t^{\pi_t}(x,a)-B_t(x,a)\Big)\right]\\
&\qquad\le o(T) + \mathbb{E}\left[\sum_{t=1}^{T} V^{\pi^\star}(x_0;b_t)\right] + \frac{1}{H}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)B_t(x,a)\right].
\end{aligned} \tag{15}$$

Then

$$\mathbb{E}\left[\text{Reg}\right] \le o(T) + 3\mathbb{E}\left[\sum_{t=1}^{T} V^{\pi_t}(x_0;b_t)\right].$$

Proof. The proof of this lemma follows that of Lemma B.1 line-by-line (with $\widehat{P} = P$), except that we take expectations in all steps. ☐

C Details Omitted in Section 4

In this section, we first discuss the implementation details of Algorithm 1 in Section C.1, then we give the complete proof of Theorem 4.1 in Section C.2.

C.1 Implementation Details

The COMP-UOB procedure is the same as Algorithm 3 of [15], which shows how to efficiently compute an upper occupancy bound. We include the algorithm in Algorithm 4 for completeness. As Algorithm 1 also needs COMP-LOB, which computes a lower occupancy bound, we provide its complete pseudocode in Algorithm 5 as well.

Fix a state x. Define $f(\tilde{x})$ to be the maximum and minimum probability of visiting x starting from state $\tilde{x}$ for COMP-UOB and COMP-LOB, respectively. Then the two algorithms almost have the same procedure to find $f(\tilde{x})$ by solving the optimization in Eq. (16) subject to $\widehat{P}$ in the confidence set $\mathcal{P}$ via a greedy approach in Algorithm 6. The difference is that COMP-UOB sets OPTIMIZE to be $\max$ while COMP-LOB sets OPTIMIZE to be $\min$, and thus in Algorithm 6, $\{f(x)\}_{x \in X_k}$ is sorted in an ascending and a descending order, respectively.

Finally, we point out that the bonus function $B_t(s,a)$ defined in Eq. (9) can clearly also be computed using a greedy procedure similar to Algorithm 6. This concludes that the entire algorithm can be implemented efficiently.

$$f(\tilde{x}) = \sum_{a \in A} \pi_t(a|\tilde{x}) \left(\underset{\widehat{P}(\cdot|\tilde{x},a)}{\textsc{Optimize}} \sum_{x' \in X_{k(\tilde{x})+1}} \widehat{P}(x'|\tilde{x},a) f(x') \right) \tag{16}$$

Algorithm 4 COMP-UOB (Algorithm 3 of [15])

Input: a policy π_t, a state-action pair (x,a) and a confidence set $\mathcal{P}$ of the form

$$\left\{ \widehat{P} : \left| \widehat{P}(x'|x,a) - \bar{P}(x'|x,a) \right| \le \epsilon(x'|x,a),\ \forall (x,a,x') \right\}$$

Initialize: for all $\tilde{x} \in X_{k(x)}$, set $f(\tilde{x}) = \mathbb{1}\{\tilde{x} = x\}$.

for $k = k(x) - 1$ ***to*** 0 **do**

  for $\forall \tilde{x} \in X_k$ **do**

    Compute $f(\tilde{x})$ based on :

$$f(\tilde{x}) = \sum_{a \in A} \pi_t(a|\tilde{x}) \cdot \textsc{Greedy}\left(f, \bar{P}(\cdot|\tilde{x},a), \epsilon(\cdot|\tilde{x},a), \max\right)$$

Return: $\pi_t(a|x) f(x_0)$.

Algorithm 5 COMP-LOB

Input: a policy π_t, a state-action pair (x,a) and a confidence set $\mathcal{P}$ of the form

$$\left\{ \widehat{P} : \left| \widehat{P}(x'|x,a) - \bar{P}(x'|x,a) \right| \le \epsilon(x'|x,a),\ \forall (x,a,x') \right\}$$

Initialize: for all $\tilde{x} \in X_{k(x)}$, set $f(\tilde{x}) = \mathbb{1}\{\tilde{x} = x\}$.

for $k = k(x) - 1$ ***to*** 0 **do**

  for $\forall \tilde{x} \in X_k$ **do**

    Compute $f(\tilde{x})$ based on :

$$f(\tilde{x}) = \sum_{a \in A} \pi_t(a|\tilde{x}) \cdot \textsc{Greedy}\left(f, \bar{P}(\cdot|\tilde{x},a), \epsilon(\cdot|\tilde{x},a), \min\right)$$

Return: $\pi_t(a|x) f(x_0)$.

Algorithm 6 GREEDY

Input: $f : X \to [0,1]$, a distribution $\bar{p}$ over n states of layer k , positive numbers $\{\epsilon(x)\}_{x\in X_k}$, objective OPTIMIZE (max for COMP-UOB and min for COMP-LOB).
Initialize: $j^- = 1, j^+ = n$, sort $\{f(x)\}_{x\in X_k}$ and find σ such that

$$f(\sigma(1)) \le f(\sigma(2)) \le \cdots \le f(\sigma(n))$$

for OPTIMIZE $=$ max, and

$$f(\sigma(1)) \ge f(\sigma(2)) \ge \cdots \ge f(\sigma(n))$$

for OPTIMIZE $=$ min.
while $j^- < j^+$ **do**
  $x^- = \sigma(j^-), x^+ = \sigma(j^+)$
  $\delta^- = \min\{\bar{p}(x^-), \epsilon(x^-)\}$
  $\delta^+ = \min\{1-\bar{p}(x^+), \epsilon(x^+)\}$
  $\bar{p}(x^-) \leftarrow \bar{p}(x^-) - \min\{\delta^-, \delta^+\}$
  $\bar{p}(x^+) \leftarrow \bar{p}(x^+) + \min\{\delta^-, \delta^+\}$
  if $\delta_- \le \delta_+$ **then**
    $\epsilon(x^+) \leftarrow \epsilon(x^+) - \delta^-$
    $j^- \leftarrow j^- + 1$
  else
    $\epsilon(x^-) \leftarrow \epsilon(x^-) - \delta^+$
    $j^+ \leftarrow j^+ - 1$
Return: $\sum_{j=1}^n \bar{p}(\sigma(j)) f(\sigma(j))$

C.2 Omitted Proofs

To prove Theorem 4.1, as discussed in the analysis sketch of Section 4, we decompose the left-hand side of Eq. (5) as:

$$\begin{aligned}
&\sum_{t=1}^T \sum_x q^\star(x) \langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), Q_t(x,\cdot) - B_t(x,\cdot)\rangle \\
&= \underbrace{\sum_{t=1}^T \sum_x q^\star(x) \left\langle \pi_t(\cdot|x), Q_t(x,\cdot) - \widehat{Q}_t(x,\cdot)\right\rangle}_{\textsc{Bias-1}} + \underbrace{\sum_{t=1}^T \sum_x q^\star(x) \left\langle \pi^\star(\cdot|x), \widehat{Q}_t(x,\cdot) - Q_t(x,\cdot)\right\rangle}_{\textsc{Bias-2}} \\
&\quad + \underbrace{\sum_{t=1}^T \sum_x q^\star(x) \left\langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), \widehat{Q}_t(x,\cdot) - B_t(x,\cdot)\right\rangle}_{\textsc{Reg-Term}}. \qquad (17)
\end{aligned}$$

We bound each term in a corresponding lemma. Specifically, We show a high probability bound of BIAS-1 in Lemma C.1, a high probability bound of BIAS-2 in Lemma C.2, and a high-probability bound of REG-TERM in Lemma C.3. Finally, we show how to combine all terms with the definition of b_t in Theorem C.5, which is a restatement of Theorem 4.1.

Lemma C.1 (BIAS-1)**.** *With probability at least* $1 - 5\delta$*,*

$$\textsc{Bias-1} \le \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + \sum_{t=1}^T \sum_{x,a} q^\star(x)\pi_t(a|x) \left(\frac{2\gamma H + H\left(\overline{q}_t(x,a) - \underline{q}_t(x,a)\right)}{\overline{q}_t(x,a) + \gamma}\right).$$

Proof. In the proof, we assume that $P \in \mathcal{P}_k$ for all k, with holds with probability at least $1-4\delta$ as already shown in [15, Lemma 2]. Under this event, $\underline{q}_t(x,a) \le q_t(x,a) \le \overline{q}_t(x,a)$ for all t, x, a.

Let $Y_t = \sum_{x\in X} q^\star(x) \left\langle \pi_t(\cdot|x), \widehat{Q}_t(x,\cdot)\right\rangle$. First, we decompose BIAS-1 as

$$\sum_{t=1}^T (\mathbb{E}_t[Y_t] - Y_t) + \left(\sum_x q^\star(x) \langle \pi_t(\cdot|x), Q_t(x,\cdot)\rangle - \mathbb{E}_t[Y_t]\right). \qquad (18)$$

We will bound the first Martingale sequence using Freedman's inequality. Note that we have

$$\begin{aligned}
\mathrm{Var}_t[Y_t] &\le \mathbb{E}_t\left[\left(\sum_x q^\star(x)\left\langle \pi_t(\cdot|x), \widehat{Q}_t(x,\cdot)\right\rangle\right)^2\right] \\
&\le \mathbb{E}_t\left[\left(\sum_{x,a} q^\star(x)\pi_t(a|x)\right)\left(\sum_{x,a} q^\star(x)\pi_t(a|x)\widehat{Q}_t(x,a)^2\right)\right] && \text{(Cauchy-Schwarz)} \\
&= H\sum_{x,a} q^\star(x)\pi_t(a|x)\frac{L_{t,h}^2\mathbb{E}_t[\mathbb{1}_t(x,a)]}{(\overline{q}_t(x,a)+\gamma)^2} && \left(\textstyle\sum_{x,a} q^\star(x)\pi_t(a|x) = H\right) \\
&\le H\sum_{x,a} q^\star(x)\pi_t(a|x)\frac{q_t(x,a)H^2}{(\overline{q}_t(x,a)+\gamma)^2} && (L_{t,h}\le H \text{ and } \mathbb{E}_t[\mathbb{1}_t(x,a)] = q_t(s,a)) \\
&\le \sum_{x,a} q^\star(x)\pi_t(a|x)\frac{H^3}{\overline{q}_t(x,a)+\gamma} && (q_t(s,a)\le \overline{q}_t(x,a))
\end{aligned}$$

and $|Y_t| \le H\sup_{x,a}|\widehat{Q}(x,a)| \le \frac{H^2}{\gamma}$.

Moreover, for every t, the second term in Eq. (18) can be bounded as

$$\begin{aligned}
&\sum_x q^\star(x)\left\langle \pi_t(\cdot|x), Q_t(x,\cdot)\right\rangle - \mathbb{E}_t\left[\sum_x q^\star(x)\left\langle \pi_t(\cdot|x), \widehat{Q}_t(x,\cdot)\right\rangle\right] \\
&= \sum_{x,a} q^\star(x)\pi_t(a|x)Q_t(x,a)\left(1 - \frac{q_t(x,a)}{\overline{q}_t(x,a)+\gamma}\right) \\
&\le \sum_{x,a} q^\star(x)\pi_t(a|x)H\left(\frac{\overline{q}_t(x,a) - q_t(x,a) + \gamma}{\overline{q}_t(x,a)+\gamma}\right) && (Q_t(x,a)\le H) \\
&\le \sum_{x,a} q^\star(x)\pi_t(a|x)H\left(\frac{\overline{q}_t(x,a) - \underline{q}_t(x,a) + \gamma}{\overline{q}_t(x,a)+\gamma}\right). && (\underline{q}_t(x,a)\le q_t(x,a))
\end{aligned}$$

Combining them, and using Freedman's inequality (Lemma A.1), we have that with probability at least $1-5\delta$,

$$\begin{aligned}
\textsc{Bias-1} &= \sum_{t=1}^T\sum_x q^\star(x)\left\langle \pi_t(\cdot|x), Q_t(x,\cdot) - \widehat{Q}_t(x,\cdot)\right\rangle \\
&\le \sum_{t=1}^T\sum_{x,a} q^\star(x)\pi_t(a|x)H\left(\frac{\left(\overline{q}_t(x,a) - \underline{q}_t(x,a)\right)+\gamma}{\overline{q}_t(x,a)+\gamma}\right) \\
&\quad + \frac{\gamma}{H^2}\sum_{t=1}^T\sum_{x,a} q^\star(x)\pi_t(a|x)\frac{H^3}{\overline{q}_t(x,a)+\gamma} + \frac{H^2}{\gamma}\ln\frac{1}{\delta} \\
&\le \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + \sum_{t=1}^T\sum_{x,a} q^\star(x)\pi_t(a|x)\left(\frac{2\gamma H + H\left(\overline{q}_t(x,a) - \underline{q}_t(x,a)\right)}{\overline{q}_t(x,a)+\gamma}\right),
\end{aligned}$$

where we use $\gamma = 2\eta H$. □

Next, we bound $\textsc{Bias-2}$.

Lemma C.2 ($\textsc{Bias-2}$)**.** *With probability at least* $1-5\delta$*,* $\textsc{Bias-2} \le \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right)$.

Proof. We invoke Lemma A.2 with $z_t(x,a) = q^\star(x)\pi^\star(a|x)Q_t(x,a)$ and $Z_t(x,a) = q^\star(x)\pi^\star(a|x)\left(\mathbb{1}_t(x,a)L_t(x,a) + (1-\mathbb{1}_t(x,a))Q_t(x,a)\right)$. Then we get that with probability at least $1-\delta$ (recalling the definition $\widehat{Q}_t(x,a) = \frac{L_{t,h}}{\overline{q}_t(x,a)+\gamma}\mathbb{1}_t(x,a)$),

$$\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi^\star(a|x)\left(\widehat{Q}_t(x,a) - \frac{q_t(x,a)}{\overline{q}_t(x,a)}Q_t(x,a)\right) \le \frac{H^2}{2\gamma}\ln\frac{H}{\delta}, \tag{19}$$

Since with probability at least $1-4\delta$, $q_t(x,a) \le \overline{q}_t(x,a)$ for all t, x, a (by [15, Lemma 2]), Eq. (19) further implies that with probability at least $1-5\delta$,

$$\textsc{Bias-2} = \sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi^\star(x,a)\left(\widehat{Q}_t(x,a) - Q_t(x,a)\right) \le \frac{H^2}{2\gamma}\ln\frac{H}{\delta}.$$

Noting that $\gamma = 2\eta H$ finishes the proof. $\square$

We continue to bound $\textsc{Reg-Term}$.

Lemma C.3 ($\textsc{Reg-Term}$)**.** *With probability at least $1-5\delta$,*

$$\textsc{Reg-Term} \le \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + \sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)\left(\frac{\gamma H}{\overline{q}_t(x,a)+\gamma} + \frac{B_t(x,a)}{H}\right).$$

Proof. The algorithm runs individual exponential weight updates on each state with loss vectors $\widehat{Q}_t(x,\cdot) - B_t(x,\cdot)$, so we can apply standard results for exponential weight updates. Specifically, we can apply Lemma A.4 on each state x, and get

$$\sum_{t=1}^{T}\left\langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), \widehat{Q}_t(x,\cdot) - B_t(x,\cdot)\right\rangle \le \frac{\ln|A|}{\eta} + \eta\sum_{t=1}^{T}\sum_{a\in A}\pi_t(a|x)\left(\widehat{Q}_t(x,a) - B_t(x,a)\right)^2. \tag{20}$$

The condition required by Lemma A.4 (i.e., $\eta|\widehat{Q}_t(x,a) - B_t(x,a)| \le 1$) is verified in Lemma C.4. Summing Eq. (20) over states with weights $q^\star(x)$, we get

$$\begin{aligned}\textsc{Reg-Term} &\le \frac{H\ln|A|}{\eta} + \eta\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)\left(\widehat{Q}_t(x,a) - B_t(x,a)\right)^2 \\ &\le \frac{H\ln|A|}{\eta} + 2\eta\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)\widehat{Q}_t(x,a)^2 + 2\eta\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)B_t(x,a)^2.\end{aligned} \tag{21}$$

Below, we focus on the last two terms on the right-hand side of Eq. (21). First, we have

$$\begin{aligned}2\eta\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)\widehat{Q}_t(x,a)^2 &\le 2\eta\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)\frac{H^2\mathbb{1}_t(x,a)}{(\overline{q}_t(x,a)+\gamma)^2} \\ &= 2\eta H^2\sum_{t=1}^{T}\sum_{x,a}\frac{q^\star(x)\pi_t(a|x)}{\overline{q}_t(x,a)+\gamma}\cdot\frac{\mathbb{1}_t(x,a)}{\overline{q}_t(x,a)+\gamma} \\ &\le 2\eta H^2\sum_{t=1}^{T}\sum_{x,a}\frac{q^\star(x)\pi_t(a|x)}{\overline{q}_t(x,a)+\gamma}\cdot\frac{q_t(x,a)}{\overline{q}_t(x,a)} + 2\eta H^2 \times \frac{\frac{H}{\gamma}\ln\frac{H}{\delta}}{2\gamma} \\ &\le \frac{H}{4\eta}\ln\frac{H}{\delta} + \sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)\frac{\gamma H}{\overline{q}_t(x,a)+\gamma},\end{aligned}$$

where the third step happens with probability at least $1-\delta$ by Lemma A.2 with $z_t(x,a) = Z_t(x,a) = \frac{q^\star(x)\pi_t(a|x)}{\overline{q}_t(x,a)+\gamma} \le \frac{1}{\gamma}$, and the last step uses $\gamma = 2\eta H$ and $q_t(x,a) \le \overline{q}_t(x,a)$ (which happens with probability at least $1-4\delta$). For the second term in Eq. (21), note that

$$2\eta\sum_{t=1}^{T}\sum_{a\in A}\pi_t(a|x)B_t(x,a)^2 \le \frac{1}{H}\sum_{t=1}^{T}\sum_{a\in A}\pi_t(a|x)B_t(x,a)$$

due to the fact $\eta B_t(x,a) \le \frac{1}{2H}$ by Lemma C.4. Combining everything finishes the proof. $\square$

In Lemma C.3, as required by Lemma A.4, we control the magnitude of $\eta \widehat{Q}_t(x,a)$ and $\eta B_t(x,a)$ by setting γ and η properly, shown in the following technical lemma.

Lemma C.4. *$\eta \widehat{Q}_t(x,a) \le \frac{1}{2}$ and $\eta B_t(x,a) \le \frac{1}{2H}$.*

Proof. Recall that $\gamma = 2\eta H$ and $\eta \le \frac{1}{24H^3}$. Thus,

$$\eta \widehat{Q}_t(x,a) \le \frac{\eta H}{\gamma} = \frac{\eta H}{2\eta H} = \frac{1}{2},$$

$$\eta b_t(x,a) = \frac{3\eta\gamma H + \eta H(\overline{q}_t(x,a) - \underline{q}_t(x,a))}{\overline{q}_t(x,a) + \gamma} \le 3\eta H + \eta H \le \frac{1}{6H^2}.$$

By the definition of $B_t(x,a)$ in Eq. (9), we have

$$\eta B_t(x,a) \le H\left(1+\frac{1}{H}\right)^H \eta \sup_{x',a'} b_t(x',a') \le 3H \times \frac{1}{6H^2} = \frac{1}{2H}.$$

This finishes the proof. $\square$

Now we are ready to prove Theorem 4.1. For convenience, we state the theorem again here and show the proof.

Theorem C.5. *Algorithm 1 ensures that with probability $1-\mathcal{O}(\delta)$, $\mathrm{Reg} = \widetilde{\mathcal{O}}\left(|X|H^2\sqrt{AT} + H^4\right)$.*

Proof. Combining BIAS-1, BIAS-2, REG-TERM, we get that with probability at least $1-\mathcal{O}(\delta)$,

$$\begin{aligned}
&\text{BIAS-1} + \text{BIAS-2} + \text{REG-TERM} \\
&\le \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + \sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)\left(\frac{3\gamma H + H(\overline{q}_t(x,a) - \underline{q}_t(x,a))}{\overline{q}_t(x,a)+\gamma} + \frac{1}{H}B_t(x,a)\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + \sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi^\star(a|x) b_t(x,a) + \frac{1}{H}\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)B_t(x,a),
\end{aligned}$$

which is of the form specified in Eq. (5). By the definition of $B_t(x,a)$ in Eq. (9), we see that Eq. (13) also holds with probability at least $1-\mathcal{O}(\delta)$ for all t, x, a.

Therefore, by Lemma B.1, we can bound the regret as (let $\widehat{P}_t$ be the optimistic transition function chosen in Eq. (9) at episode t)

$$\begin{aligned}
\mathrm{Reg} &= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \sum_{t=1}^{T}\sum_{x,a} q^{\widehat{P}_t,\pi_t}(x,a) b_t(x,a)\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \sum_{t=1}^{T}\sum_{x,a} q^{\widehat{P}_t,\pi_t}(x,a) \frac{H(\overline{q}_t(x,a) - \underline{q}_t(x,a)) + \gamma H}{\overline{q}_t(x,a)+\gamma}\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \sum_{t=1}^{T}\sum_{x,a}\left(H(\overline{q}_t(x,a) - \underline{q}_t(x,a)) + \eta H^2\right)\right) \\
&\hspace{15em} (q^{\widehat{P}_t,\pi_t}(x,a) \le \overline{q}_t(x,a) \text{ and } \gamma = 2\eta H) \\
&\le \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + |X|H^2\sqrt{AT} + \eta|X||A|H^2T\right),
\end{aligned}$$

where the last inequality is due to [15, Lemma 4]. Plugging in the specified value for η, the regret can be further upper bounded by $\widetilde{\mathcal{O}}\left(|X|H^2\sqrt{AT} + H^4\right)$. $\square$

D Details Omitted in Section 5

In this section, our goal is to analyze Algorithm 2 and prove Theorem 5.1. Before conducting regret analysis, we first analyze the GEOMETRICRESAMPLING algorithm in Appendix D.1, which mostly follows [24].

D.1 GEOMETRICRESAMPLING and Its Analysis

The GEOMETRICRESAMPLING algorithm is shown in Algorithm 7, which is almost the same as that in [24] except that we repeat the same procedure for M times and average the outputs (see the extra outer loop). This extra step is added to deal with some technical difficulties in the analysis.

Algorithm 7 GEOMETRICRESAMPLING(t, M, N, γ)

Let $c = \frac{1}{2}$.
for $m = 1, \dots, M$ **do**
  for $n = 1, \dots, N$ **do**
    Generate path $(x_{n,0}, a_{n,0}), \dots, (x_{n,H-1}, a_{n,H-1})$ using policy π_t and the simulator.
    For all h, compute $Y_{n,h} = \gamma I + \phi(x_{n,h}, a_{n,h})\phi(x_{n,h}, a_{n,h})^\top$.
    For all h, compute $Z_{n,h} = \Pi_{j=1}^{n}(I - cY_{j,h})$.
  For all h, set $\widehat{\Sigma}_{t,h}^{+(m)} = cI + c\sum_{n=1}^{N} Z_{n,h}$.
For all h, set $\widehat{\Sigma}_{t,h}^{+} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\Sigma}_{t,h}^{+(m)}$.
return $\widehat{\Sigma}_{t,h}^{+}$ for all $h = 0, \dots, H-1$.

Lemma D.1. *Let* $M = \left\lceil \frac{24\ln(dHT)}{\epsilon^2\gamma^2} \right\rceil$, $N = \left\lceil \frac{2}{\gamma} \ln \frac{1}{\epsilon\gamma} \right\rceil$ *for some* $\epsilon, \gamma > 0$. *Then* GEOMETRICRESAMPLING *(Algorithm 7) with input* (t, M, N, γ) *ensures the following for all* h*:*

$$\left\| \widehat{\Sigma}_{t,h}^{+} \right\|_{\text{op}} \le \frac{1}{\gamma}. \tag{22}$$

$$\left\| \mathbb{E}_t\left[\widehat{\Sigma}_{t,h}^{+} \right] - (\gamma I + \Sigma_{t,h})^{-1} \right\|_{\text{op}} \le \epsilon, \tag{23}$$

$$\left\| \widehat{\Sigma}_{t,h}^{+} - (\gamma I + \Sigma_{t,h})^{-1} \right\|_{\text{op}} \le 2\epsilon, \tag{24}$$

$$\left\| \widehat{\Sigma}_{t,h}^{+} \Sigma_{t,h} \right\|_{\text{op}} \le 1 + 2\epsilon, \tag{25}$$

where $\|\cdot\|_{\text{op}}$ *represents the spectral norm and the last two properties Eq. (24) and Eq. (25) hold with probability at least* $1 - \frac{1}{T^3}$.

Proof. To prove Eq. (22), notice that each one of $\widehat{\Sigma}_{t,h}^{+(m)}$, $m = 1, \dots, M$, is a sum of $N+1$ terms. Furthermore, the n-th term of them ($cZ_{n,h}$ in Algorithm 7) has an operator norm upper bounded by $c(1 - c\gamma)^n$. Therefore,

$$\left\| \widehat{\Sigma}_{t,h}^{+(m)} \right\|_{\text{op}} \le \sum_{n=0}^{N} c(1 - c\gamma)^n \le \frac{1}{\gamma}. \tag{26}$$

Since $\widehat{\Sigma}_{t,h}^{+}$ is an average of $\widehat{\Sigma}_{t,h}^{+(m)}$, this implies Eq. (22).

To show Eq. (23), observe that $\mathbb{E}_t[Y_{n,h}] = \gamma I + \Sigma_{t,h}$ and $\{Y_{n,h}\}_{n=1}^{N}$ are independent. Therefore, we a have

$$\begin{aligned}
\mathbb{E}_t\left[\widehat{\Sigma}_{t,h}^{+} \right] = \mathbb{E}_t\left[\widehat{\Sigma}_{t,h}^{+(m)} \right] &= cI + c\sum_{i=1}^{N} \left(I - c\left(\gamma I + \Sigma_{t,h}\right)\right)^i \\
&= (\gamma I + \Sigma_{t,h})^{-1}\left(I - \left(I - c\left(\gamma I + \Sigma_{t,h}\right)\right)^{N+1}\right)
\end{aligned}$$

where the last step uses the formula: $\left(I+\sum_{i=1}^{N} A^i\right)=(I-A)^{-1}(I-A^{N+1})$ with $A=I-c(\gamma I+\Sigma_{t,h})$. Thus,

$$\left\|\mathbb{E}_t\left[\widehat{\Sigma}^+_{t,h}\right]-(\gamma I+\Sigma_{t,h})^{-1}\right\|_{\text{op}}=\left\|(\gamma I+\Sigma_{t,h})^{-1}\left(I-c\left(\gamma I+\Sigma_{t,h}\right)\right)^{N+1}\right\|_{\text{op}}$$
$$\leq\frac{(1-c\gamma)^{N+1}}{\gamma}\leq\frac{e^{-(N+1)c\gamma}}{\gamma}\leq\epsilon,$$

where the first inequality is by $0\prec I-c(\gamma I+I)\preceq I-c(\gamma I+\Sigma_{t,h})\preceq I-c\gamma I$, and the last inequality is by our choice of N and that $c=\frac{1}{2}$.

To show Eq. (24), we only further need

$$\left\|\widehat{\Sigma}^+_{t,h}-\mathbb{E}_t\left[\widehat{\Sigma}^+_{t,h}\right]\right\|_{\text{op}}\leq\epsilon$$

and combine it with Eq. (23). This can be shown by applying Lemma A.3 with $X_k=\widehat{\Sigma}^{+(k)}_{t,h}-\mathbb{E}_t\left[\widehat{\Sigma}^{+(k)}_{t,h}\right]$, $A_k=\frac{1}{\gamma}I$ (recall Eq. (26) and thus $X_k^2\preceq A_k^2$), $\sigma=\frac{1}{\gamma}$, $\tau=\epsilon$, and $n=M$. This gives the following statement: the event $\left\|\widehat{\Sigma}^+_{t,h}-\mathbb{E}_t\left[\widehat{\Sigma}^+_{t,h}\right]\right\|_{\text{op}}>\epsilon$ holds with probability less than

$$d\exp\left(-M\times\epsilon^2\times\frac{1}{8}\times\gamma^2\right)\leq\frac{1}{d^2H^3T^3}\leq\frac{1}{HT^3}$$

by our choice of M. The conclusion follows by a union bound over h.

To prove Eq. (25), observe that with Eq. (24), we have

$$\left\|\widehat{\Sigma}^+_{t,h}\Sigma_{t,h}\right\|_{\text{op}}\leq\left\|(\gamma I+\Sigma_{t,h})^{-1}\Sigma_{t,h}\right\|_{\text{op}}+\left\|\left(\widehat{\Sigma}^+_{t,h}-(\gamma I+\Sigma_{t,h})^{-1}\right)\Sigma_{t,h}\right\|_{\text{op}}\leq 1+2\epsilon$$

since $\|\Sigma_{t,h}\|_{\text{op}}\leq 1$. □

D.2 Regret Analysis

In the analysis, we require that $\pi_t(a|x)$ and $B_t(x,a)$ be defined for all x,a,t, but in Algorithm 2, they are only explicitly defined if the learner has ever visited state x. Below, we construct a virtual process that is equivalent to Algorithm 2, but with all $\pi_t(a|x)$ and $B_t(x,a)$ well-defined.

Imagine a virtual process where at the end of episode t (a moment when $\widehat{\Sigma}^+_t$ has been defined), $\textsc{Bonus}(t,x,a)$ is called once for every (x,a), in an order from layer $H-1$ to layer 0. Observe that within $\textsc{Bonus}(t,x,a)$, other $\textsc{Bonus}(t',x',a')$ might be called, but either $t'<t$, or x' is in a later layer. Therefore, in this virtual process, every recursive call will soon be returned in the third line of Algorithm 3 because they have been called previously and the values of them are already determined. Given that $\textsc{Bonus}(t,x,a)$ are all called once, at the beginning of episode $t+1$, π_{t+1} will be well-defined for all states since it only depends on $\textsc{Bonus}(t',x',a')$ with $t'\leq t$ and other quantities that are well-defined before episode $t+1$.

Comparing the virtual process and the real process, we see that the virtual process calculates all entries of $\textsc{Bonus}(t,x,a)$, while the real process only calculates a subset of them that are necessary for constructing π_t and $\widehat{\Sigma}^+_t$. However, they define exactly the same policies as long as the random seeds we use for each entry of $\textsc{Bonus}(t,x,a)$ are the same for both processes. Therefore, we can define $B_t(x,a)$ unambiguously as the value returned by $\textsc{Bonus}(t,x,a)$ in the virtual process, and $\pi_t(a|x)$ as shown in (11) with $\textsc{Bonus}(\tau,x,a)$ replaced by $B_\tau(x,a)$.

Now, we follow the exactly same regret decomposition as described in Section 4 (see also Eq. (17)), with the new definition of $\widehat{Q}_t(x,a)\triangleq\phi(x,a)^\top\widehat{\theta}_{t,h}$ (for $x\in X_h$) and $B_t(x,a)$ described above, and then bound $\mathbb{E}[\textsc{Bias-1}+\textsc{Bias-2}]$ and $\mathbb{E}[\textsc{Reg-Term}]$ in Lemma D.2 and Lemma D.3 respectively.

Lemma D.2. *If $\beta\leq H$, then $\mathbb{E}[\textsc{Bias-1}+\textsc{Bias-2}]$ is upper bounded by*

$$\frac{\beta}{4}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{(x,a)\in X_h\times A}q^\star(x)\big(\pi_t(a|x)+\pi^\star(a|x)\big)\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}}\right]+\mathcal{O}\left(\frac{\gamma dH^3T}{\beta}+\epsilon H^2T\right).$$

Proof of Lemma D.2. Consider a specific (t, x, a). Let h be such that $x \in X_h$. Then we proceed as

$$
\begin{aligned}
&\mathbb{E}_t\left[Q_t^{\pi_t}(x,a) - \widehat{Q}_t(x,a)\right] \\
&= \phi(x,a)^\top \left(\theta_{t,h}^{\pi_t} - \mathbb{E}_t\left[\widehat{\theta}_{t,h}\right]\right) \\
&= \phi(x,a)^\top \left(\theta_{t,h}^{\pi_t} - \mathbb{E}_t\left[\widehat{\Sigma}_{t,h}^{+}\right]\mathbb{E}_t\left[\phi(x_{t,h},a_{t,h})L_{t,h}\right]\right) && \text{(definition of } \widehat{\theta}_{t,h}) \\
&= \phi(x,a)^\top \left(\theta_{t,h}^{\pi_t} - (\gamma I + \Sigma_{t,h})^{-1}\mathbb{E}_t\left[\phi(x_{t,h},a_{t,h})L_{t,h}\right]\right) + \mathcal{O}(\epsilon H) \\
&\qquad \text{(by Eq. (23) of Lemma D.1 and that } \|\phi(x,a)\| \le 1 \text{ for all } x, a \text{ and } L_{t,h} \le H) \\
&= \phi(x,a)^\top \left(\theta_{t,h}^{\pi_t} - (\gamma I + \Sigma_{t,h})^{-1}\Sigma_{t,h}\theta_{t,h}^{\pi_t}\right) + \mathcal{O}(\epsilon H) && (\mathbb{E}[L_{t,h}] = \phi(x_{t,h},a_{t,h})^\top \theta_{t,h}^{\pi_t}) \\
&= \gamma\phi(x,a)^\top(\gamma I + \Sigma_{t,h})^{-1}\theta_{t,h}^{\pi_t} + \mathcal{O}(\epsilon H) && (\theta_{t,h}^{\pi_t} = (\gamma I + \Sigma_{t,h})^{-1}(\gamma I + \Sigma_{t,h})\,\theta_{t,h}^{\pi_t}) \\
&\le \gamma\|\phi(x,a)\|^2_{(\gamma I + \Sigma_{t,h})^{-1}}\|\theta_{t,h}^{\pi_t}\|^2_{(\gamma I + \Sigma_{t,h})^{-1}} + \mathcal{O}(\epsilon H) && \text{(Cauchy-Schwarz inequality)} \\
&\le \frac{\beta}{4}\|\phi(x,a)\|^2_{(\gamma I + \Sigma_{t,h})^{-1}} + \frac{\gamma^2}{\beta}\|\theta_{t,h}^{\pi_t}\|^2_{(\gamma I + \Sigma_{t,h})^{-1}} + \mathcal{O}(\epsilon H) && \text{(AM-GM inequality)} \\
&\le \frac{\beta}{4}\mathbb{E}_t\left[\|\phi(x,a)\|^2_{\widehat{\Sigma}_{t,h}^{+}}\right] + \frac{\gamma d H^2}{\beta} + \mathcal{O}\left(\epsilon(H+\beta)\right) && (27)
\end{aligned}
$$

where in the last inequality we use Eq. (23) again and also $\|\theta_{t,h}^{\pi}\|^2 \le dH^2$ according to Assumption 1. Summing the above over t, x, a with weights $q^\star(x)\pi_t(a|x)$, and taking expectation, we get

$$
\mathbb{E}[\textsc{Bias-1}] \le \frac{\beta}{4}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{(x,a)\in X_h\times A} q^\star(x)\pi_t(a|x)\|\phi(x,a)\|^2_{\widehat{\Sigma}_{t,h}^{+}}\right] + \mathcal{O}\left(\frac{\gamma d H^3 T}{\beta} + \epsilon H^2 T\right).
$$
$$\text{(using } \beta \le H)$$

By the same argument, we can show that $\mathbb{E}_t[\widehat{Q}_t(x,a) - Q_t^{\pi_t}(x,a)]$ is also upper bounded by the right-hand side of Eq. (27), and thus

$$
\mathbb{E}[\textsc{Bias-2}] \le \frac{\beta}{4}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{(x,a)\in X_h\times A} q^\star(x)\pi^\star(a|x)\|\phi(x,a)\|^2_{\widehat{\Sigma}_{t,h}^{+}}\right] + \mathcal{O}\left(\frac{\gamma d H^3 T}{\beta} + \epsilon H^2 T\right).
$$

Summing them up finishes the proof. □

Lemma D.3. *If $\eta\beta \le \frac{\gamma}{12H^2}$ and $\eta \le \frac{\gamma}{2H}$, then $\mathbb{E}[\textsc{Reg-Term}]$ is upper bounded by*

$$
\frac{H\ln|A|}{\eta} + 2\eta H^2\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{(x,a)\in X_h\times A} q^\star(x)\pi_t(a|x)\|\phi(x,a)\|^2_{\widehat{\Sigma}_{t,h}^{+}}\right]
$$
$$
+ \frac{1}{H}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)B_t(x,a)\right] + \mathcal{O}\left(\eta\epsilon H^3 T + \frac{\eta H^3}{\gamma^2 T^2}\right).
$$

Proof of Lemma D.3. Again, we will apply the regret bound of the exponential weight algorithm Lemma A.4 to each state. We start by checking the required condition: $\eta|\phi(x,a)^\top\widehat{\theta}_{\tau,h} - B_t(x,a)| \le 1$. This can be seen by that

$$
\begin{aligned}
\eta\left|\phi(x,a)^\top\widehat{\theta}_{\tau,h}\right| &= \eta\left|\phi(x,a)^\top\widehat{\Sigma}_{t,h}^{+}\phi(x_{t,h},a_{t,h})L_{t,h}\right| \\
&\le \eta \times \left\|\widehat{\Sigma}_{t,h}^{+}\right\|_{\text{op}} \times L_{t,h} \le \frac{\eta H}{\gamma} \le \frac{1}{2}, \qquad \text{(Eq. (22) and the condition } \eta \le \tfrac{\gamma}{2H})
\end{aligned}
$$

and that by the definition of $\textsc{Bonus}(t,x,a)$, we have

$$
\eta B_t(x,a) \le \eta \times H\left(1 + \frac{1}{H}\right)^H \times 2\beta\sup_{x,a,h}\|\phi(x,a)\|^2_{\widehat{\Sigma}_{t,h}^{+}} \le \frac{6\eta\beta H}{\gamma} \le \frac{1}{2H}, \tag{28}
$$

where the last inequality is by Eq. (22) again and the condition $\eta\beta \le \frac{\gamma}{12H^2}$.

Thus, by Lemma A.4, we have for any x,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\left(\pi_t(a|x)-\pi^\star(a|x)\right)\widehat{Q}_t(x,a)\right] \\
&\le \frac{\ln|A|}{\eta} + 2\eta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\pi_t(a|x)\widehat{Q}_t(x,a)^2\right] + 2\eta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\pi_t(a|x)B_t(x,a)^2\right]. \qquad (29)
\end{aligned}$$

The last term in Eq. (29) can be upper bounded by $\mathbb{E}\left[\frac{1}{H}\sum_{t=1}^{T}\sum_a \pi_t(a|x)B_t(x,a)\right]$ because $\eta B_t(x,a) \le \frac{1}{2H}$ as we verified in Eq. (28). To bound the second term in Eq. (29), we use the following: for $(x,a)\in X_h\times A$,

$$\begin{aligned}
\mathbb{E}_t\left[\widehat{Q}_t(x,a)^2\right] &\le H^2\mathbb{E}_t\left[\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\phi(x_{t,h},a_{t,h})\phi(x_{t,h},a_{t,h})^\top\widehat{\Sigma}^+_{t,h}\phi(x,a)\right] \\
&= H^2\mathbb{E}_t\left[\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\Sigma_{t,h}\widehat{\Sigma}^+_{t,h}\phi(x,a)\right] \\
&\le H^2\mathbb{E}_t\left[\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\Sigma_{t,h}\left(\gamma I+\Sigma_{t,h}\right)^{-1}\phi(x,a)\right] + \mathcal{O}\left(\epsilon H^2+\frac{H^2}{\gamma^2T^3}\right) \qquad (*) \\
&\le H^2\phi(x,a)^\top\left(\gamma I+\Sigma_{t,h}\right)^{-1}\Sigma_{t,h}\left(\gamma I+\Sigma_{t,h}\right)^{-1}\phi(x,a) + \mathcal{O}\left(\epsilon H^2+\frac{H^2}{\gamma^2T^3}\right) \\
&\hspace{30em}\text{(by Eq. (23))} \\
&\le H^2\phi(x,a)^\top\left(\gamma I+\Sigma_{t,h}\right)^{-1}\phi(x,a) + \mathcal{O}\left(\epsilon H^2+\frac{H^2}{\gamma^2T^3}\right) \\
&\le H^2\mathbb{E}_t\left[\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\phi(x,a)\right] + \mathcal{O}\left(\epsilon H^2+\frac{H^2}{\gamma^2T^3}\right) \qquad \text{(by Eq. (23) again)} \\
&= H^2\mathbb{E}_t\left[\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}}\right] + \mathcal{O}\left(\epsilon H^2+\frac{H^2}{\gamma^2T^3}\right)
\end{aligned}$$

where $(*)$ is because by Eq. (24) and Eq. (25), $\|(\gamma I+\Sigma_{t,h})^{-1}-\widehat{\Sigma}^+_{t,h}\|_{\text{op}} \le 2\epsilon$ and $\|\widehat{\Sigma}^+_{t,h}\Sigma_{t,h}\|_{\text{op}} \le 1+2\epsilon$ hold with probability $1-\frac{1}{T^3}$; for the remaining probability, we upper bound $H^2\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\Sigma_{t,h}\widehat{\Sigma}^+_{t,h}\phi(x,a)$ by $\frac{H^2}{\gamma^2}$. Combining them with Eq. (29) and summing over states with weights $q^\star(x)$ finishes the proof. ☐

With Lemma D.2 and Lemma D.3, we can now prove Theorem 5.1.

Proof of Theorem 5.1. Combining Lemma D.2 and Lemma D.3, we get (under the required conditions of the parameters):

$$\begin{aligned}
&\mathbb{E}\left[\textsc{Bias-1}+\textsc{Bias-2}+\textsc{Reg-Term}\right] \\
&\le \mathcal{O}\left(\frac{H\ln|A|}{\eta}+\frac{\gamma dH^3T}{\beta}+\epsilon H^2T+\eta\epsilon H^3T+\frac{\eta H^3}{\gamma^2T^2}\right) \\
&\quad + \left(2\eta H^2+\frac{\beta}{4}\right)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{(x,a)\in X_h\times A} q^\star(x)\Big(\pi_t(a|x)+\pi^\star(a|x)\Big)\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}}\right] \\
&\quad + \frac{1}{H}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)B_t(x,a)\right].
\end{aligned}$$

We see that Eq. (15) is satisfied in expectation as long as we have $2\eta H^2+\frac{\beta}{4} \le \beta$ and define $b_t(x,a) \triangleq \beta\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}} + \beta\sum_{a'}\pi_t(a'|x)\|\phi(x,a')\|^2_{\widehat{\Sigma}^+_{t,h}}$ (for $x\in X_h$). By the definition of Algorithm 3, Eq. (14) is also satisfied with this choice of $b_t(x,a)$. Therefore, we can apply

Lemma B.2 to obtain a regret bound. To simply the presentation, we first pick $\epsilon = \frac{1}{H^3T}$ so that all ϵ-related terms become $\mathcal{O}(1)$. Then we have

$$
\begin{aligned}
&\mathbb{E}[\mathrm{Reg}] \\
&= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\gamma d H^3 T}{\beta} + \frac{\eta H^3}{\gamma^2 T^2} + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,a} q_t(x,a) b_t(x,a)\right]\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\gamma d H^3 T}{\beta} + \frac{\eta H^3}{\gamma^2 T^2} + \beta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h}\sum_{(x,a)\in X_h\times A} q_t(x,a)\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}}\right]\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\gamma d H^3 T}{\beta} + \frac{\eta H^3}{\gamma^2 T^2} + \beta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h}\sum_{(x,a)\in X_h\times A} q_t(x,a)\|\phi(x,a)\|^2_{(\gamma I+\Sigma_{t,h})^{-1}}\right]\right) \\
&\hspace{22em} \text{(Eq. (23) and } \beta \le H) \\
&= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\gamma d H^3 T}{\beta} + \frac{\eta H^3}{\gamma^2 T^2} + \beta d H T\right),
\end{aligned}
$$

where the last step uses the fact

$$
\begin{aligned}
\mathbb{E}_t\left[\sum_{h}\sum_{(x,a)\in X_h\times A} q_t(x,a)\|\phi(x,a)\|^2_{(\gamma I+\Sigma_{t,h})^{-1}}\right] &\le \mathbb{E}_t\left[\sum_{h}\sum_{(x,a)\in X_h\times A} q_t(x,a)\|\phi(x,a)\|^2_{\Sigma^{-1}_{t,h}}\right] \\
&= \sum_h \left\langle \Sigma_{t,h}, \Sigma^{-1}_{t,h}\right\rangle = dH. \qquad (30)
\end{aligned}
$$

Finally, choosing the parameters under the specified constraints as:

$$
\begin{aligned}
&\gamma = (dT)^{-\frac{2}{3}}, \qquad \beta = H(dT)^{-\frac{1}{3}}, \qquad \epsilon = \frac{1}{H^3 T}, \\
&\eta = \min\left\{\frac{\gamma}{2H}, \frac{3\beta}{8H^2}, \frac{\gamma}{12\beta H^2}\right\},
\end{aligned}
$$

we further bound the regret by $\widetilde{\mathcal{O}}\left(H^2(dT)^{\frac{2}{3}} + H^4(dT)^{\frac{1}{3}}\right)$. □

E Details for Linear-Q with a Simulator and an Exploratory Policy

The main algorithm (Algorithm 8) follows the same idea as Algorithm 2. The main difference is that we can leverage π_0 to perform exploration. To do so, in Step 1 of the algorithm, we draw a Bernoulli random variable $Y_t \sim \textsc{Bernoulli}(\delta_e)$ (for some $\delta_e \in (0,1)$) to indicate whether in this round the learner should use π_0. If Y_t is 1, then the learner further randomly draw h_t^* from $0,\ldots,H-1$. Then she walks from x_0 to layer h_t^* using π_0, and then continues with π_t to the end. In this way, the learner can explicitly explore the state space on every layer, which facilitates estimating $\theta^{\pi_t}_{t,h}$ with less bias.

Because we mix the exploration into the policy, we perform a slightly different procedure $\textsc{GeometricResampling-Mixture}$ in Step 2, which does not incorporate the γ parameter as in $\textsc{GeometricResampling}$. Instead, it will estimate the inverse of $\Sigma^{\text{mix}}_{t,h} = (1-\delta_e)\Sigma_{t,h} + \delta_e\Sigma^{\pi_0}_h$ where $\Sigma_{t,h}$ is the covariance matrix under π_t and $\Sigma^{\pi_0}_h$ is the covariance matrix under π_0.

The new construction of $\widehat{\theta}_{t,h}$ in Step 3 makes it an estimator of $\theta^{\pi_t}_{t,h}$ with low error. To see this, observe that

$$
\begin{aligned}
&\mathbb{E}_t\left[\left((1-Y_t) + Y_t H\mathbb{1}[h=h_t^*]\right)\phi(x_{t,h},a_{t,h})L_{t,h}\right] \\
&= \delta_e\mathbb{E}_t\left[H\mathbb{1}[h=h_t^*]\phi(x_{t,h},a_{t,h})L_{t,h} \mid Y_t = 1\right] + (1-\delta_e)\mathbb{E}_t\left[\phi(x_{t,h},a_{t,h})L_{t,h} \,\middle|\, Y_t = 0\right] \\
&= \delta_e\mathbb{E}_t\left[H\mathbb{1}[h=h_t^*]\phi(x_{t,h},a_{t,h})\phi(x_{t,h},a_{t,h})^\top\theta^{\pi_t}_{t,h} \,\middle|\, Y_t = 1\right] \\
&\qquad + (1-\delta_e)\mathbb{E}_t\left[\phi(x_{t,h},a_{t,h})\phi(x_{t,h},a_{t,h})^\top\theta^{\pi_t}_{t,h} \,\middle|\, Y_t = 0\right]
\end{aligned}
$$

Algorithm 8 Policy Optimization with Dilated Bonuses (Linear-Q Case with an Exploratory Policy)

parameters: $\lambda_{\min}, \beta, \eta, \epsilon, \delta_e$, $M = \left\lceil \frac{96\ln(dHT)\ln^2(\frac{1}{\epsilon\delta_e\lambda_{\min}})}{\epsilon^2\delta_e^2\lambda_{\min}^2} \right\rceil$, $N = \left\lceil \frac{2}{\delta_e\lambda_{\min}} \ln \frac{1}{\epsilon\delta_e\lambda_{\min}} \right\rceil$.

for $t = 1, 2, \ldots, T$ **do**

Step 1: Interact with the environment. Let π_t be defined such that for each $x \in X_h$,

$$\pi_t(a|x) \propto \exp\left(-\eta \sum_{\tau=1}^{t-1} \left(\phi(x,a)^\top \widehat{\theta}_{\tau,h} - \textsc{Bonus}(\tau, x, a)\right)\right). \tag{32}$$

Draw $Y_t \sim \textsc{Bernoulli}(\delta_e)$.

if $Y_t = 1$ **then**

Draw $h_t^* \sim \text{Uniform}\{0, \ldots, H-1\}$.

Execute π_0 in steps $0, \ldots, h_t^* - 1$; continue with π_t in steps $h_t^*, \ldots, H-1$.

else Execute π_t.

Obtain trajectory $\{(x_{t,h}, a_{t,h}, \ell_t(x_{t,h}, a_{t,h}))\}_{h=0}^{H-1}$.

Step 2: Construct covariance matrix inverse estimators.

$$\left\{\widehat{\Sigma}^+_{t,h}\right\}_{h=0}^{H-1} = \textsc{GeometricResampling-Mixture}\,(t, M, N)\,. \qquad \text{(see Algorithm 9)}$$

Step 3: Construct Q-function weight estimators. For all $h = 0, \ldots, H-1$,

$$\widehat{\theta}_{t,h} = \widehat{\Sigma}^+_{t,h}\left((1 - Y_t) + Y_t H \mathbb{1}[h = h_t^*]\right)\phi(x_{t,h}, a_{t,h}) L_{t,h}, \quad \text{where } L_{t,h} = \sum_{i=h}^{H-1} \ell_t(x_{t,i}, a_{t,i}).$$

$$\begin{aligned}
&= \delta_e \mathbb{E}_t\left[H\mathbb{1}[h = h_t^*]\right] \Sigma_h^{\pi_0}\theta^{\pi_t}_{t,h} + (1-\delta_e)\mathbb{E}_t\left[\Sigma_{t,h}\theta^{\pi_t}_{t,h}\right] \\
&= \left(\delta_e \Sigma_h^{\pi_0} + (1-\delta_e)\Sigma_{t,h}\right)\theta^{\pi_t}_{t,h} \\
&= \Sigma^{\text{mix}}_{t,h}\theta^{\pi_t}_{t,h}
\end{aligned} \tag{31}$$

and thus

$$\mathbb{E}_t\left[\widehat{\theta}_{t,h}\right] = \mathbb{E}_t\left[\widehat{\Sigma}^+_{t,h}\left((1-Y_t) + Y_t H\mathbb{1}[h = h_t^*]\right)\phi(x_{t,h}, a_{t,h})L_{t,h}\right] = \mathbb{E}_t\left[\widehat{\Sigma}^+_{t,h}\right]\Sigma^{\text{mix}}_{t,h}\theta^{\pi_t}_{t,h} \approx \theta^{\pi_t}_{t,h},$$

where the last step is because $\widehat{\Sigma}^+_{t,h}$ is approximately the inverse of $\Sigma^{\text{mix}}_{t,h}$.

In Appendix E.1, we first present the algorithm GeometricResampling-Mixture and its analysis (similar to Lemma D.1). Then in Appendix E.2, we perform regret analysis for Algorithm 8.

E.1 GeometricResampling-Mixture

Lemma E.1. *Let* $M = \left\lceil \frac{96\ln(dHT)\ln^2(\frac{1}{\epsilon\delta_e\lambda})}{\epsilon^2\delta_e^2\lambda^2} \right\rceil$, $N = \left\lceil \frac{2}{\delta_e\lambda}\ln\frac{1}{\epsilon\delta_e\lambda} \right\rceil$ *for some* $\epsilon > 0$. *Let* $\Sigma_{t,h} = \mathbb{E}_{\pi_t}[\phi(x_h,a_h)\phi(x_h,a_h)^\top]$ *and* $\Sigma_h^{\pi_0} = \mathbb{E}_{\pi_0}[\phi(x_h,a_h)\phi(x_h,a_h)^\top]$ *and* $\Sigma^{mix}_{t,h} = (1 - \delta_e)\Sigma_{t,h} + \delta_e\Sigma_h^{\pi_0}$. *Suppose that* $\lambda > 0$ *is a lower bound for the minimum eigenvalue of* $\Sigma_h^{\pi_0}$. *Then* GeometricResampling-Mixture *(Algorithm 9) with input* (t, M, N) *ensures the following for all* h*:*

$$\left\|\widehat{\Sigma}^+_{t,h}\right\|_{\text{op}} \le \frac{2}{\delta_e\lambda}\ln\frac{1}{\epsilon\delta_e\lambda}. \tag{33}$$

$$\left\|\mathbb{E}_t\left[\widehat{\Sigma}^+_{t,h}\right] - (\Sigma^{mix}_{t,h})^{-1}\right\|_{\text{op}} \le \epsilon, \tag{34}$$

$$\left\|\widehat{\Sigma}^+_{t,h} - (\Sigma^{mix}_{t,h})^{-1}\right\|_{\text{op}} \le 2\epsilon. \tag{35}$$

$$\left\|\widehat{\Sigma}^+_{t,h}\Sigma^{mix}_{t,h}\right\|_{\text{op}} \le 1 + 2\epsilon, \tag{36}$$

Algorithm 9 GEOMETRICRESAMPLING-MIXTURE(t, M, N)

Let $c = \frac{1}{2}$.
for $m = 1, \dots, M$ **do**
  for $n = 1, \dots, N$ **do**
    With probability $1-\delta_e$, generate path $(x_{n,0}, a_{n,0}), \dots, (x_{n,H-1}, a_{n,H-1})$ using π_t; otherwise, generate it using π_0.
    For all h, compute $Y_{n,h} = \phi(x_{n,h}, a_{n,h})\phi(x_{n,h}, a_{n,h})^\top$.
    For all h, compute $Z_{n,h} = \Pi_{j=1}^{n}(I - cY_{j,h})$.
  For all h, set $\widehat{\Sigma}_{t,h}^{+(m)} = cI + c\sum_{n=1}^{N} Z_{n,h}$.
For all h, set $\widehat{\Sigma}_{t,h}^{+} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\Sigma}_{t,h}^{+(m)}$.
return $\widehat{\Sigma}_{t,h}^{+}$ for all $h = 0, \dots, H-1$.

where $\|\cdot\|_{\text{op}}$ *represents the spectral norm and the last two properties* Eq. (35) *and* Eq. (36) *hold with probability at least* $1 - \frac{1}{T^3}$.

Proof. To prove Eq. (33), notice that each one of $\widehat{\Sigma}_{t,h}^{+(m)}$, $m = 1, \dots, M$, is a sum of $N+1$ terms. Furthermore, the n-th term of them ($cZ_{n,h}$ in Algorithm 9) has an operator norm upper bounded by c. Therefore,

$$\left\|\widehat{\Sigma}_{t,h}^{+(m)}\right\|_{\text{op}} \le c(N+1) = \frac{1}{2}(N+1) \le \frac{2}{\delta_e\lambda}\ln\frac{1}{\epsilon\delta_e\lambda}.$$

Since $\widehat{\Sigma}_{t,h}^{+}$ is an average of $\widehat{\Sigma}_{t,h}^{+(m)}$, this implies Eq. (33).

To show Eq. (34), observe that

$$\begin{aligned}\mathbb{E}_t\left[\widehat{\Sigma}_{t,h}^{+}\right] = \mathbb{E}_t\left[\widehat{\Sigma}_{t,h}^{+(m)}\right] &= cI + c\sum_{i=1}^{N}\left(I - c\Sigma_{t,h}^{\text{mix}}\right)^i \\ &= \left(\Sigma_{t,h}^{\text{mix}}\right)^{-1}\left(I - \left(I - c\Sigma_{t,h}^{\text{mix}}\right)^{N+1}\right)\end{aligned}$$

where we use the formula: $\left(I + \sum_{i=1}^{N} A^i\right) = (I-A)^{-1}(I - A^{N+1})$ with $A = I - c\Sigma_{t,h}^{\text{mix}}$.

Thus,

$$\begin{aligned}\left\|\mathbb{E}_t\left[\widehat{\Sigma}_{t,h}^{+}\right] - (\Sigma_{t,h}^{\text{mix}})^{-1}\right\|_{\text{op}} &= \left\|\left(\Sigma_{t,h}^{\text{mix}}\right)^{-1}\left(I - c\Sigma_{t,h}^{\text{mix}}\right)^{N+1}\right\|_{\text{op}} \\ &\le \frac{(1 - c\delta_e\lambda)^{N+1}}{\delta_e\lambda} \le \frac{e^{-(N+1)c\delta_e\lambda}}{\delta_e\lambda} \le \epsilon,\end{aligned}$$

where the last inequality is by our choice of N and that $c = \frac{1}{2}$.

To show Eq. (35), we only further need

$$\left\|\widehat{\Sigma}_{t,h}^{+} - \mathbb{E}_t\left[\widehat{\Sigma}_{t,h}^{+}\right]\right\|_{\text{op}} \le \epsilon$$

and combine it with Eq. (34). This can be shown by applying Lemma A.3 with $X_k = \widehat{\Sigma}_{t,h}^{+(k)}$, $\sigma = \frac{2}{\delta_e\lambda}\ln\frac{1}{\epsilon\delta_e\lambda}$, $\tau = \epsilon$, and $n = M$ (see the proof for Eq. (24) for the reason). This gives the following statement: the event $\left\|\widehat{\Sigma}_{t,h}^{+} - \mathbb{E}_t\left[\widehat{\Sigma}_{t,h}^{+}\right]\right\|_{\text{op}} > \epsilon$ holds with probability less than

$$d\exp\left(-M \times \epsilon^2 \times \frac{1}{8} \times \frac{\delta_e^2\lambda^2}{4\ln^2\frac{1}{\epsilon\delta_e\lambda}}\right) \le \frac{1}{d^2H^3T^3} \le \frac{1}{HT^3}$$

by our choice of M. The conclusion follows by a union bound over h.

To prove Eq. (36), observe that with Eq. (35), we have

$$\left\|\widehat{\Sigma}_{t,h}^{+}\Sigma_{t,h}^{\text{mix}}\right\|_{\text{op}} \le \left\|(\Sigma_{t,h}^{\text{mix}})^{-1}\Sigma_{t,h}^{\text{mix}}\right\|_{\text{op}} + \left\|\left(\widehat{\Sigma}_{t,h}^{+} - (\Sigma_{t,h}^{\text{mix}})^{-1}\right)\Sigma_{t,h}^{\text{mix}}\right\|_{\text{op}} \le 1 + 2\epsilon$$

since $\left\|\Sigma_{t,h}^{\text{mix}}\right\|_{\text{op}} \le 1$. ◻

E.2 Regret Analysis

The analysis follows the same outline discussed in Section D.2. In particular, we define $B_t(x,a)$ for all t, x, a again using the same virtual process, and then we follow the same regret decomposition as in Eq. (17), with $\widehat{Q}_t(x,a) \triangleq \phi(x,a)^\top \widehat{\theta}_{t,h}$ (for $x \in X_h$). We then bound $\mathbb{E}[\textsc{Bias-1} + \textsc{Bias-2}]$ and $\mathbb{E}[\textsc{Reg-Term}]$ in Lemma E.2 and Lemma E.3 respectively.

Lemma E.2. $\mathbb{E}[\textsc{Bias-1} + \textsc{Bias-2}] = \mathcal{O}(\epsilon H^3 T)$.

Proof. Consider a specific (t, x, a). Let h be such that $x \in X_h$. Then we have

$$
\begin{aligned}
&\mathbb{E}_t\left[Q_t^{\pi_t}(x,a) - \widehat{Q}_t(x,a)\right] \\
&= \phi(x,a)^\top \left(\theta_{t,h}^{\pi_t} - \mathbb{E}_t\left[\widehat{\theta}_{t,h}\right]\right) \\
&= \phi(x,a)^\top \left(\theta_{t,h}^{\pi_t} - \mathbb{E}_t\left[\widehat{\Sigma}_{t,h}^{+}\right]\mathbb{E}_t\left[\Big((1-Y_t) + Y_t\mathbb{1}[h_t^* = h]H\Big)\phi(x_{t,h}, a_{t,h})L_{t,h}\right]\right) + \mathcal{O}(\epsilon H^2) \\
&= \phi(x,a)^\top \left(\theta_{t,h}^{\pi_t} - (\Sigma_{t,h}^{\text{mix}})^{-1}\mathbb{E}_t\left[\Big((1-Y_t) + Y_t\mathbb{1}[h_t^* = h]H\Big)\phi(x_{t,h}, a_{t,h})L_{t,h}\right]\right) + \mathcal{O}(\epsilon H^2) \\
&\qquad\qquad\qquad\qquad \text{(by Lemma E.1 and that } \|\phi(x,a)\| \le 1 \text{ for all } x, a \text{ and } L_{t,h} \le H) \\
&= \phi(x,a)^\top \left(\theta_{t,h}^{\pi_t} - (\Sigma_{t,h}^{\text{mix}})^{-1}\Sigma_{t,h}^{\text{mix}}\theta_{t,h}^{\pi_t}\right) + \mathcal{O}(\epsilon H^2) \qquad \text{(Eq. (31))} \\
&= \mathcal{O}\left(\epsilon H^2\right).
\end{aligned}
$$

Similarly, one can show $\mathbb{E}_t\left[\widehat{Q}_t(x,a) - Q_t^{\pi_t}(x,a)\right] = \mathcal{O}\left(\epsilon H^2\right)$. Summing them up over t, x, a with weights $q^\star(x)\pi^\star(a|x)$ and $q^\star(x)\pi_t(a|x)$ respectively finishes the proof. $\square$

Lemma E.3. *If* $\eta\beta \le \frac{\delta_e \lambda_{\min}}{24H^2 \ln(\frac{1}{\epsilon\delta_e\lambda_{\min}})}$ *and* $\eta \le \frac{\delta_e\lambda_{\min}}{4H^2\ln(\frac{1}{\epsilon\delta_e\lambda_{\min}})}$, *then* $\mathbb{E}[\textsc{Reg-Term}]$ *is upper bounded by*

$$
\frac{H\ln|A|}{\eta} + 2\eta H^3 \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{(x,a)\in X_h\times A} q^\star(x)\pi_t(a|x)\|\phi(x,a)\|^2_{\widehat{\Sigma}_{t,h}^{+}}\right] + \frac{1}{H}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)B_t(x,a)\right] + \widetilde{\mathcal{O}}\left(\eta\epsilon H^4 T + \frac{\eta H^4}{\delta_e^2\lambda_{\min}^2 T^2}\right).
$$

Proof. The proof is similar to that of Lemma D.3. Again, we will apply the regret bound of the exponential weight algorithm Lemma A.4 for each state. We start by checking the required condition: $\eta|\phi(x,a)^\top\widehat{\theta}_{\tau,h} - B_t(x,a)| \le 1$. This can be seen by

$$
\begin{aligned}
\eta\left|\phi(x,a)^\top\widehat{\theta}_{\tau,h}\right| &= \eta\left|\phi(x,a)^\top\widehat{\Sigma}_{t,h}^{+}\phi(x_{t,h}, a_{t,h})L_{t,h}\right| \times ((1-Y_t) + Y_t\mathbb{1}[h = h^*]H) \\
&\le \eta \times \left\|\widehat{\Sigma}_{t,h}^{+}\right\|_{\text{op}} \times L_{t,h} \times H \\
&\le \eta \times \frac{2}{\delta_e\lambda_{\min}} \ln\frac{1}{\epsilon\delta_e\lambda_{\min}} \times H^2 && \text{(by Lemma E.1)} \\
&\le \frac{1}{2}, && \text{(condition of the lemma)}
\end{aligned}
$$

and that by the definition of $\textsc{Bonus}(t,x,a)$, we have

$$
\begin{aligned}
\eta B_t(x,a) &\le \eta \times H\left(1 + \frac{1}{H}\right)^H \times 2\beta \sup_{x,a,h}\|\phi(x,a)\|^2_{\widehat{\Sigma}_{t,h}^{+}} \\
&\le 6\eta\beta \times \frac{2H}{\delta_e\lambda_{\min}}\ln\frac{1}{\epsilon\delta_e\lambda_{\min}} && \text{(by Lemma E.1 again)} \\
&\le \frac{1}{2H}, && (37)
\end{aligned}
$$

where the last inequality is by the first condition of the lemma.

Thus, by Lemma A.4, we have for any x,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\left(\pi_t(a|x)-\pi^\star(a|x)\right)\widehat{Q}_t(x,a)\right] \\
&\leq \frac{\ln|A|}{\eta}+2\eta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\pi_t(a|x)\widehat{Q}_t(x,a)^2\right]+2\eta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\pi_t(a|x)B_t(x,a)^2\right]. \qquad (38)
\end{aligned}$$

The last term in Eq. (38) can be upper bounded by $\mathbb{E}\left[\frac{1}{H}\sum_{t=1}^{T}\sum_a \pi_t(a|x)B_t(x,a)\right]$ because $\eta B_t(x,a)\leq \frac{1}{2H}$ as we verified in Eq. (37). To bound the second term in Eq. (38), we use the following: for $(x,a)\in X_h\times A$,

$$\begin{aligned}
&\mathbb{E}_t\left[\widehat{Q}_t(x,a)^2\right] \\
&\leq H^2\mathbb{E}_t\left[\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\left(((1-Y_t)+Y_tH\mathbb{1}[h=h_t^*])^2\phi(x_{t,h},a_{t,h})\phi(x_{t,h},a_{t,h})^\top\right)\widehat{\Sigma}^+_{t,h}\phi(x,a)\right] \\
&= H^2\mathbb{E}_t\left[\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\left((1-\delta_e)\Sigma_{t,h}+\delta_e H\Sigma_h^{\pi_0}\right)\widehat{\Sigma}^+_{t,h}\phi(x,a)\right] \\
&\leq H^3\mathbb{E}_t\left[\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\Sigma^{\text{mix}}_{t,h}\widehat{\Sigma}^+_{t,h}\phi(x,a)\right] \\
&\leq H^3\mathbb{E}_t\left[\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\Sigma^{\text{mix}}_{t,h}(\Sigma^{\text{mix}}_{t,h})^{-1}\phi(x,a)\right]+\widetilde{\mathcal{O}}\left(\epsilon H^3+\frac{H^3}{\delta_e^2\lambda_{\min}^2T^3}\right) \qquad (*) \\
&= H^3\mathbb{E}_t\left[\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}}\right]+\widetilde{\mathcal{O}}\left(\epsilon H^3+\frac{H^3}{\delta_e^2\lambda_{\min}^2T^3}\right), \qquad (39)
\end{aligned}$$

where $(*)$ is because by Eq. (35) and Eq. (36), $\left\|\widehat{\Sigma}^+_{t,h}-(\Sigma^{\text{mix}}_{t,h})^{-1}\right\|_{\text{op}}\leq 2\epsilon$ and $\left\|\widehat{\Sigma}^+_{t,h}\Sigma^{\text{mix}}_{t,h}\right\|_{\text{op}}\leq 1+2\epsilon$ hold with probability $1-\frac{1}{T^3}$; for the remaining probability, we upper bound $H^3\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\Sigma^{\text{mix}}_{t,h}\widehat{\Sigma}^+_{t,h}\phi(x,a)$ by $\frac{4H^3}{\delta_e^2\lambda_{\min}^2}\ln^2\left(\frac{1}{\epsilon\delta_e\lambda_{\min}}\right)$ using Eq. (33).

Combining them with Eq. (38) and summing over states with weights $q^\star(x)$ finishes the proof. $\square$

Finally, we are ready to prove the regret bound.

Proof of Theorem 6.1. Combining Lemma E.2 and Lemma E.3, we see that if we choose $\beta=2\eta H^3$, then

$$\begin{aligned}
&\mathbb{E}[\textsc{Bias-1}+\textsc{Bias-2}+\textsc{Reg-Term}] \\
&\leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\epsilon H^3T+\eta\epsilon H^4T+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2T^2}\right)+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{(x,a)\in X_h\times A}q^\star(x)\pi_t(a|x)b_t(x,a)\right] \\
&\qquad +\frac{1}{H}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,a}q^\star(x)\pi_t(a|x)B_t(x,a)\right].
\end{aligned}$$

Hence, by Lemma B.2, we obtain the following bound, where we first set $\epsilon=\frac{1}{H^4T}$ so that all ϵ-related terms are $\widetilde{\mathcal{O}}(1)$:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}V_t^{\pi_t}(x_0)\right]-\sum_{t=1}^{T}V_t^{\pi^\star}(x_0) \\
&\leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2T^2}+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,a}q_t(x,a)b_t(x,a)\right]\right) \\
&\leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2T^2}+\beta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,a}q_t(x,a)\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}}\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2 T^2}+\beta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,a} q_t(x,a)\|\phi(x,a)\|^2_{(\Sigma_{t,h}^{\text{mix}})^{-1}}\right]\right)\\
&\leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2 T^2}+\frac{\beta}{1-\delta_e}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,a} q_t(x,a)\|\phi(x,a)\|^2_{\Sigma_{t,h}^{-1}}\right]\right)\\
&\leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2 T^2}+\beta dHT\right) && \text{(Eq. (30))}\\
&= \widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2 T^2}+\eta dH^4 T\right). && (\beta = 2\eta H^3)
\end{aligned} \tag{40}
$$

Since we explore with probability δ_e, the final regret is

$$
\mathbb{E}[\text{Reg}] = \widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2 T^2}+\eta dH^4 T+\delta_e HT\right).
$$

Considering the constraints specified in Lemma E.3, we choose the parameters as follows:

$$
\begin{aligned}
\eta &= \min\left\{\frac{\delta_e\lambda_{\min}}{4H^2\ln(\frac{1}{\epsilon\delta_e\lambda_{\min}})}, \sqrt{\frac{\delta_e\lambda_{\min}}{48H^5\ln(\frac{1}{\epsilon\delta_e\lambda_{\min}})}}\right\},\\
\delta_e &= \min\left\{\sqrt{\frac{H^2}{\lambda_{\min}T(\lambda_{\min}dH+1)}}, \frac{1}{2}\right\},\\
\epsilon &= \frac{1}{H^4T}.
\end{aligned}
$$

Then the regret can be bounded by $\mathcal{O}\left(\sqrt{H^5dT}+H^2\sqrt{\frac{T}{\lambda_{\min}}}\right)$. $\square$

F Details for Linear MDP with an Exploratory Policy

The algorithm for linear MDP with an exploratory policy π_0 is presented in Algorithm 10, which is based on the similar idea as Algorithm 8. Instead of changing policies on every episode, the algorithm proceeds in *epochs*, each of which consists of W consecutive episodes, and the algorithm only updates its policy between epochs. We index epoch with k. The definitions of π_k, $\widehat{\Sigma}^+_{k,h}$, $B_k(x,a)$ are analogous to those of π_t, $\widehat{\Sigma}^+_{t,h}$, $B_t(x,a)$ in previous sections.

To deal with the epoch-based update, we define the following quantities (notice that the k-th epoch consists of episodes $(k-1)W+1,\ldots,kW$):

Definition 1.

$$
\begin{aligned}
\overline{\ell}_k(x,a) &\triangleq \frac{1}{W}\sum_{t=(k-1)W+1}^{kW}\ell_t(x,a)\\
\overline{Q}^{\pi}_k(x,a) &\triangleq Q^{\pi}(x,a;\overline{\ell}_k)\\
\overline{\theta}^{\pi}_{k,h} &\triangleq \frac{1}{W}\sum_{t=(k-1)W+1}^{kW}\theta^{\pi}_{t,h}
\end{aligned}
$$

Recall that the main difference between Algorithm 10 and Algorithm 8 is that in Algorithm 10 we use linear function approximation to calculate the bonus. The bonus $B_k(x,a)$ and the *estimated* bonus $\widehat{\Lambda}_k(x,a)$ are defined in Definition 2.

Definition 2.

$$
\begin{aligned}
B_k(x,a) &\triangleq b_k(x,a)+\left(1+\frac{1}{H}\right)\mathbb{E}_{x'\sim P(\cdot|x,a)}\mathbb{E}_{a'\sim\pi_t(\cdot|x')}[B_t(x',a')] && (\text{with } B_k(x_H,a)\triangleq 0)\\
\widehat{B}_k(x,a) &\triangleq b_k(x,a)+\phi(x,a)^\top\widehat{\Lambda}_{k,h} && (\text{for } x\in X_h)
\end{aligned}
$$

where $b_k(x,a)$ and $\widehat{\Lambda}_{k,h}$ are defined in Algorithm 10.

Algorithm 10 Policy Optimization with Dilated Bonuses (Linear MDP with an Exploratory Policy)

Parameters: $\lambda_{\min}, \beta, \eta, \epsilon$, $M = \left\lceil \frac{96 \ln(dHT) \ln^2(\frac{1}{\epsilon \delta_e \lambda_{\min}})}{\epsilon^2 \delta_e^2 \lambda_{\min}^2} \right\rceil$, $N = \left\lceil \frac{2}{\delta_e \lambda_{\min}} \ln \frac{1}{\epsilon \delta_e \lambda_{\min}} \right\rceil$, $W = 2MN$

for $k = 1, 2, \ldots, T/W$ **do**

1) Interact with the environment: Define π_k as the following for $x \in X_h$:

$$\pi_k(a|x) \propto \exp\left(-\eta \sum_{\tau=1}^{k-1} \left(\phi(x,a)^\top \widehat{\theta}_{\tau,h} - \phi(x,a)^\top \widehat{\Lambda}_{\tau,h} - b_\tau(x,a)\right)\right) \tag{41}$$

where

$$b_\tau(x,a) \triangleq \beta \|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{\tau,h}} + \beta \sum_{a'} \pi_\tau(a'|x) \|\phi(x,a')\|^2_{\widehat{\Sigma}^+_{\tau,h}}.$$

Randomly divide $[(k-1)W + 1, \ldots, kW]$ into two parts: S and S', such that $|S| = |S'| = W/2$.

for $t = (k-1)W + 1, \ldots, kW$ **do**

Draw $Y_t \sim$ BERNOULLI(δ_e).

if $Y_t = 1$ *and* $t \in S$ **then** Execute π_0

else if $Y_t = 1$ *and* $t \in S'$ **then**

Draw $h_t^* \sim \text{Uniform}\{0, \ldots, H-1\}$.

Execute π_0 in steps $0, \ldots, h_t^* - 1$; continue with π_t in steps $h_t^*, \ldots, H-1$.

else Execute π_t

Collect trajectories $\{(x_{t,h}, a_{t,h}, \ell_t(x_{t,h}, a_{t,h}))\}_{h=0}^{H-1}$

2) Construct inverse covariance matrix estimators: Use the samples in S to calculate the following (note that $|S| = W/2 = MN$ and the GEOMETRICRESAMPLING-MIXTURE requires exactly MN episodes of samples. We simply view these MN episodes as calls within GEOMETRICRESAMPLING-MIXTURE):

$$\left\{\widehat{\Sigma}^+_{k,h}\right\}_{h=0}^{H-1} = \text{GEOMETRICRESAMPLING-MIXTURE}(k, M, N). \tag{42}$$

3) Construct Q-function estimators: Define for all t, h:

$$L_{t,h} \triangleq \sum_{i=h}^{H-1} \ell_t(x_{t,i}, a_{t,i})$$

and

$$\widehat{\theta}_{k,h} \triangleq \widehat{\Sigma}^+_{k,h} \left(\frac{1}{|S'|} \sum_{t \in S'} ((1 - Y_t) + Y_t H \mathbb{1}[h = h_t^*]) \phi(x_{t,h}, a_{t,h}) L_{t,h}\right). \tag{43}$$

4) Construct bonus function estimators: Define for all t, h:

$$D_{t,h} \triangleq \begin{cases} 0 & \text{if } h = H-1, \\ \displaystyle\sum_{i=h+1}^{H-1} \left(1 + \frac{1}{H}\right)^{i-h} b_t(x_{t,i}, a_{t,i}) & \text{otherwise;} \end{cases}$$

and

$$\widehat{\Lambda}_{k,h} \triangleq \widehat{\Sigma}^+_{k,h} \left(\frac{1}{|S'|} \sum_{t \in S'} ((1 - Y_t) + Y_t H \mathbb{1}[h = h_t^*]) \phi(x_{t,h}, a_{t,h}) D_{t,h}\right). \tag{44}$$

F.1 Regret Analysis

The regret decomposition for this section is slightly different from those in previous sections. Since we also use function approximation on the bonus $B_t(x,a)$, we need to also account for its estimation error, resulting in two extra bias terms:

$$
\begin{aligned}
&\sum_{k=1}^{T/W}\sum_{x} q^\star(x)\left\langle \pi_k(\cdot|x)-\pi^\star(\cdot|x), \overline{Q}_k^{\pi_k}(x,\cdot)-B_k(x,\cdot)\right\rangle \\
&=\underbrace{\sum_{k=1}^{T/W}\sum_{x} q^\star(x)\left\langle \pi_k(\cdot|x), \overline{Q}_k^{\pi_k}(x,\cdot)-\widehat{Q}_k(x,\cdot)\right\rangle}_{\textsc{Bias-1}}+\underbrace{\sum_{k=1}^{T/W}\sum_{x} q^\star(x)\left\langle \pi^\star(\cdot|x), \widehat{Q}_k(x,\cdot)-\overline{Q}_k^{\pi_k}(x,\cdot)\right\rangle}_{\textsc{Bias-2}} \\
&\quad+\underbrace{\sum_{k=1}^{T/W}\sum_{x} q^\star(x)\left\langle \pi_k(\cdot|x), \widehat{B}_k(x,\cdot)-B_k(x,\cdot)\right\rangle}_{\textsc{Bias-3}}+\underbrace{\sum_{k=1}^{T/W}\sum_{x} q^\star(x)\left\langle \pi^\star(\cdot|x), B_k(x,\cdot)-\widehat{B}_k(x,\cdot)\right\rangle}_{\textsc{Bias-4}} \\
&\quad+\underbrace{\sum_{k=1}^{T/W}\sum_{x} q^\star(x)\left\langle \pi_k(\cdot|x)-\pi^\star(\cdot|x), \widehat{Q}_k(x,\cdot)-\widehat{B}_k(x,\cdot)\right\rangle}_{\textsc{Reg-Term}}
\end{aligned}
$$

In the following lemmas, we bound each term separately:

Lemma F.1.

$$
\mathbb{E}[\textsc{Bias-1}+\textsc{Bias-2}]\le\mathcal{O}\left(\frac{\epsilon H^3 T}{W}\right).
$$

Proof. The proof of this lemma is almost identical to that of Lemma E.2, except that we replace T by T/W, and consider the averaged loss $\bar{\ell}_k$ in an epoch instead of the single episode loss ℓ_t:

$$
\begin{aligned}
&\mathbb{E}_k\left[\overline{Q}_k^{\pi_k}(x,a)-\widehat{Q}_k(x,a)\right] \\
&=\phi(x,a)^\top\left(\overline{\theta}_{k,h}^{\pi_k}-\mathbb{E}_k\left[\widehat{\theta}_{k,h}\right]\right) \\
&=\phi(x,a)^\top\left(\overline{\theta}_{k,h}^{\pi_k}-\mathbb{E}_k\left[\widehat{\Sigma}_{k,h}^{+}\right]\mathbb{E}_k\left[\frac{1}{|S'_k|}\sum_{t\in S'_k}((1-Y_t)+Y_tH\mathbb{1}[h=h_t^*])\phi(x_{t,h},a_{t,h})L_{t,h}\right]\right) \\
&\hspace{20em}(S'_k \text{ is the } S' \text{ in Algorithm 10 within epoch } k) \\
&=\phi(x,a)^\top\left(\overline{\theta}_{k,h}^{\pi_k}-\left(\Sigma_{k,h}^{\text{mix}}\right)^{-1}\mathbb{E}_k\left[\frac{1}{|S'_k|}\sum_{t\in S'_k}((1-Y_t)+Y_tH\mathbb{1}[h=h_t^*])\phi(x_{t,h},a_{t,h})L_{t,h}\right]\right)+\mathcal{O}(\epsilon H^2) \\
&\hspace{12em}(\text{by Lemma E.1 and that } \|\phi(x,a)\|\le 1 \text{ for all } x,a \text{ and } L_{t,h}\le H) \\
&=\phi(x,a)^\top\left(\overline{\theta}_{k,h}^{\pi_k}-\left(\Sigma_{k,h}^{\text{mix}}\right)^{-1}\mathbb{E}_k\left[\frac{1}{|S'_k|}\sum_{t\in S'_k}\Sigma_{k,h}^{\text{mix}}\theta_{t,h}^{\pi_k}\right]\right)+\mathcal{O}(\epsilon H^2) \qquad (45) \\
&=\phi(x,a)^\top\left(\overline{\theta}_{k,h}^{\pi_k}-\left(\Sigma_{k,h}^{\text{mix}}\right)^{-1}\mathbb{E}_k\left[\frac{1}{W}\sum_{t=(k-1)W+1}^{kW}\Sigma_{k,h}^{\text{mix}}\theta_{t,h}^{\pi_k}\right]\right)+\mathcal{O}(\epsilon H^2) \\
&\hspace{24em}(S'_k \text{ is randomly chosen from epoch } k) \\
&=\phi(x,a)^\top\left(\overline{\theta}_{k,h}^{\pi_k}-\left(\Sigma_{k,h}^{\text{mix}}\right)^{-1}\Sigma_{k,h}^{\text{mix}}\overline{\theta}_{k,h}^{\pi_k}\right)+\mathcal{O}(\epsilon H^2) \\
&=\mathcal{O}\left(\epsilon H^2\right) \qquad (46)
\end{aligned}
$$

Similarly, $\mathbb{E}_k\left[\widehat{Q}_k(x,a)-\overline{Q}_k^{\pi_k}(x,a)\right]=\mathcal{O}\left(\epsilon H^2\right)$. Summing them over k,x,a using weights $q^\star(x)\pi_k(a|x)$ and $q^\star(x)\pi^\star(a|x)$ respectively finishes the proof. $\square$

Lemma F.2.

$$\mathbb{E}[\textsc{Bias-3}+\textsc{Bias-4}]\le\widetilde{\mathcal{O}}\left(\frac{\epsilon H^3T}{W}\times\frac{\beta}{\delta_e\lambda_{\min}}\right).$$

Proof. The proof is almost identical to that of the previous lemma. Recall the definition of $\Lambda_{k,h}^{\pi_k}$ in Section 6. Then we have

$$\begin{aligned}
&\mathbb{E}_k\left[B_k(x,a)-\widehat{B}_k(x,a)\right]\\
&=\phi(x,a)^\top\left(\Lambda_{k,h}^{\pi_k}-\mathbb{E}_k\left[\widehat{\Lambda}_{k,h}\right]\right)\\
&=\phi(x,a)^\top\left(\Lambda_{k,h}^{\pi_k}-\mathbb{E}_k\left[\widehat{\Sigma}_{k,h}^{+}\right]\mathbb{E}_k\left[\frac{1}{|S'_k|}\sum_{t\in S'_k}((1-Y_t)+Y_tH\mathbb{1}[h=h_t^*])\phi(x_{t,h},a_{t,h})D_{t,h}\right]\right)\\
&\qquad\qquad(S'_k\text{ is the }S'\text{ in Algorithm 10 within epoch }k)\\
&=\phi(x,a)^\top\left(\Lambda_{k,h}^{\pi_k}-\left(\Sigma_{k,h}^{\text{mix}}\right)^{-1}\mathbb{E}_k\left[\frac{1}{|S'_k|}\sum_{t\in S'_k}((1-Y_t)+Y_tH\mathbb{1}[h=h_t^*])\phi(x_{t,h},a_{t,h})D_{t,h}\right]\right)\\
&\qquad+\widetilde{\mathcal{O}}\left(\epsilon H^2\times\frac{\beta}{\delta_e\lambda_{\min}}\right)\\
&\qquad(\text{by Lemma E.1 and that }\|\phi(x,a)\|\le1\text{ for all }x,a\text{ and }D_{t,h}=\mathcal{O}(H\beta\sup\|\widehat{\Sigma}_{k,h}^{+}\|_{\text{op}})=\widetilde{\mathcal{O}}(\tfrac{H\beta}{\delta_e\lambda_{\min}}))\\
&=\phi(x,a)^\top\left(\Lambda_{k,h}^{\pi_k}-\left(\Sigma_{k,h}^{\text{mix}}\right)^{-1}\mathbb{E}_k\left[\frac{1}{|S'_k|}\sum_{t\in S'_k}\Sigma_{k,h}^{\text{mix}}\Lambda_{k,h}^{\pi_k}\right]\right)+\widetilde{\mathcal{O}}\left(\epsilon H^2\times\frac{\beta}{\delta_e\lambda_{\min}}\right)\\
&=\phi(x,a)^\top\left(\Lambda_{k,h}^{\pi_k}-\left(\Sigma_{k,h}^{\text{mix}}\right)^{-1}\Sigma_{k,h}^{\text{mix}}\Lambda_{k,h}^{\pi_k}\right)+\widetilde{\mathcal{O}}\left(\epsilon H^2\times\frac{\beta}{\delta_e\lambda_{\min}}\right)\\
&=\widetilde{\mathcal{O}}\left(\epsilon H^2\times\frac{\beta}{\delta_e\lambda_{\min}}\right).
\end{aligned}\tag{47}$$

Similar for $\mathbb{E}_k[\widehat{B}_k(x,a)-B_k(x,a)]$. Summing them over k,x,a using weights $q^\star(x)\pi^\star(a|x)$ and $q^\star(x)\pi_t(a|x)$ respectively /finishes the proof. $\square$

Lemma F.3. *Let* $\frac{\eta\beta}{\delta_e^2\lambda_{\min}^2}\le\frac{1}{160H^4\ln(\frac{1}{\epsilon\delta_e\lambda_{\min}})^2}$ *and* $\frac{\eta}{\delta_e\lambda_{\min}}\le\frac{1}{4H^2\ln(\frac{1}{\epsilon\delta_e\lambda_{\min}})}$. *Then*

$$\begin{aligned}
&\mathbb{E}[\textsc{Reg-Term}]\\
&=\widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta\epsilon H^4T}{W}+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2T^2W}+\frac{\eta\epsilon\beta^2H^4T}{\delta_e^2\lambda_{\min}^2W}+\frac{\eta H^4\beta^2}{\delta_e^3\lambda_{\min}^3T^2W}\right)\\
&\qquad+2\eta H^3\mathbb{E}\left[\sum_{k,x,a}q^\star(x)\pi_k(x,a)\|\phi(x,a)\|^2_{\widehat{\Sigma}_{k,h}^{+}}\right]+\frac{1}{H}\mathbb{E}\left[\sum_{k,x,a}q^\star(x)\pi_k(a|x)B_k(x,a)\right].
\end{aligned}$$

Proof. We first check the condition for Lemma A.4: $\eta\left|\widehat{Q}_k(x,a)-\widehat{B}_t(x,a)\right|\le1$. In our case,

$$\begin{aligned}
\eta\left|\widehat{Q}_k(x,a)\right|&=\eta\left|\phi(x,a)^\top\widehat{\Sigma}_{k,h}^{+}\left(\frac{1}{|S'|}\sum_{t\in S'}((1-Y_t)+Y_tH\mathbb{1}[h=h_t^*])\phi(x_{t,h},a_{t,h})L_{t,h}\right)\right|\\
&\le\eta\times\|\widehat{\Sigma}_{k,h}^{+}\|_{\text{op}}\times H\times\sup_{t\in S'}L_{t,h}\\
&\le\eta\times\frac{2}{\delta_e\lambda_{\min}}\ln\frac{1}{\epsilon\delta_e\lambda_{\min}}\times H^2 &&(\text{by Lemma E.1})
\end{aligned}$$

$$\leq \frac{1}{2} \qquad \text{(by the condition specified in the lemma)}$$

and

$$\begin{aligned}
\eta\left|\widehat{B}_k(x,a)\right| &\leq \eta\left|b_k(x,a)\right| + \eta\left|\phi(x,a)^\top\widehat{\Sigma}^+_{k,h}\left(\frac{1}{|S'|}\sum_{t\in S'}((1-Y_t)+Y_tH\mathbb{1}[h=h_t^*])\phi(x_{t,h},a_{t,h})D_{t,h}\right)\right| \\
&\leq \eta\times 2\beta\times\|\widehat{\Sigma}^+_{k,h}\|_{\text{op}} + \eta\times\|\widehat{\Sigma}^+_{k,h}\|_{\text{op}}\times H\times\sup_{t\in S'}D_{t,h} \\
&\leq \eta\times 2\beta\times\|\widehat{\Sigma}^+_{k,h}\|_{\text{op}} + \eta\times\|\widehat{\Sigma}^+_{k,h}\|_{\text{op}}\times H\times(H-1)\left(1+\frac{1}{H}\right)^H\times 2\beta\|\widehat{\Sigma}^+_{k,h}\|_{\text{op}} \\
&\leq 8\eta\beta H^2\times\|\widehat{\Sigma}^+_{k,h}\|^2_{\text{op}} \\
&\leq 8\eta\beta H^2\left(\frac{2}{\delta_e\lambda_{\min}}\ln\frac{1}{\epsilon\delta_e\lambda_{\min}}\right)^2 && \text{(by Lemma E.1)} \\
&\leq \frac{1}{2H}. && \text{(by the condition specified in the lemma)}
\end{aligned}$$

An upper bound for $\mathbb{E}_k\left[\widehat{Q}_k(x,a)^2\right]$ follows the same calculation as in Eq. (39):

$$\begin{aligned}
&\mathbb{E}_k\left[\widehat{Q}_k(x,a)^2\right] \\
&\leq \mathbb{E}_k\left[\frac{1}{|S'_k|}\sum_{t\in S'_k}H^2\phi(x,a)^\top\widehat{\Sigma}^+_{k,h}\Big(((1-Y_t)+Y_tH\mathbb{1}[h=h_t^*])^2\phi(x_{t,h},a_{t,h})\phi(x_{t,h},a_{t,h})^\top\Big)\widehat{\Sigma}^+_{k,h}\phi(x,a)\right] \\
& \hspace{30em} (*) \\
&= \mathbb{E}_k\left[H^2\phi(x,a)^\top\widehat{\Sigma}^+_{k,h}\Big((1-\delta_e)\Sigma_{k,h}+\delta_eH\Sigma_h^{\pi_0}\Big)\widehat{\Sigma}^+_{k,h}\phi(x,a)\right] \\
&\leq H^3\mathbb{E}_k\left[\phi(x,a)^\top\widehat{\Sigma}^+_{k,h}\Sigma^{\text{mix}}_{k,h}\widehat{\Sigma}^+_{k,h}\phi(x,a)\right] \\
&\leq H^3\mathbb{E}_k\left[\phi(x,a)^\top\widehat{\Sigma}^+_{k,h}\Sigma^{\text{mix}}_{k,h}(\Sigma^{\text{mix}}_{k,h})^{-1}\phi(x,a)\right] + \widetilde{\mathcal{O}}\left(\epsilon H^3+\frac{H^3}{\delta_e^2\lambda_{\min}^2T^3}\right) \\
& \hspace{25em} \text{(by Eq. (35) and Eq. (36))} \\
&= H^3\mathbb{E}_k\left[\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right] + \widetilde{\mathcal{O}}\left(\epsilon H^3+\frac{H^3}{\delta_e^2\lambda_{\min}^2T^3}\right), \qquad (48)
\end{aligned}$$

where in $(*)$ we use $\left(\frac{1}{|S'_k|}\sum_{t\in S'_k}v_t\right)^2 \leq \frac{1}{|S'_k|}\sum_{t\in S'_k}v_t^2$ with $v_t = \phi(x,a)^\top\widehat{\Sigma}^+_{k,h}\left((1-Y_t)+Y_tH\mathbb{1}[h=h_t^*]\right)\phi(x_{t,h},a_{t,h})L_{t,h}$.

Next, we bound $\mathbb{E}_t\left[\widehat{B}_t(x,a)^2\right]$:

$$\begin{aligned}
&\mathbb{E}_k\left[\widehat{B}_k(x,a)^2\right] \\
&\leq 2\mathbb{E}_k\left[b_k(x,a)^2\right] + 2\mathbb{E}_k\left[(\phi(x,a)^\top\widehat{\Lambda}_{k,h})^2\right] \\
&\leq 2\left(\beta\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}} + \beta\mathbb{E}_{a'\sim\pi_k(\cdot|x)}\left[\|\phi(x,a')\|^2_{\widehat{\Sigma}^+_{k,h}}\right]\right)^2 \\
&\qquad + 2H^3\left(\frac{6\beta}{\delta_e\lambda_{\min}}\ln\frac{1}{\epsilon\delta_e\lambda_{\min}}\right)^2\mathbb{E}_k\left[\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right] + \widetilde{\mathcal{O}}\left(\epsilon H^3+\frac{H^3}{\delta_e^2\lambda_{\min}^2T^3}\right)\times\left(\frac{6\beta}{\delta_e\lambda_{\min}}\ln\frac{1}{\epsilon\delta_e\lambda_{\min}}\right)^2 \\
&\leq \frac{8\beta}{\delta_e\lambda_{\min}}\ln\frac{1}{\epsilon\delta_e\lambda_{\min}}\left(\beta\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}} + \beta\mathbb{E}_{a'\sim\pi_k(\cdot|x)}\left[\|\phi(x,a')\|^2_{\widehat{\Sigma}^+_{k,h}}\right]\right) \\
&\qquad + 2H^3\left(\frac{6\beta}{\delta_e\lambda_{\min}}\ln\frac{1}{\epsilon\delta_e\lambda_{\min}}\right)^2\mathbb{E}_k\left[\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right] + \widetilde{\mathcal{O}}\left(\epsilon H^3+\frac{H^3}{\delta_e^2\lambda_{\min}^2T^3}\right)\times\left(\frac{6\beta}{\delta_e\lambda_{\min}}\ln\frac{1}{\epsilon\delta_e\lambda_{\min}}\right)^2
\end{aligned}$$

$$\leq \left(\frac{8\beta}{\delta_e\lambda_{\min}}\ln\frac{1}{\epsilon\delta_e\lambda_{\min}} + \frac{72H^3\beta}{\delta_e^2\lambda_{\min}^2}\ln^2\left(\frac{1}{\epsilon\delta_e\lambda_{\min}}\right)\right)b_k(x,a) + \widetilde{\mathcal{O}}\left(\frac{\epsilon\beta^2H^3}{\delta_e^2\lambda_{\min}^2} + \frac{H^3\beta^2}{\delta_e^4\lambda_{\min}^4T^3}\right)$$
$$\leq \frac{80H^3\beta}{\delta_e^2\lambda_{\min}^2}\ln^2\left(\frac{1}{\epsilon\delta_e\lambda_{\min}}\right)b_k(x,a) + \widetilde{\mathcal{O}}\left(\frac{\epsilon\beta^2H^3}{\delta_e^2\lambda_{\min}^2} + \frac{H^3\beta^2}{\delta_e^4\lambda_{\min}^4T^3}\right),$$

where in the second inequality we bound $\mathbb{E}_k\left[(\phi(x,a)^\top\widehat{\Lambda}_{k,h})^2\right]$ similarly as we bound $\mathbb{E}_k\left[\widehat{Q}_k(x,a)^2\right]$ in Eq. (48), except that we replace the upper bound for $L_{t,h}$ as H by the upper bound for $D_{t,h}$ as $H\left(1+\frac{1}{H}\right)^H\beta\|\widehat{\Sigma}^+_{k,h}\|_{\text{op}} \leq 3H\times\beta\times\frac{2}{\delta_e\lambda_{\min}}\ln\frac{1}{\epsilon\delta_e\lambda_{\min}}$ by Eq. (33). In the third inequality, we use that

$$\beta\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}} \leq \beta\times\frac{2}{\delta_e\lambda_{\min}}\ln\frac{1}{\epsilon\delta_e\lambda_{\min}}. \qquad \text{(also by Eq. (33))}$$

Thus, by Lemma A.4, we have

$\mathbb{E}[\text{REG-TERM}]$

$$\leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + 2\eta\sum_{k,x,a}q^\star(x)\pi_k(a|x)\widehat{Q}_k(x,a)^2 + 2\eta\sum_{k,x,a}q^\star(x)\pi_k(a|x)B_k(x,a)^2$$
$$\leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\eta\epsilon H^4T}{W} + \frac{\eta H^4}{\delta_e^2\lambda_{\min}^2T^2W} + \frac{\eta\epsilon\beta^2H^4T}{\delta_e^2\lambda_{\min}^2W} + \frac{\eta H^4\beta^2}{\delta_e^4\lambda_{\min}^4T^2W}\right)$$
$$+ 2\eta H^3\mathbb{E}\left[\sum_{k,x,a}q^\star(x)\pi_k(x,a)\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right] + \frac{160H^3\eta\beta}{\delta_e^2\lambda_{\min}^2}\ln^2\left(\frac{1}{\epsilon\delta_e\lambda_{\min}}\right)\mathbb{E}\left[\sum_{k,x,a}q^\star(x)\pi_k(a|x)b_k(x,a)\right]$$
$$\leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\eta\epsilon H^4T}{W} + \frac{\eta H^4}{\delta_e^2\lambda_{\min}^2T^2W} + \frac{\eta\epsilon\beta^2H^4T}{\delta_e^2\lambda_{\min}^2W} + \frac{\eta H^4\beta^2}{\delta_e^4\lambda_{\min}^4T^2W}\right)$$
$$+ 2\eta H^3\mathbb{E}\left[\sum_{k,x,a}q^\star(x)\pi_k(x,a)\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right] + \frac{1}{H}\mathbb{E}\left[\sum_{k,x,a}q^\star(x)\pi_k(a|x)B_k(x,a)\right]$$

where in the last inequality we use the condition specified in the lemma and that $B_k(x,a)\geq b_k(x,a)$.

□

Proof of Theorem 6.2. Now we combine the bounds in Lemma F.2, Lemma F.2, Lemma F.3. We get

$\mathbb{E}[\text{BIAS-1} + \text{BIAS-2} + \text{BIAS-3} + \text{BIAS-4} + \text{REG-TERM}]$

$$= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\eta\epsilon H^4T}{W} + \frac{\eta H^4}{\delta_e^2\lambda_{\min}^2T^2W} + \frac{\eta\epsilon\beta^2H^4T}{\delta_e^2\lambda_{\min}^2W} + \frac{\eta H^4\beta^2}{\delta_e^4\lambda_{\min}^4T^2W} + \frac{\epsilon H^3T}{W}\times\frac{\beta}{\delta_e\lambda_{\min}} + \frac{\epsilon H^3T}{W}\right)$$
$$+ \sum_{k,x,a}q^\star(x)\pi_k(a|x)b_k(x,a)$$
$$+ \frac{1}{H}\sum_{k,x,a}q^\star(x)\pi_k(a|x)B_k(x,a)$$

where we use $b_k = 2\eta H^3$. By picking $\beta\leq\delta_e\lambda_{\min}$, the first term above can be further upper bounded by

$$\widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\eta\epsilon H^4T}{W} + \frac{\eta H^4}{\delta_e^2\lambda_{\min}^2T^2W} + \frac{\epsilon H^3T}{W}\right).$$

By Lemma B.2, we have

$$\mathbb{E}\left[\sum_{k=1}^{T/W}V^{\pi_k}(x_0;\overline{\ell}_k)\right] - \sum_{k=1}^{T/W}V^{\pi^\star}(x_0;\overline{\ell}_k)$$

$$\leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta\epsilon H^4 T}{W}+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2 T^2 W}+\frac{\epsilon H^3 T}{W}+\beta\mathbb{E}\left[\sum_{k=1}^{T/W}\sum_{x,a} q_k(x)\pi_k(a|x)\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right]\right)$$

$$= \widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta\epsilon H^4 T}{W}+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2 T^2 W}+\frac{\epsilon H^3 T}{W}+\frac{\beta d H T}{W}\right)$$

(by similar calculation as in Eq. (40))

$$= \widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta\epsilon H^4 T}{W}+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2 T^2 W}+\frac{\epsilon H^3 T}{W}+\frac{\eta d H^4 T}{W}\right) \qquad (\beta = 2\eta H^3)$$

Multiplying back with W, and considering the exploration rate δ_e, we see that the true expected regret is upper bounded by

$$\widetilde{\mathcal{O}}\left(\frac{HW}{\eta}+\eta\epsilon H^4 T+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2 T^2}+\epsilon H^3 T+\eta d H^4 T+\delta_e H T\right)$$

$$= \widetilde{\mathcal{O}}\left(\frac{H}{\eta\epsilon^2\delta_e^3\lambda_{\min}^3}+\eta\epsilon H^4 T+\frac{\eta H^4}{\delta_e^2\lambda_{\min}^2 T^2}+\epsilon H^3 T+\eta d H^4 T+\delta_e H T\right)$$

where we use the specified value of M and N and that $W = 2MN$.

Considering the constraints in Lemma F.3 and that $\beta = 2\eta H^3$, we pick

$$\eta = \frac{\delta_e\lambda_{\min}}{20H^{3.5}\ln\left(\frac{1}{\epsilon\delta_e\lambda_{\min}}\right)}$$

which also makes $\beta \leq \delta_e\lambda_{\min}$ as we assumed previously.

With this η, the regret can be simplified as

$$\widetilde{\mathcal{O}}\left(\frac{H^{4.5}}{\epsilon^2\delta_e^4\lambda_{\min}^4}+\delta_e\lambda_{\min}\epsilon\sqrt{H}T+\frac{\sqrt{H}}{\delta_e\lambda_{\min}T^2}+\epsilon H^3 T+\delta_e\lambda_{\min} d\sqrt{H}T+\delta_e H T\right)$$

$$= \widetilde{\mathcal{O}}\left(\frac{H^{4.5}}{\epsilon^2\delta_e^4\lambda_{\min}^4}+\epsilon H^3 T+\delta_e\lambda_{\min} d\sqrt{H}T+\delta_e H T\right)$$

By picking

$$\epsilon = \left(\frac{H^{1.5}}{\delta_e^4\lambda_{\min}^4 T}\right)^{\frac{1}{3}}, \qquad \delta_e = \left(\frac{H^9}{T\lambda_{\min}^4(\lambda_{\min}d+\sqrt{H})^3}\right)^{\frac{1}{7}},$$

we get a regret bound of

$$\widetilde{\mathcal{O}}\left(\left(H^{12.5}\left(\frac{d\lambda_{\min}+\sqrt{H}}{\lambda_{\min}}\right)^4 T^6\right)^{\frac{1}{7}}\right).$$

□