Policy Optimization in Adversarial MDPs: Improved Exploration via Dilated Bonuses

Haipeng Luo* **Chen-Yu Wei*** **Chung-Wei Lee**

{haipengl, chenyu.wei, leechung}@usc.edu

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 (Shani et al., 2020). 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 (Neu & Olkhovskaya, 2020) while importantly removing the need of an exploratory policy that their algorithm requires. When a simulator is unavailable, we further consider a linear MDP setting and obtain $\widetilde{\mathcal{O}}(T^{14/15})$ regret, which is the first result for linear MDPs with adversarial losses and bandit feedback.

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 (Schulman et al., 2017) and robotics (Levine & Koltun, 2013). 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., (Abbasi-Yadkori et al., 2019; Agarwal et al., 2020b; Neu & Olkhovskaya, 2020; Wei et al., 2021)), making it theoretically less appealing compared to other methods.

Motivated by this issue, a line of recent works (Cai et al., 2020; Shani et al., 2020; Agarwal et al., 2020a; Zanette et al., 2021) 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 misspecification). Despite these important progresses, however, many limitations still exist, including worse regret rates comparing to the best value-based or model-based approaches (Shani et al., 2020; Agarwal et al., 2020a; Zanette et al., 2021), or requiring full-information feedback on the entire loss function (as opposed to the more realistic bandit feedback) (Cai et al., 2020).

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 (Shani et al., 2020), 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 (Jin et al., 2020a; Lee et al., 2020) 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 A)
- 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 (Neu & Olkhovskaya, 2020).

We obtain the same $\widetilde{\mathcal{O}}(T^{2/3})$ regret while importantly removing their exploratory assumption. (Section B)

- Finally, to remove the need of a sampling oracle, we further consider linear MDPs, a special case where the transition kernel is also linear in the features. To our knowledge, the only existing works that consider adversarial losses in this setup are (Cai et al., 2020), which obtains $\widetilde{\mathcal{O}}(\sqrt{T})$ regret but requires full-information feedback on the loss functions, and (Neu & Olkhovskaya, 2021) (an updated version of (Neu & Olkhovskaya, 2020)), which obtains $\widetilde{\mathcal{O}}(\sqrt{T})$ regret under bandit feedback but requires perfect knowledge of the transition as well as an exploratory assumption. We propose the first algorithm for the most challenging setting with bandit feedback and unknown transition, which achieves $\widetilde{\mathcal{O}}(T^{14/15})$ regret without any exploratory assumption. (Section C)

We emphasize that unlike the tabular setting (where we improve existing regret rates of policy optimization), in the two adversarial linear function approximation settings with bandit feedback that we consider, researchers have not been able to show *any* sublinear regret for policy optimization without exploratory assumptions before our work, which shows the critical role of our proposed dilated bonuses. In fact, there are simply no existing algorithms with sublinear regret *at all* for these two settings, be it policy-optimization-type or not. This shows the advantage of policy optimization over other approaches, when combined with our dilated bonuses.

Related work. In the tabular setting, except for (Shani et al., 2020), most algorithms apply the occupancy-measure-based framework to handle adversarial losses (e.g., (Rosenberg & Mansour, 2019; Jin et al., 2020a; Chen et al., 2021; Chen & Luo, 2021)), which as mentioned is computationally expensive. For stochastic losses, there are many more different approaches such as model-based ones (Jaksch et al., 2010; Dann & Brunskill, 2015; Azar et al., 2017; Fruit et al., 2018; Zanette & Brunskill, 2019) and value-based ones (Jin et al., 2018; Dong et al., 2019).

Theoretical studies for linear function approximation have gained increasing interest recently (Yang & Wang, 2020; Zanette et al., 2020; Jin et al., 2020b). Most of them study stochastic/stationary losses, with the exception of (Cai et al., 2020; Neu & Olkhovskaya, 2020; 2021). Our algorithm for the linear MDP setting bears some similarity to those of (Agarwal et al., 2020a; Zanette et al., 2021) which consider stationary losses. However, 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: $V(x_H; \ell) = 0$,

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

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).[1] 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 $\widetilde{\mathcal{O}}(H|X|\sqrt{|A|T})$ (Jin et al., 2020a), with an extra $\sqrt{X}$ fac-

[1] 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.

tor compared to the best existing lower bound (Jin et al., 2018).

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 $\mathrm{Var}_t[\cdot]$ 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:

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

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 $Q_t^{\pi_t}(x,a)$ as the loss of action a in state x. Since $Q_t^{\pi_t}(x,a)$ is unknown, however, the algorithm has to use some estimator of $Q_t^{\pi_t}(x,a)$ 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 on the assumption that some *distribution mismatch coefficient* related to $q^\star(x,a)/q_t(x,a)$ is bounded (see e.g., (Agarwal et al., 2020b; Wei et al., 2020)).

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:

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

Then by linearity and rearranging, we have

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

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 (Kakade & Langford, 2002):

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

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

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

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* $\mathrm{Reg} \le o(T) + 3\sum_{t=1}^{T} V^{\pi_t}(x_0;b_t)$.

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 E.1) or a version where Eq. (4) and Eq. (5) only hold in expectation (see Lemma E.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.

Due to the space limit, we provide the instantiation of the idea of dilated bonuses and the proofs in the appendix.

4. Conclusions and Future Directions

In this work, we propose the general idea of dilated bonuses and demonstrate how it leads to improved exploration and regret bounds for policy optimization in various settings. One future direction is to further improve our results in the function approximation setting, including reducing the number of simulator calls in the linear-Q setting and improving the regret bound for the linear MDP setting (which is currently far from optimal). A potential idea for the latter is to reuse data across different epochs, an idea adopted by several recent works (Zanette et al., 2021; Lazic et al., 2021) for different problems. Another key future direction is to investigate whether the idea of dilated bonuses is applicable beyond the finite-horizon setting (e.g. whether it is applicable to the more general stochastic shortest path model or the infinite-horizon setting).

Acknowledgments We thank Gergely Neu and Julia Olkhovskaya for discussions on the technical details of their GEOMETRICRESAMPLING procedure.

References

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

Agarwal, A., Henaff, M., Kakade, S., and Sun, W. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020a.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020b.

Azar, M. G., Munos, R., and Kappen, B. On the sample complexity of reinforcement learning with a generative model. In *International Conference on Machine Learning*, 2012.

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

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, 2011.

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

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

Chen, L., Luo, H., and Wei, C.-Y. Minimax regret for stochastic shortest path with adversarial costs and known transition. In *Conference On Learning Theory*, 2021.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. *arXiv preprint arXiv:1510.08906*, 2015.

Dong, K., Wang, Y., Chen, X., and Wang, L. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. *arXiv preprint arXiv:1901.09311*, 2019.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pp. 1578–1586. PMLR, 2018.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? In *Advances in neural information processing systems*, pp. 4863–4873, 2018.

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

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

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

Kakade, S. M. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Lazic, N., Yin, D., Abbasi-Yadkori, Y., and Szepesvari, C. Improved regret bound and experience replay in regularized policy iteration. *arXiv preprint arXiv:2102.12611*, 2021.

Lee, C.-W., Luo, H., Wei, C.-Y., and Zhang, M. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in Neural Information Processing Systems*, 2020.

Levine, S. and Koltun, V. Guided policy search. In *International conference on machine learning*, pp. 1–9. PMLR, 2013.

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

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

Neu, G. and Olkhovskaya, J. Online learning in mdps with linear function approximation and bandit feedback. *arXiv preprint arXiv:2007.01612v1*, 2020.

Neu, G. and Olkhovskaya, J. Online learning in mdps with linear function approximation and bandit feedback. *arXiv preprint arXiv:2007.01612v2*, 2021.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Sidford, A., Wang, M., Wu, X., Yang, L. F., and Ye, Y. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pp. 5192–5202, 2018.

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

Wang, R., Du, S. S., Yang, L. F., and Salakhutdinov, R. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International Conference on Machine Learning*, pp. 10170–10180. PMLR, 2020.

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

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

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

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

Zanette, A., Cheng, C.-A., and Agarwal, A. Cautiously optimistic policy optimization and exploration with linear function approximation. *arXiv preprint arXiv:2103.12923*, 2021.

A. 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 (Shani et al., 2020). See Algorithm 1 for the complete pseudocode.

Algorithm 1 Policy Optimization with Dilated Bonuses (Tabular Case)

Parameters: $\delta \in (0,1)$, $\eta = \min\left\{1/24H^3, 1/\sqrt{|X||A|HT}\right\}$, $\gamma = 2\eta H$.

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

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. $\forall h < H$, $N_k(x_{t,h}, a_{t,h}) \xleftarrow{+} 1$, $N_k(x_{t,h}, a_{t,h}, x_{t,h+1}) \xleftarrow{+} 1$.[2]

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:

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

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

---

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

Algorithm design. First, to handle unknown transition, we follow the common practice (dating back to (Jaksch et al., 2010)) 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 (Kakade, 2001; Agarwal et al., 2020b; Wei et al., 2021)), 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 (Jin et al., 2020a) (see also Appendix F.1); and finally γ is a parameter used to control the maximum magnitude of $\widehat{Q}_t(x,a)$. 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 (Jin et al., 2020a).

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

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

Theorem A.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 (Shani et al., 2020). It almost matches the best existing upper bound for this problem, which is $\widetilde{\mathcal{O}}(H|X|\sqrt{|A|T})$ (Jin et al., 2020a). 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 (Jin et al., 2020a), 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 F. 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), $\sum_x q^\star(x)\sum_t \langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), Q_t^{\pi_t}(x,\cdot) - B_t(x,\cdot)\rangle$, as BIAS-1 + BIAS-2 + REG-TERM, where

- BIAS-1 $= \sum_x q^\star(x)\sum_t\langle \pi_t(\cdot|x), Q_t^{\pi_t}(x,\cdot) - \widehat{Q}_t(x,\cdot)\rangle$ 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 F.1);
- BIAS-2 $= \sum_x q^\star(x)\sum_t\langle \pi^\star(\cdot|x), \widehat{Q}_t(x,\cdot) - Q_t^{\pi_t}(x,\cdot)\rangle$ 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 F.2);
- 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 F.3).

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

parameters: $\gamma, \beta, \eta, \epsilon \in (0, \frac{1}{2})$, $M = \left\lceil \frac{24 \ln(dHT)}{\epsilon^2 \gamma^2} \right\rceil$, $N = \left\lceil \frac{2}{\gamma} \ln \frac{1}{\epsilon\gamma} \right\rceil$.

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. Collect MN trajectories using the simulator and π_t. Let $\mathcal{T}_t$ be the set of trajectories. Compute

$$\left\{\widehat{\Sigma}^+_{t,h}\right\}_{h=0}^{H-1} = \textsc{GeometricResampling}\left(\mathcal{T}_t, M, N, \gamma\right). \qquad \text{(see Algorithm 4)}$$

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

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

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

B. 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 $Q^\pi_t(x, a)$ 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 $\theta^\pi_{t,h} \in \mathbb{R}^d$ such that for all $(x, a) \in X_h \times A$, $Q^\pi_t(x, a) = \phi(x, a)^\top \theta^\pi_{t,h}$. Without loss of generality, we assume $\|\phi(x, a)\| \leq 1$ for all (x, a) and $\|\theta^\pi_{t,h}\| \leq \sqrt{d}H$ for all t, h, π.*

For justification on the last condition on norms, see (Wei et al., 2021, Lemma 8). This linear-Q assumption has been made in several recent works with stationary losses (Abbasi-Yadkori et al., 2019; Wei et al., 2021) and also in (Neu & Olkhovskaya, 2020) with the same adversarial losses.[3] It is weaker than the linear MDP assumption (see Section C) 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 (Neu & Olkhovskaya, 2020) and more earlier works with stationary losses (see e.g., (Azar et al., 2012; Sidford et al., 2018)).[4] In this setting, we propose a new policy optimization algorithm with $\widetilde{\mathcal{O}}(T^{2/3})$ regret. See Algorithm 2 for the pseudocode.

[3] The assumption in (Neu & Olkhovskaya, 2020) 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.

[4] The simulator required by (Neu & Olkhovskaya, 2020) 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 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 $\beta\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}} + \mathbb{E}_{j\sim\pi_t(\cdot|x)}\left[\beta\|\phi(x,j)\|^2_{\widehat{\Sigma}^+_{t,h}}\right] + \left(1+\frac{1}{H}\right)\textsc{Bonus}(t, x', a')$.

Algorithm 4 $\textsc{GeometricResampling}(\mathcal{T}, M, N, \gamma)$

Denote the MN trajectories in $\mathcal{T}$ as: $\{(x_{i,0}, a_{i,0}, \ldots, x_{i,H-1}, a_{i,H-1})\}_{i=1,\ldots,MN}$. Let $c = \frac{1}{2}$.
for $m = 1, \ldots, M$ **do**
 for $n = 1, \ldots, N$ **do**
  $i = (m-1)N + n$.
  For all h, compute $Y_{n,h} = \gamma I + \phi(x_{i,h}, a_{i,h})\phi(x_{i,h}, a_{i,h})^\top$.
  For all h, compute $Z_{n,h} = \Pi_{j=1}^{n}(I - cY_{j,h})$.
 For all h, set $\widehat{\Sigma}^{+(m)}_h = cI + c\sum_{n=1}^{N} Z_{n,h}$.
For all h, set $\widehat{\Sigma}^+_h = \frac{1}{M}\sum_{m=1}^{M}\widehat{\Sigma}^{+(m)}_h$.
return $\widehat{\Sigma}^+_h$ for all $h = 0, \ldots, H-1$.

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 $Q^{\pi_t}_t(x,a) = \phi(x,a)^\top\theta^{\pi_t}_{t,h}$, 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 (Neu & Olkhovskaya, 2020) (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 $Q^{\pi_t}_t(x_{t,h}, a_{t,h}) = \phi(x_{t,h}, a_{t,h})^\top\theta^{\pi_t}_{t,h}$. Therefore, 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 4) of (Neu & Olkhovskaya, 2020) to compute $\widehat{\Sigma}^+_{t,h}$, which requires 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 H for the proof.)

Theorem B.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 (Neu & Olkhovskaya, 2020, Theorem 1), without the need of their assumption which essentially says that the learner is given an exploratory policy to start with.[5] To our knowledge, this is the first no-regret algorithm for the linear-Q setting (with adversarial losses and bandit feedback) when no exploratory assumptions are made.

C. The Linear MDP Case

To remove the need of a simulator, we further consider the linear MDP case, a special case of the linear-Q setting. It is equivalent to Assumption 1 plus the extra assumption that the transition function also has a low-rank structure, formally stated below.

Assumption 3 (Linear MDP). *The MDP satisfies Assumption 1 and that for any h and $x' \in X_{h+1}$, there exists an unknown 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 (Cai et al., 2020) 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, except for a concurrent work (Neu & Olkhovskaya, 2021) which assumes perfect knowledge of the transition and an exploratory condition. We propose the first algorithm with sublinear regret for this problem with unknown transition and bandit feedback, shown in Algorithm 5. The structure of Algorithm 5 is similar to that of Algorithm 2, but importantly with the following modifications.

A succinct representation of dilated bonuses Our definition of b_t remains the same as in the linear-Q case. However, due to the low-rank transition structure in linear MDPs, 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 3 $B_t(x,a)$ 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, following the similar idea of using $\widehat{\theta}_{t,h}$ to estimate $\theta_{t,h}^{\pi_t}$ as we did in Algorithm 2, we can construct $\widehat{\Lambda}_{t,h}$ to estimate $\Lambda_{t,h}^{\pi_t}$ as well, thus succinctly representing $B_t(x,a)$ for all (x,a).

Epoch schedule 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 of the algorithm is to construct $\widehat{\Sigma}_{t,h}^{+}$ using *online* samples. To do so, we divide the entire horizon (or more accurately the last $T-T_0$ rounds since the first T_0 rounds are reserved for some other purpose to be discussed next) into epochs with equal length W, and only update the policy optimization algorithm at the beginning of an epoch. We index an epoch by k, and thus $\theta_{t,h}^{\pi_t}$, $\Lambda_{t,h}^{\pi_t}$, $\widehat{\Sigma}_{t,h}^{+}$ are now denoted by $\theta_{k,h}^{\pi_k}$, $\Lambda_{k,h}^{\pi_k}$, $\widehat{\Sigma}_{k,h}^{+}$. Within an epoch, we keep executing the same policy π_k (up to a small exploration probability δ_e) and collect W trajectories, which are then used to construct $\widehat{\Sigma}_{k,h}^{+}$ as well as $\theta_{k,h}^{\pi_k}$ and $\Lambda_{k,h}^{\pi_k}$. To decouple their dependence, we uniformly at random partition these W trajectories into two sets S and S' with equal size, and use data from S to construct $\widehat{\Sigma}_{k,h}^{+}$ in **Step 2** via the same GEOMETRICRESAMPLING procedure and data from S' to construct $\theta_{k,h}^{\pi_k}$ and $\Lambda_{k,h}^{\pi_k}$ in **Step 3** and **Step 4** respectively.

Exploration with a policy cover Unfortunately, some technical difficulty arises when bounding the estimation error and the variance of $\widehat{\Lambda}_{k,h}$. Specifically, they can be large if the magnitude of the bonus term $b_k(x,a)$ is large for some (x,a); furthermore, since $\widehat{\Lambda}_{k,h}$ is constructed using empirical samples, its variance can be even larger in those directions of the feature space that are rarely visited. Overall, due to the combined effect of these two facts, we are unable to prove any sublinear regret with only the ideas described so far.

To address this issue, we adopt the idea of *policy cover*, recently introduced in (Agarwal et al., 2020a; Zanette et al., 2021). Specifically, we spend the first T_0 rounds to find an exploratory (mixture) policy π_{cov} (called policy cover) which tends to reach all possible directions of the feature space. This is done via the procedure POLICYCOVER (Algorithm 6) (to be discussed in detail soon), which also returns $\widehat{\Sigma}_h^{\text{cov}}$ for each layer h, an estimator of the true covariance matrix Σ_h^{cov} of the

[5]Under an even stronger assumption that every policy is exploratory, they also improve the regret to $\widetilde{\mathcal{O}}(\sqrt{T})$; see (Neu & Olkhovskaya, 2020, Theorem 2).

Algorithm 5 Policy Optimization with Dilated Bonuses (Linear MDP Case)

Parameters: $\gamma, \beta, \eta, \epsilon, \delta_e \in (0, \frac{1}{2}), \delta$, $M = \left\lceil \frac{96 \ln(dHT)}{\epsilon^2\gamma^2} \right\rceil$, $N = \left\lceil \frac{2}{\gamma} \ln \frac{1}{\epsilon\gamma} \right\rceil$, $W = 2MN$, $\alpha = \frac{\delta_e}{6\beta}$, $M_0 = \lceil \alpha^2 dH^2 \rceil$, $N_0 = \frac{100 M_0^4 \log(T/\delta)}{\alpha^2}$, $T_0 = M_0 N_0$.

Construct a mixture policy π_{cov} and its estimated covariance matrices (which requires interacting with the environment for the first T_0 rounds using Algorithm 6):

$$\pi_{\text{cov}}, \left\{\widehat{\Sigma}_h^{\text{cov}}\right\}_{h=0,\ldots,H-1} \leftarrow \textsc{PolicyCover}(M_0, N_0, \alpha, \delta).$$

Define known state set $\mathcal{K} = \left\{x \in X : \forall a \in A, \|\phi(x,a)\|^2_{(\widehat{\Sigma}_h^{\text{cov}})^{-1}} \leq \alpha \text{ where } h \text{ is such that } x \in X_h\right\}$.

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

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

$$\pi_k(a|x) \propto \exp\left(-\eta \sum_{\tau=1}^{k-1} \left(\phi(x,a)^\top \widehat{\theta}_{\tau,h} - \phi(x,a)^\top \widehat{\Lambda}_{\tau,h} - b_\tau(x,a)\right)\right) \tag{13}$$

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

Randomly partition $\{T_0 + (k-1)W + 1, \ldots, T_0 + kW\}$ into two parts: S and S', such that $|S| = |S'| = W/2$.

for $t = T_0 + (k-1)W + 1, \ldots, T_0 + kW$ **do**

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

if $Y_t = 1$ **then**

if $t \in S$ **then** Execute π_{cov}.

else Draw $h_t^* \overset{\text{unif.}}{\sim} \{0, \ldots, H-1\}$; execute π_{cov} in steps $0, \ldots, h_t^* - 1$ and π_k in steps $h_t^*, \ldots, H-1$.

else Execute π_k.

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

Step 2: Construct inverse covariance matrix estimators. Let

$$\mathcal{T}_k = \{(x_{t,0}, a_{t,0}, \ldots, x_{t,H-1}, a_{t,H-1})\}_{t\in S}, \qquad \text{(the trajectories in } S\text{)}$$

$$\left\{\widehat{\Sigma}^+_{k,h}\right\}_{h=0}^{H-1} = \textsc{GeometricResampling}(\mathcal{T}_k, M, N, \gamma). \tag{14}$$

Step 3: Construct Q-function weight estimators. Computer for all h (with $L_{t,h} = \sum_{i=h}^{H-1} \ell_t(x_{t,i}, a_{t,i})$):

$$\widehat{\theta}_{k,h} = \widehat{\Sigma}^+_{k,h}\left(\frac{1}{|S'|}\sum_{t\in S'}((1-Y_t) + Y_t H \mathbb{1}[h = h_t^*])\phi(x_{t,h}, a_{t,h}) L_{t,h}\right). \tag{15}$$

Step 4: Construct bonus function weight estimators. Computer for all h :

$$\widehat{\Lambda}_{k,h} = \widehat{\Sigma}^+_{k,h}\left(\frac{1}{|S'|}\sum_{t\in S'}((1-Y_t) + Y_t H \mathbb{1}[h = h_t^*])\phi(x_{t,h}, a_{t,h}) D_{t,h}\right), \tag{16}$$

where $D_{t,h} = \sum_{i=h+1}^{H-1} \left(1 + \frac{1}{H}\right)^{i-h} b_k(x_{t,i}, a_{t,i})$.

---

policy cover π_{cov}. $\textsc{PolicyCover}$ guarantees that with high probability, for any policy π and h we have

$$\Pr_{x_h\sim\pi}\left[\exists a,\ \|\phi(x_h,a)\|^2_{(\widehat{\Sigma}_h^{\text{cov}})^{-1}} \geq \alpha\right] \leq \widetilde{\mathcal{O}}\left(\frac{dH}{\alpha}\right) \tag{17}$$

where $x_h \in X_h$ is sampled from executing π; see Lemma G.4. This motivates us to only focus on x such that

$\|\phi(x,a)\|^2_{(\widehat{\Sigma}^{\text{cov}}_h)^{-1}} \leq \alpha$ for all a (h is the layer to which x belongs). This would not incur much regret because no policy would visit other states often enough. We call such state a *known* state and denote by $\mathcal{K}$ the set of all known states. To implement the idea above, we simply introduce an indicator $\mathbb{1}[x \in \mathcal{K}]$ in the definition of b_k (that is, no bonus at all for unknown states).

The benefit of doing so is that the aforementioned issue of $b_k(x,a)$ having a large magnitude is now alleviated as long as we explore using π_{cov} with some small probability in each episode. Specifically, in each episode of epoch k, with probability $1-\delta_e$ we execute π_k suggested by policy optimization, otherwise we explore using π_{cov}. The way we explore differs slightly for episodes in S and those in S' (recall that an epoch is partitioned evenly into S and S', where S is used to estimate $\widehat{\Sigma}^+_{k,h}$ and S' is used to estimate $\theta^{\pi_k}_{k,h}$ and $\Lambda^{\pi_k}_{k,h}$). For an episode in S, we simply explore by executing π_{cov} for the entire episode, so that $\widehat{\Sigma}^+_{k,h}$ is an estimation of the inverse of $\gamma I + \delta_e \Sigma^{\text{cov}}_h + (1-\delta_e)\mathbb{E}_{(x_h,a)\sim\pi_k}[\phi(x_h,a)\phi(x_h,a)^\top]$, and thus by its definition $b_k(x,a)$ is bounded by roughly $\frac{\alpha\beta}{\delta_e}$ for all (x,a) (this improves over the trivial bound $\frac{\beta}{\gamma}$ by our choice of parameters; see Lemma I.1). On the other hand, for an episode in S', we first uniformly at random draw a step h^*_t, then we execute π_{cov} for the first h^*_t steps and continue with π_k for the rest. This leads to a slightly different form of the estimators $\theta^{\pi_k}_{k,h}$ and $\Lambda^{\pi_k}_{k,h}$ compared to Eq. (12) (see Eq. (15) and Eq. (16), where the definition of $D_{t,h}$ is in light of Eq. (4)), which is important to ensure their (almost) unbiasedness. This also concludes the description of **Step 1**.

We note that the idea of dividing states into known and unknown parts is related to those of (Agarwal et al., 2020a; Zanette et al., 2021). However, our case is more challenging because we are only allowed to mix a small amount of π_{cov} into our policy in order to get sublinear regret against an adversary, while their algorithms can always start by executing π_{cov} in each episode to maximally explore the feature space.

Constructing the policy cover Finally, we describe how Algorithm 6 finds a policy cover π_{cov}. It is a procedure very similar to Algorithm 1 of (Wang et al., 2020). Note that the focus of (Wang et al., 2020) is reward-free exploration in linear MDPs, but it turns out that the same idea can be used for our purpose, and it is also related to the exploration strategy introduced in (Agarwal et al., 2020a; Zanette et al., 2021).

More specifically, PolicyCover interacts with the environment for $T_0 = M_0 N_0$ rounds. At the beginning of episode $(m-1)N_0 + 1$ for every $m = 1, \ldots, M_0$, it computes a policy π_m using the LSVI-UCB algorithm of (Jin et al., 2020b) but with a fake reward function Eq. (18) (ignoring the true loss feedback from the environment). This fake reward function is designed to encourage the learner to explore unseen state-action pairs and to ensure Eq. (17) eventually. For this purpose, we could have set the fake reward for (x,a) to be $\mathbf{1}\left[\|\phi(x,a)\|^2_{\Gamma^{-1}_{m,h}} \geq \frac{\alpha}{2M_0}\right]$. However, for technical reasons the analysis requires the reward function to be Lipschitz, and thus we approximate the indicator function above using a ramp function (with a large slope T). With π_m in hand, the algorithm then interacts with the environment for N_0 episodes, collecting trajectories to construct a good estimator of the covariance matrix of π_m. The design of the fake reward function and this extra step of covariance estimation are the only differences compared to Algorithm 1 of (Wang et al., 2020). At the end of the procedure, PolicyCover construct π_{cov} as a uniform mixture of $\{\pi_m\}_{m=1,\ldots,M_0}$. This means that whenever we execute π_{cov}, we first uniformly at random sample $m \in [M_0]$, and then execute the (pure) policy π_m.

Regret guarantee With all these elements, we successfully remove the need of a simulator and prove the following regret guarantee.

Theorem C.1. *Under Assumption 3, Algorithm 5 with appropriate choices of the parameters ensures* $\mathbb{E}[\text{Reg}] = \widetilde{\mathcal{O}}\left(d^2 H^4 T^{14/15}\right)$.

Although our regret rate is significantly worse than that in the full-information setting (Cai et al., 2020), in the stochastic setting (Zanette et al., 2021), or in the case when the transition is known (Neu & Olkhovskaya, 2021), we emphasize again that our algorithm is the first with provable sublinear regret guarantee for this challenging adversarial setting with bandit feedback and unknown transition.

D. Auxiliary Lemmas

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

Lemma D.1 ((A special form of) Freedman's inequality, Theorem 1 of (Beygelzimer et al., 2011))**.** *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, $\mathbb{E}[X_i|\mathcal{F}_i] = 0$, $|X_i| \leq b$, and*

Algorithm 6 PolicyCover $(M_0, N_0, \alpha, \delta)$

Let $\xi = 60dH\sqrt{\log(T/\delta)}$.
Let $\Gamma_{1,h}$ be the identity matrix in $\mathbb{R}^{d\times d}$ for all h.
for $m = 1, \ldots, M_0$ **do**
  Let $\widehat{V}_m(x_H) = 0$.
  for $h = H-1, H-2, \ldots, 0$ **do**
    For all $(x,a) \in X_h \times A$, compute

$$\widehat{Q}_m(x,a) = \min\left\{r_m(x,a) + \xi\|\phi(x,a)\|_{\Gamma_{m,h}^{-1}} + \phi(x,a)^\top\widehat{\theta}_{m,h},\ \ H\right\},$$
$$\widehat{V}_m(x) = \max_{a'}\widehat{Q}_m(x,a'),$$
$$\pi_m(a|x) = \mathbf{1}\left[a = \operatorname*{argmax}_{a'}\widehat{Q}_m(x,a')\right], \qquad \text{(break tie in argmax arbitrarily)}$$

    with

$$r_m(x,a) = \text{ramp}_{\frac{1}{T}}\left(\|\phi(x,a)\|^2_{\Gamma_{m,h}^{-1}} - \frac{\alpha}{M_0}\right), \tag{18}$$
$$\widehat{\theta}_{m,h} = \Gamma_{m,h}^{-1}\left(\frac{1}{N_0}\sum_{t=1}^{(m-1)N_0}\phi(x_{t,h},a_{t,h})\widehat{V}_m(x_{t,h+1})\right),$$

    where $\text{ramp}_z(y) = \begin{cases} 0 & \text{if } y \le -z, \\ 1 & \text{if } y \ge 0, \\ \frac{y}{z}+1 & \text{if } -z < y < 0. \end{cases}$
  for $t = (m-1)N_0+1, \ldots, mN_0$ **do**
    Execute π_m in episode t and collect trajectory $\{x_{t,h}, a_{t,h}\}_{h=0}^{H-1}$.
  Compute

$$\Gamma_{m+1,h} = \Gamma_{m,h} + \frac{1}{N_0}\sum_{t=(m-1)N_0+1}^{mN_0}\phi(x_{t,h},a_{t,h})\phi(x_{t,h},a_{t,h})^\top.$$

Let $\pi_{\text{cov}} = \text{Uniform}\left(\{\pi_m\}_{m=1}^{M_0}\right)$ and $\widehat{\Sigma}_h^{\text{cov}} = \frac{1}{M_0}\Gamma_{M_0+1,h}$ for all h.
return π_{cov} and $\left\{\widehat{\Sigma}_h^{\text{cov}}\right\}_{h=0,\ldots,H-1}$.

$\sum_{i=1}^n \mathbb{E}[X_i^2|\mathcal{F}_i] \le V$ *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 D.2 (Adapted from Lemma 11 of (Jin et al., 2020a))**.** *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 D.3 (Matrix Azuma, Theorem 7.1 of (Tropp, 2012))**.** *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*

$$\mathbb{E}_k[X_k] = 0 \ \textit{and} \ \ X_k^2 \preceq A_k^2 \ \textit{almost surely}$$

Define the variance parameter

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

Then, for all $\tau > 0$,

$$\Pr\left\{ \left\| \frac{1}{n} \sum_{k=1}^{n} X_k \right\|_{op} \geq \tau \right\} \leq de^{-n\tau^2/8\sigma^2}.$$

Next, we show a classic regret bound for the exponential weight algorithm, which can be found, for example, in (Luo, 2017).

Lemma D.4 (Regret bound of exponential weight, extracted from Theorem 1 of (Luo, 2017))**.** *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 $a \in A$:*

$$\begin{aligned}
\pi_1(a) &= \frac{1}{|A|}, \\
\pi_{t+1}(a) &= \frac{\pi_t(a)e^{-\eta \ell_t(a)}}{\sum_{a' \in A} \pi_t(a')e^{-\eta \ell_t(a')}}, \\
|\eta \ell_t(a)| &\leq 1.
\end{aligned}$$

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

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

E. Proofs Omitted in Section 3

In this section, we prove Lemma 3.1. In fact, we prove two generalized versions of it. Lemma E.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 E.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:*

$$\begin{aligned}
B_t(x, a) &= b_t(x, a) + \left(1 + \frac{1}{H}\right) \mathbb{E}_{x' \sim \widehat{P}(\cdot|x,a)} \mathbb{E}_{a' \sim \pi_t(\cdot|x')} \left[B_t(x', a')\right] \\
&\geq b_t(x, a) + \left(1 + \frac{1}{H}\right) \mathbb{E}_{x' \sim P(\cdot|x,a)} \mathbb{E}_{a' \sim \pi_t(\cdot|x')} \left[B_t(x', a')\right]
\end{aligned} \tag{19}$$

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 E.1. By rearranging Eq. (5), we see that

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

We first focus on **term**$_3$, and focus on a single layer $0 \le h \le H-1$ and a single t:

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

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

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

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 $x \in X_{h-1}$,

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

$$= \left(1 + \frac{1}{H}\right)^{H-h} \widehat{V}^{\pi_t}(x; b_t),$$

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

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

Lemma E.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:*

$$\mathbb{E}_t \left[B_t(x,a)\right] = b_t(x,a) + \left(1 + \frac{1}{H}\right) \mathbb{E}_{x' \sim P(\cdot|x,a)} \mathbb{E}_{a' \sim \pi_t(\cdot|x')} \mathbb{E}_t \Big[B_t(x', a')\Big] \tag{20}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\sum_x q^\star(x) \sum_{t=1}^T \sum_a \Big(\pi_t(a|x) - \pi^\star(a|x)\Big)\Big(Q_t^{\pi_t}(x,a) - B_t(x,a)\Big)\right] \\
&\qquad \leq o(T) + \mathbb{E}\left[\sum_{t=1}^T V^{\pi^\star}(x_0; b_t)\right] + \frac{1}{H}\mathbb{E}\left[\sum_{t=1}^T \sum_{x,a} q^\star(x)\pi_t(a|x) B_t(x,a)\right].
\end{aligned} \tag{21}$$

Then

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

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

F. Details Omitted in Section A

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

F.1. Implementation Details

The COMP-UOB procedure is the same as Algorithm 3 of (Jin et al., 2020a), which shows how to efficiently compute an upper occupancy bound. We include the algorithm in Algorithm 7 for completeness. As Algorithm 1 also needs COMP-LOB, which computes a lower occupancy bound, we provide its complete pseudocode in Algorithm 8 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. (22) subject to $\widehat{P}$ in the confidence set $\mathcal{P}$ via a greedy approach in Algorithm 9. The difference is that COMP-UOB sets OPTIMIZE to be $\max$ while COMP-LOB sets OPTIMIZE to be $\min$, and thus in Algorithm 9, $\{f(x)\}_{x \in X_k}$ 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 9. This concludes that the entire algorithm can be implemented efficiently.

$$f(\tilde{x}) = \sum_{a \in A} \pi_t(a|\tilde{x}) \left(\underset{\widehat{P}(\cdot|\tilde{x},a)}{\text{OPTIMIZE}} \sum_{x' \in X_{k(\tilde{x})+1}} \widehat{P}(x'|\tilde{x}, a) f(x') \right) \tag{22}$$

Algorithm 7 COMP-UOB (Algorithm 3 of (Jin et al., 2020a))

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

F.2. Omitted Proofs

To prove Theorem A.1, as discussed in the analysis sketch of Section A, we decompose the left-hand side of Eq. (5) as:

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{x} q^\star(x)\left\langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), Q_t^{\pi_t}(x,\cdot) - B_t(x,\cdot)\right\rangle \\
&= \underbrace{\sum_{t=1}^{T}\sum_{x} q^\star(x)\left\langle \pi_t(\cdot|x), Q_t^{\pi_t}(x,\cdot) - \widehat{Q}_t(x,\cdot)\right\rangle}_{\textsc{Bias-1}} + \underbrace{\sum_{t=1}^{T}\sum_{x} q^\star(x)\left\langle \pi^\star(\cdot|x), \widehat{Q}_t(x,\cdot) - Q_t^{\pi_t}(x,\cdot)\right\rangle}_{\textsc{Bias-2}} \\
&\quad + \underbrace{\sum_{t=1}^{T}\sum_{x} q^\star(x)\left\langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), \widehat{Q}_t(x,\cdot) - B_t(x,\cdot)\right\rangle}_{\textsc{Reg-Term}}.
\end{aligned} \tag{23}$$

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

Algorithm 9 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: $\sum_{j=1}^n \bar{p}(\sigma(j))f(\sigma(j))$

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

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

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 (Jin et al., 2020a, 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

$$\sum_{t=1}^T (\mathbb{E}_t[Y_t] - Y_t) + \left(\sum_x q^\star(x)\left\langle \pi_t(\cdot|x), Q_t^{\pi_t}(x,\cdot)\right\rangle - \mathbb{E}_t[Y_t]\right). \tag{24}$$

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

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

$$\le \sum_{x,a} q^\star(x)\pi_t(a|x)\frac{H^3}{\overline{q}_t(x,a)+\gamma} \qquad (q_t(s,a) \le \overline{q}_t(x,a))$$

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. (24) can be bounded as

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

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

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

where we use $\gamma = 2\eta H$. $\square$

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

Lemma F.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 D.2 with $z_t(x,a) = q^\star(x)\pi^\star(a|x)Q_t^{\pi_t}(x,a)$ and

$$Z_t(x,a) = q^\star(x)\pi^\star(a|x)\left(\mathbb{1}_t(x,a)L_t(x,a) + (1-\mathbb{1}_t(x,a))Q_t^{\pi_t}(x,a)\right).$$

Then we get that with probability at least $1-\delta$ (recalling the definition $\widehat{Q}_t(x,a) = \frac{L_{t,h}}{\overline{q}_t(x,a)+\gamma}\mathbb{1}_t(x,a)$),

$$\sum_{t=1}^T\sum_{x,a} q^\star(x)\pi^\star(a|x)\left(\widehat{Q}_t(x,a) - \frac{q_t(x,a)}{\overline{q}_t(x,a)}Q_t^{\pi_t}(x,a)\right) \le \frac{H^2}{2\gamma}\ln\frac{H}{\delta}, \tag{25}$$

Since with probability at least $1-4\delta$, $q_t(x,a) \le \overline{q}_t(x,a)$ for all t,x,a (by (Jin et al., 2020a, Lemma 2)), Eq. (25) further implies that with probability at least $1-5\delta$,

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

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

We continue to bound REG-TERM.

Lemma F.3 (REG-TERM)**.** *With probability at least* $1 - 5\delta$*,*

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

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 D.4 on each state x, and get

$$\sum_{t=1}^{T}\left\langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), \widehat{Q}_t(x,\cdot) - B_t(x,\cdot)\right\rangle \leq \frac{\ln|A|}{\eta} + \eta\sum_{t=1}^{T}\sum_{a\in A}\pi_t(a|x)\left(\widehat{Q}_t(x,a) - B_t(x,a)\right)^2. \tag{26}$$

The condition required by Lemma D.4 (i.e., $\eta|\widehat{Q}_t(x,a) - B_t(x,a)| \leq 1$) is verified in Lemma F.4. Summing Eq. (26) over states with weights $q^\star(x)$, we get

$$\begin{aligned}
\text{REG-TERM} &\leq \frac{H\ln|A|}{\eta} + \eta\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)\left(\widehat{Q}_t(x,a) - B_t(x,a)\right)^2 \\
&\leq \frac{H\ln|A|}{\eta} + 2\eta\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)\widehat{Q}_t(x,a)^2 + 2\eta\sum_{t=1}^{T}\sum_{x,a} q^\star(x)\pi_t(a|x)B_t(x,a)^2.
\end{aligned}\tag{27}$$

Below, we focus on the last two terms on the right-hand side of Eq. (27). First, we have

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

where the third step happens with probability at least $1-\delta$ by Lemma D.2 with $z_t(x,a) = Z_t(x,a) = \frac{q^\star(x)\pi_t(a|x)}{\overline{q}_t(x,a)+\gamma} \leq \frac{1}{\gamma}$, and the last step uses $\gamma = 2\eta H$ and $q_t(x,a) \leq \overline{q}_t(x,a)$ (which happens with probability at least $1-4\delta$). For the second term in Eq. (27), note that

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

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

In Lemma F.3, as required by Lemma D.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 F.4. $\eta\widehat{Q}_t(x,a) \leq \frac{1}{2}$ *and* $\eta B_t(x,a) \leq \frac{1}{2H}$.

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

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

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

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

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

This finishes the proof. $\square$

Now we are ready to prove Theorem A.1. For convenience, we state the theorem again here and show the proof.

Theorem F.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 $1-\mathcal{O}(\delta)$,

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

which is of the form specified in Eq. (5). By the definition of $B_t(x,a)$ in Eq. (9), we see that Eq. (19) also holds with probability at least $1-\mathcal{O}(\delta)$ for all t,x,a.

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

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

where the last inequality is due to (Jin et al., 2020a, Lemma 4). Plugging in the specified value for η, the regret can be further upper bounded by $\widetilde{\mathcal{O}}\left(|X|H^2\sqrt{AT} + H^4\right)$. $\square$

G. Analysis for Auxiliary Procedures

In this section, we analyze two important auxiliary procedures for the linear function approximation settings: GEOMETRICRESAMPLING and POLICYCOVER.

G.1. The Guarantee of GEOMETRICRESAMPLING

The GEOMETRICRESAMPLING algorithm is shown in Algorithm 4, which is almost the same as that in (Neu & Olkhovskaya, 2020) 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. The following lemma summarizes some useful guarantees of this procedure. For generality, we present the lemma assuming a lower bound on the minimum eigenvalue λ of the covariance matrix, but it will simply be 0 in all our applications of this lemma in this work.

Lemma G.1. *Let π be a policy (possibly a mixture policy) with a covariance matrix $\Sigma_h = \mathbb{E}_\pi[\phi(x_h, a_h)\phi(x_h, a_h)^\top] \succeq \lambda I$ for layer h and some constant $\lambda \geq 0$. Further let $\epsilon > 0$ and $\gamma \geq 0$ be two parameters satisfying $0 < \gamma + \lambda < 1$. Define $M = \left\lceil \frac{24\ln(dHT)}{\epsilon^2} \min\left\{\frac{1}{\gamma^2}, \frac{4}{\lambda^2}\ln^2\frac{1}{\epsilon\lambda}\right\}\right\rceil$ and $N = \left\lceil \frac{2}{\gamma+\lambda}\ln\frac{1}{\epsilon(\gamma+\lambda)}\right\rceil$. Let $\mathcal{T}$ be a set of MN trajectories generated by π. Then* GeometricResampling *(Algorithm 4) with input $(\mathcal{T}, M, N, \gamma)$ ensures the following for all h:*

$$\left\|\widehat{\Sigma}_h^+\right\|_{\text{op}} \leq \min\left\{\frac{1}{\gamma}, \frac{2}{\lambda}\ln\frac{1}{\epsilon\lambda}\right\}. \tag{28}$$

$$\left\|\mathbb{E}\left[\widehat{\Sigma}_h^+\right] - (\gamma I + \Sigma_h)^{-1}\right\|_{\text{op}} \leq \epsilon, \tag{29}$$

$$\left\|\widehat{\Sigma}_h^+ - (\gamma I + \Sigma_h)^{-1}\right\|_{\text{op}} \leq 2\epsilon, \tag{30}$$

$$\left\|\widehat{\Sigma}_h^+\Sigma_h\right\|_{\text{op}} \leq 1 + 2\epsilon, \tag{31}$$

where $\|\cdot\|_{\text{op}}$ represents the spectral norm and the last two properties Eq. (30) and Eq. (31) hold with probability at least $1 - \frac{1}{T^3}$.

Proof. To prove Eq. (28), notice that each one of $\widehat{\Sigma}_h^{+(m)}$, $m = 1, \ldots, M$, is a sum of $N+1$ terms. Furthermore, the n-th term of them ($cZ_{n,h}$ in Algorithm 4) has an operator norm upper bounded by $c(1-c\gamma)^n$. Therefore,

$$\left\|\widehat{\Sigma}_h^{+(m)}\right\|_{\text{op}} \leq \sum_{n=0}^{N} c(1-c\gamma)^n \leq \min\left\{\frac{1}{\gamma}, c(N+1)\right\} \leq \min\left\{\frac{1}{\gamma}, \frac{2}{\lambda}\ln\frac{1}{\epsilon\lambda}\right\} \tag{32}$$

by the definition of N and that $c = \frac{1}{2}$. Since $\widehat{\Sigma}_h^+$ is an average of $\widehat{\Sigma}_h^{+(m)}$, this implies Eq. (28).

To show Eq. (29), observe that $\mathbb{E}_t[Y_{n,h}] = \gamma I + \Sigma_h$ and $\{Y_{n,h}\}_{n=1}^N$ are independent. Therefore, we a have

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

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,

$$\begin{aligned}
\left\|\mathbb{E}_t\left[\widehat{\Sigma}_h^+\right] - (\gamma I + \Sigma_h)^{-1}\right\|_{\text{op}} &= \left\|(\gamma I + \Sigma_h)^{-1}\left(I - c\left(\gamma I + \Sigma_h\right)\right)^{N+1}\right\|_{\text{op}} \\
&\leq \frac{(1 - c(\gamma+\lambda))^{N+1}}{\gamma+\lambda} \leq \frac{e^{-(N+1)c(\gamma+\lambda)}}{\gamma+\lambda} \leq \epsilon,
\end{aligned}$$

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

To show Eq. (30), we only further need

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

and combine it with Eq. (29). This can be shown by applying Lemma D.3 with $X_k = \widehat{\Sigma}_h^{+(k)} - \mathbb{E}\left[\widehat{\Sigma}_h^{+(k)}\right]$, $A_k = \min\left\{\frac{1}{\gamma}, \frac{2}{\lambda}\ln\frac{1}{\epsilon\lambda}\right\} I$ (recall Eq. (32) and thus $X_k^2 \preceq A_k^2$), $\sigma = \min\left\{\frac{1}{\gamma}, \frac{2}{\lambda}\ln\frac{1}{\epsilon\lambda}\right\}$, $\tau = \epsilon$, and $n = M$. This gives the following statement: the event $\left\|\widehat{\Sigma}_h^+ - \mathbb{E}_t\left[\widehat{\Sigma}_h^+\right]\right\|_{\text{op}} > \epsilon$ holds with probability less than

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

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

To prove Eq. (31), observe that with Eq. (30), we have

$$\left\|\widehat{\Sigma}_h^+\Sigma_h\right\|_{\text{op}} \le \left\|(\gamma I+\Sigma_h)^{-1}\Sigma_h\right\|_{\text{op}} + \left\|\left(\widehat{\Sigma}_h^+ - (\gamma I+\Sigma_h)^{-1}\right)\Sigma_h\right\|_{\text{op}} \le 1+2\epsilon$$

since $\|\Sigma_h\|_{\text{op}} \le 1$. □

G.2. The Guarantee of PolicyCover

In this section, we analyze Algorithm 6, which returns a policy cover and its estimated covariance matrices. The final guarantee of the policy cover is provided in Lemma G.4, but we need to establish a couple of useful lemmas before introducing that. Note that Algorithm 6 bears some similarity with (Wang et al., 2020, Algorithm 1) (except for the design of the reward function r_t), and thus the analysis is also similar to theirs.

We first define the following definitions, using notations defined in Algorithm 6 and Assumption 3.

Definition 1. *For any π and m, define V_m^π to be the state value function for π with respect to reward function r_m. Precisely, this means $V_m^\pi(x_H)=0$ and for $(x,a)\in X_h\times A$, $h=H-1,\dots,0$: $V_m^\pi(x)=\sum_a \pi(a|x)Q_m^\pi(x,a)$ where*

$$Q_m^\pi(x,a)=r_m(x,a)+\phi(x,a)^\top\theta_{m,h}^\pi \quad \textit{and} \quad \theta_{m,h}^\pi=\int_{x'\in X_{h+1}} V_m^\pi(x')\nu_h^{x'}\mathrm{d}x'.$$

Furthermore, let π_m^* be the optimal policy satisfying $\pi_m^*=\operatorname{argmax}_\pi V_m^\pi(x)$ for all x, and define shorthands $V_m^*(x)=V_m^{\pi_m^*}(x)$, $Q_m^*(x,a)=Q_m^{\pi_m^*}(x,a)$, and $\theta_{m,h}^*=\theta_{m,h}^{\pi_m^*}$.

The following lemma characterizes the optimistic nature of Algorithm 6.

Lemma G.2. *With probability at least $1-\delta$, for all h, all $(x,a)\in X_h\times A$, and all π, Algorithm 6 ensures*

$$0\le \widehat{Q}_m(x,a)-Q_m^\pi(x,a)\le \mathbb{E}_{x'\sim P(\cdot|x,a)}\left[\widehat{V}_m(x')-V_m^\pi(x')\right]+2\xi\|\phi(x,a)\|_{\Gamma_{m,h}^{-1}}.$$

Proof. The proof mostly follows that of (Wei et al., 2021, Lemma 4). For notational convenience, denote $\phi(x_{\tau,h},a_{\tau,h})$ as $\phi_{\tau,h}$, and $x'\sim P(\cdot|x_{\tau,h},a_{\tau,h})$ as $x'\sim(\tau,h)$. We then have

$$\begin{aligned}
&\widehat{\theta}_{m,h}-\theta_{m,h}^\pi\\
&=\Gamma_{m,h}^{-1}\left(\frac{1}{N_0}\sum_{\tau=1}^{(m-1)N_0}\phi_{\tau,h}\widehat{V}_m(x_{\tau,h+1})\right)-\Gamma_{m,h}^{-1}\left(\theta_{m,h}^\pi+\frac{1}{N_0}\sum_{\tau=1}^{(m-1)N_0}\phi_{\tau,h}\phi_{\tau,h}^\top\theta_{m,h}^\pi\right)\\
&=\Gamma_{m,h}^{-1}\left(\frac{1}{N_0}\sum_{\tau=1}^{(m-1)N_0}\phi_{\tau,h}\widehat{V}_m(x_{\tau,h+1})\right)-\Gamma_{m,h}^{-1}\left(\frac{1}{N_0}\sum_{\tau=1}^{(m-1)N_0}\phi_{\tau,h}\mathbb{E}_{x'\sim(\tau,h)}\left[V_m^\pi(x')\right]\right)-\Gamma_{m,h}^{-1}\theta_{m,h}^\pi\\
&=\Gamma_{m,h}^{-1}\left(\frac{1}{N_0}\sum_{\tau=1}^{(m-1)N_0}\phi_{\tau,h}\mathbb{E}_{x'\sim(\tau,h)}\left[\widehat{V}_m(x')-V_m^\pi(x')\right]\right)+\zeta_{m,h}-\Gamma_{t,h}^{-1}\theta_{t,h}^\pi\\
&\qquad\qquad\left(\text{define } \zeta_{m,h}=\tfrac{1}{N_0}\Gamma_{m,h}^{-1}\textstyle\sum_{\tau=1}^{(m-1)N_0}\left(\widehat{V}_m(x_{\tau,h+1})-\mathbb{E}_{x'\sim(\tau,h)}\widehat{V}_m(x')\right)\right)\\
&=\Gamma_{m,h}^{-1}\left(\frac{1}{N_0}\sum_{\tau=1}^{(m-1)N_0}\phi_{\tau,h}\phi_{\tau,h}^\top\int_{x'\in X_{h+1}}\nu_h^{x'}\left(\widehat{V}_m(x')-V_m^\pi(x')\right)\mathrm{d}x'\right)+\zeta_{m,h}-\Gamma_{m,h}^{-1}\theta_{m,h}^\pi\\
&=\int_{x'\in X_{h+1}}\nu_h^{x'}\left(\widehat{V}_m(x')-V_m^\pi(x')\right)\mathrm{d}x'+\zeta_{m,h}-\Gamma_{m,h}^{-1}\theta_{m,h}^\pi-\Gamma_{m,h}^{-1}\int_{x'\in X_{h+1}}\nu_h^{x'}\left(\widehat{V}_m(x')-V_m^\pi(x')\right)\mathrm{d}x'.
\end{aligned}$$

Therefore, for $x\in X_h$,

$$\widehat{Q}_m(x,a)-Q_m^\pi(x,a)$$

$$
\begin{aligned}
&= \phi(x,a)^\top \left(\widehat{\theta}_{m,h} - \theta^\pi_{m,h}\right) + \xi\|\phi(x,a)\|_{\Gamma^{-1}_{m,h}} \\
&= \phi(x,a)^\top \int_{x'\in X_{h+1}} \nu_h^{x'} \left(\widehat{V}_m(x') - V^\pi_m(x')\right) \mathrm{d}x' + \underbrace{\phi(x,a)^\top \zeta_{m,h}}_{\textbf{term}_1} + \xi\|\phi(x,a)\|_{\Gamma^{-1}_{m,h}} \\
&\qquad \underbrace{-\phi(x,a)^\top \Gamma^{-1}_{m,h} \int_{x'\in X_{h+1}} \nu_h^{x'} \left(\widehat{V}_m(x') - V^\pi_m(x')\right) \mathrm{d}x'}_{\textbf{term}_2} \underbrace{-\phi(x,a)^\top \Gamma^{-1}_{m,h}\theta^\pi_{m,h}}_{\textbf{term}_3} \\
&= \mathbb{E}_{x'\sim p(\cdot|x,a)} \left[\widehat{V}_m(x') - V^\pi_m(x')\right] + \xi \|\phi(x,a)\|_{\Gamma^{-1}_{m,h}} + \textbf{term}_1 + \textbf{term}_2 + \textbf{term}_3. \qquad (33)
\end{aligned}
$$

It remains to bound $|\textbf{term}_1 + \textbf{term}_2 + \textbf{term}_3|$. To do so, we follow the exact same arguments as in (Wei et al., 2021, Lemma 4) to bound each of the three terms.

Bounding $\textbf{term}_1$. First we have $|\textbf{term}_1| \le \|\zeta_{m,h}\|_{\Gamma_{m,h}} \|\phi(x,a)\|_{\Gamma^{-1}_{m,h}}$. To bound $\|\zeta_{m,h}\|_{\Gamma_{m,h}}$, we use the exact same argument of (Wei et al., 2021, Lemma 4) to arrive at (with probability at least $1-\delta$)

$$
\begin{aligned}
\|\zeta_{m,h}\|_{\Gamma_{m,h}} &= \left\| \frac{1}{N_0} \sum_{\tau=1}^{(m-1)N_0} \left(\widehat{V}_m(x_{\tau,h+1}) - \mathbb{E}_{x'\sim(\tau,h)} \widehat{V}_m(x')\right)\right\|_{\Gamma^{-1}_{m,h}} \\
&\le 2H\sqrt{\frac{d}{2}\log(M_0+1) + \log\frac{\mathcal{N}_\varepsilon}{\delta}} + \sqrt{8M_0^2\varepsilon^2}, \qquad (34)
\end{aligned}
$$

where $\mathcal{N}_\varepsilon$ is the ε-cover of the function class that $\widehat{V}_m(\cdot)$ lies in. Notice that for all m, $\widehat{V}_m(\cdot)$ can be expressed as the following:

$$
\widehat{V}_m(x) = \min\left\{ \max_a \left\{ \mathrm{ramp}_{\frac{1}{T}} \left(\|\phi(x,a)\|_Z^2 - \frac{\alpha}{M_0}\right) + \xi\|\phi(x,a)\|_Z + \phi(x,a)^\top\theta \right\},\ \ H \right\}
$$

for some positive definite matrix $Z \in \mathbb{R}^{d\times d}$ with $\frac{1}{1+M_0} I \preceq Z \preceq I$ and vector $\theta\in\mathbb{R}^d$ with $\|\theta\| \le \sup_{m,\tau,h} \left\|\Gamma^{-1}_{m,h}\right\|_{\mathrm{op}} \times M_0\|\phi_{\tau,h}\|H \le M_0 H$. Therefore, we can write the class of functions that $\widehat{V}_m(\cdot)$ lies in as the following set:

$$
\begin{aligned}
\mathcal{V} = \Bigg\{ &\min\left\{ \max_a \left\{ \mathrm{ramp}_{\frac{1}{T}} \left(\|\phi(x,a)\|_Z^2 - \frac{\alpha}{M_0}\right) + \xi\|\phi(x,a)\|_Z + \phi(x,a)^\top\theta \right\},\ \ H \right\} : \\
&\theta\in\mathbb{R}^d : \|\theta\| \le M_0 H,\ \ Z\in\mathbb{R}^{d\times d} : \frac{1}{1+M_0} I \preceq Z \preceq I \Bigg\}.
\end{aligned}
$$

Now we apply Lemma 12 of (Wei et al., 2021) to $\mathcal{V}$, with the following choices of parameters: $P = d^2 + d$, $\varepsilon = \frac{1}{T^3}$, $B = M_0 H$, and $L = T + \xi\sqrt{1+M_0} + 1 \le 3T$ (without loss of generality, we assume that T is large enough so that the last inequality holds). The value of the Lipschitzness parameter L is according to the following calculation that is similar to (Wei et al., 2021): for any $\Delta Z = \epsilon \mathbf{e}_i \mathbf{e}_j^\top$,

$$
\begin{aligned}
&\frac{1}{|\epsilon|} \left| \sqrt{\phi(x,a)^\top (Z+\Delta Z)\phi(x,a)} - \sqrt{\phi(x,a)^\top Z\phi(x,a)} \right| \\
&\le \frac{\left|\phi(x,a)^\top \mathbf{e}_i\mathbf{e}_j^\top \phi(x,a)\right|}{\sqrt{\phi(x,a)^\top Z\phi(x,a)}} \qquad \left(\sqrt{u+v} - \sqrt{u} \le \tfrac{|v|}{\sqrt{u}}\right) \\
&\le \frac{\phi(x,a)^\top \left(\frac{1}{2}\mathbf{e}_i\mathbf{e}_i^\top + \frac{1}{2}\mathbf{e}_j\mathbf{e}_j^\top\right)\phi(x,a)}{\sqrt{\phi(x,a)^\top Z\phi(x,a)}} \\
&\le \frac{\phi(x,a)^\top\phi(x,a)}{\sqrt{\phi(x,a)^\top Z\phi(x,a)}} \le \sqrt{\frac{1}{\lambda_{\min}(Z)}} \le \sqrt{1+M_0};
\end{aligned}
$$

$\frac{1}{|\epsilon|}\left|\|\phi(x,a)\|_{Z+\Delta Z}^2 - \|\phi(x,a)\|_Z^2\right| = |\mathbf{e}_i^\top \phi(x,a)\phi(x,a)^\top \mathbf{e}_j| \leq 1$; and that $\text{ramp}_{\frac{1}{T}}(\cdot)$ has a slope of T (this is why we need to use the ramp function to approximate an indication function that is not Lipschitz). Overall, this leads to $\log \mathcal{N}_\varepsilon \leq 20(d^2+d)\log T$. Using this fact in Eq. (34), we get

$$\|\zeta_{m,h}\|_{\Gamma_{m,h}} \leq 20H\sqrt{d^2 \log\left(\frac{T}{\delta}\right)} \leq \frac{1}{3}\xi,$$

and thus $|\mathbf{term}_1| \leq \frac{\xi}{3}\|\phi(x,a)\|_{\Gamma_{m,h}^{-1}}$.

Bounding $\mathbf{term}_2$ and $\mathbf{term}_3$. This is exactly the same as (Wei et al., 2021, Lemma 4), and we omit the details. In summary, we can also prove $|\mathbf{term}_2| \leq \frac{\xi}{3}\|\phi(x,a)\|_{\Gamma_{m,h}^{-1}}$ and $\mathbf{term}_3 \leq \frac{\xi}{3}\|\phi(x,a)\|_{\Gamma_{m,h}^{-1}}$.

In sum, we can bound

$$|\mathbf{term}_1 + \mathbf{term}_2 + \mathbf{term}_3| \leq |\mathbf{term}_1| + |\mathbf{term}_2| + |\mathbf{term}_3| \leq \xi\|\phi(x,a)\|_{\Gamma_{m,h}^{-1}}$$

for all m, h and (s,a) with probability at least $1-\delta$.

Combining this with Eq. (33), we get

$$\widehat{Q}_m(x,a) - Q_m^\pi(x,a) \leq \mathbb{E}_{x'\sim p(\cdot|x,a)}\left[\widehat{V}_m(x') - V_m^\pi(x')\right] + 2\xi\|\phi(x,a)\|_{\Gamma_{m,h}^{-1}}, \tag{35}$$

$$\widehat{Q}_m(x,a) - Q_m^\pi(x,a) \geq \mathbb{E}_{x'\sim p(\cdot|x,a)}\left[\widehat{V}_m(x') - V_m^\pi(x')\right], \tag{36}$$

where Eq. (35) proves the second inequality in the lemma. To prove the first inequality in the lemma, we use and induction to show that $\widehat{V}_m(x) \geq V_m^\pi(x)$ for all x, which combined with Eq. (36) finishes the proof. Recall that we define $\widehat{V}_m(x_H) = V_m^\pi(x_H) = 0$. Assume that $\widehat{V}_m(x) \geq V_m^\pi(x)$ holds for $x \in X_{h+1}$. Then by Eq. (36), $\widehat{Q}_m(x,a) - Q_m^\pi(x,a) \geq 0$ for all $(x,a) \in X_h \times A$. Thus, $\widehat{V}_m(x) - V_m^\pi(x) = \max_a \widehat{Q}_m(x,a) - \sum_a \pi(a|x)Q_m^\pi(x,a) \geq 0$, finishing the induction. □

The next lemma provides a "regret guarantee" for Algorithm 6 with respect to the fake rewards.

Lemma G.3. *With probability at least $1-2\delta$, Algorithm 6 ensures*

$$\sum_{m=1}^{M_0} V_m^*(x_0) - \sum_{m=1}^{M_0} V_m^{\pi_m}(x_0) = \widetilde{\mathcal{O}}\left(d^{3/2}H^2\sqrt{M_0}\right).$$

Proof. For any $t \in [(m-1)N_0+1, mN_0]$ and any h,

$$\begin{aligned}
&\widehat{V}_m(x_{t,h}) - V_m^{\pi_m}(x_{t,h}) \\
&= \max_a \widehat{Q}_m(x_{t,h},a) - Q_m^{\pi_m}(x_{t,h},a_{t,h}) && (\pi_m \text{ is a deterministic policy}) \\
&= \widehat{Q}_m(x_{t,h},a_{t,h}) - Q_m^{\pi_m}(x_{t,h},a_{t,h}) \\
&\leq \mathbb{E}_{x'\sim(x_{t,h},a_{t,h})}\left[\widehat{V}_m(x') - V_m^{\pi_m}(x')\right] + 2\xi\|\phi(x_{t,h},a_{t,h})\|_{\Gamma_{m,h}^{-1}} && (\text{Lemma G.2}) \\
&= \widehat{V}_m(x_{t,h+1}) - V_m^{\pi_m}(x_{t,h+1}) + e_{t,h} + 2\xi\|\phi(x_{t,h},a_{t,h})\|_{\Gamma_{m,h}^{-1}}. && (\text{define } e_{t,h} \text{ to be the difference})
\end{aligned}$$

Thus,

$$\widehat{V}_m(x_0) - V_m^{\pi_m}(x_0) \leq \sum_h \left(2\xi\|\phi(x_{t,h},a_{t,h})\|_{\Gamma_{m,h}^{-1}} + e_{t,h}\right).$$

Summing over t, and using the fact $V_m^*(x_0) \leq \widehat{V}_m(x_0)$ (from Lemma G.2), we get

$$\frac{1}{M_0}\sum_{m=1}^{M_0}\left(V_m^*(x_0) - V_m^{\pi_m}(x_0)\right)$$

$$\le \frac{1}{M_0N_0}\sum_{t=1}^{M_0N_0}\sum_h \left(2\xi\left\|\phi(x_{t,h},a_{t,h})\right\|_{\Gamma_{m,h}^{-1}} + e_{t,h}\right)$$

$$\le \frac{2\xi}{\sqrt{M_0N_0}}\sum_h \sqrt{\sum_{t=1}^{M_0N_0}\left\|\phi(x_{t,h},a_{t,h})\right\|^2_{\Gamma_{m,h}^{-1}}} + \frac{1}{M_0N_0}\sum_{t=1}^{M_0N_0}\sum_h e_{t,h}. \qquad \text{(Cauchy-Schwarz inequality)}$$

Further using the fact $\sum_{t=1}^{M_0N_0}\left\|\phi(x_{t,h},a_{t,h})\right\|^2_{\Gamma_{m,h}^{-1}} = N_0\sum_{m=1}^{M_0}\left\langle \Gamma_{m+1,h}-\Gamma_{m,h}, \Gamma_{m,h}^{-1}\right\rangle = \widetilde{\mathcal{O}}\left(N_0 d\right)$ (see e.g., (Jin et al., 2020b, Lemma D.2)), we bound the first term above by $\widetilde{\mathcal{O}}\left(\xi H\sqrt{d/M_0}\right) = \widetilde{\mathcal{O}}\left(H^2\sqrt{d^3/M_0}\right)$. For the second term, note that $\sum_{t=1}^{M_0N_0} e_{t,h}$ is the sum of a martingale difference sequence. By Azuma's inequality, the entire second term is thus of order $\widetilde{\mathcal{O}}\left(\frac{H^2\log(1/\delta)}{\sqrt{M_0N_0}}\right)$ with probability at least $1-\delta$. This finishes the proof. $\square$

Finally, we are ready to show the guarantee of the returned policy cover. Recall our definition of known state set:

$$\mathcal{K} = \left\{x \in X : \forall a \in A, \left\|\phi(x,a)\right\|^2_{(\widehat{\Sigma}_h^{\text{cov}})^{-1}} \le \alpha \text{ where } h \text{ is such that } x \in X_h\right\}.$$

Lemma G.4. *For any $h = 0, \ldots, H-1$, with probability at least $1-4\delta$ (over the randomness in the first T_0 rounds), the covariance matrices $\widehat{\Sigma}_h^{cov}$ returned by Algorithm 6 satisfies that for any policy π,*

$$\Pr_{x_h\sim\pi}\left[x_h \notin \mathcal{K}\right] \le \widetilde{\mathcal{O}}\left(\frac{dH}{\alpha}\right).$$

where $x_h \in X_h$ is sampled from executing π.

Proof. We define an auxiliary policy π' which only differs from π for unknown states in layer h. Specifically, for $x \in X_h$ not in $\mathcal{K}$, let a be such that $\left\|\phi(x,a)\right\|^2_{(\widehat{\Sigma}_h^{\text{cov}})^{-1}} \ge \alpha$ (which must exist by the definition of $\mathcal{K}$), then $\pi'(a'|x) = \mathbb{1}[a'=a]$ for all $a' \in A$. By doing so, we have

$$\Pr_{x_h\sim\pi}\left[x_h \notin \mathcal{K}\right]$$

$$= \Pr_{(x_h,a)\sim\pi'}\left[\left\|\phi(x_h,a)\right\|^2_{(\widehat{\Sigma}_h^{\text{cov}})^{-1}} \ge \alpha\right]$$

$$= \Pr_{(x_h,a)\sim\pi'}\left[\left\|\phi(x_h,a)\right\|^2_{\Gamma_{M_0+1,h}^{-1}} \ge \frac{\alpha}{M_0}\right]$$

$$\le \frac{1}{M_0}\sum_{m=1}^{M_0}\Pr_{(x_h,a)\sim\pi'}\left[\left\|\phi(x_h,a)\right\|^2_{\Gamma_{m,h}^{-1}} \ge \frac{\alpha}{M_0}\right] \qquad (\Gamma_{m,h} \preceq \Gamma_{M_0+1,h})$$

$$\le \frac{1}{M_0}\sum_{m=1}^{M_0}\mathbb{E}_{(x_h,a)\sim\pi'}\left[\text{ramp}_{\frac{1}{T}}\left(\left\|\phi(x,a)\right\|^2_{\Gamma_{m,h}^{-1}} - \frac{\alpha}{M_0}\right)\right] \qquad (\mathbf{1}[y\ge 0] \le \text{ramp}_z(y))$$

$$\le \frac{1}{M_0}\sum_{m=1}^{M_0} V_m^{\pi'}(x_0) \qquad \text{(rewards } r_m(\cdot,\cdot) \text{ are non-negative)}$$

$$\le \frac{1}{M_0}\sum_{m=1}^{M_0} V_m^{\pi_m}(x_0) + \frac{1}{M_0}\times\widetilde{\mathcal{O}}\left(d^{3/2}H^2\sqrt{M_0}\right) \qquad \text{(Lemma G.3)}$$

$$\le \frac{1}{M_0N_0}\sum_{t=1}^{M_0N_0}\sum_{h=0}^{H-1} r_m(x_{t,h},a_{t,h}) + \widetilde{\mathcal{O}}\left(\frac{H}{\sqrt{M_0N_0}}\right) + \widetilde{\mathcal{O}}\left(\frac{d^{3/2}H^2}{\sqrt{M_0}}\right) \qquad \text{(by Azuma's inequality)}$$

$$\le \frac{1}{M_0N_0}\times\frac{1}{\frac{\alpha}{M_0}}\sum_{t=1}^{M_0N_0}\sum_{h=0}^{H-1}\left\|\phi(x_{t,h},a_{t,h})\right\|^2_{\Gamma_{m,h}^{-1}} + \widetilde{\mathcal{O}}\left(\frac{d^{3/2}H^2}{\sqrt{M_0}}\right) \qquad (\text{ramp}_z(y-y') \le \tfrac{y}{y'} \text{ for } y>0, y'>z>0)$$

$$\le \frac{1}{M_0N_0}\times\frac{1}{\frac{\alpha}{M_0}}\times\widetilde{\mathcal{O}}\left(N_0 dH\right) + \widetilde{\mathcal{O}}\left(\frac{d^{3/2}H^2}{\sqrt{M_0}}\right) \qquad \text{(same calculation as done in the proof of Lemma G.3)}$$

$$\leq \widetilde{\mathcal{O}}\left(\frac{dH}{\alpha}+\frac{d^{3/2}H^2}{\sqrt{M_0}}\right).$$

Finally, using the definition of M_0 finishes the proof. $\square$

H. Details Omitted in Section B

In this section, we analyze Algorithm 2 and prove Theorem B.1. In the analysis, we require that $\pi_t(a|x)$ and $B_t(x,a)$ are 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 (the 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 A, 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:

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{x_h\sim\pi^\star}\left[\langle\pi_t(\cdot|x_h)-\pi^\star(\cdot|x_h),Q_t^{\pi_t}(x_h,\cdot)-B_t(x_h,\cdot)\rangle\right]\\
&=\underbrace{\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{x_h\sim\pi^\star}\left[\left\langle\pi_t(\cdot|x_h),Q_t^{\pi_t}(x_h,\cdot)-\widehat{Q}_t(x_h,\cdot)\right\rangle\right]}_{\textsc{Bias-1}}+\underbrace{\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{x_h\sim\pi^\star}\left[\left\langle\pi^\star(\cdot|x_h),\widehat{Q}_t(x_h,\cdot)-Q_t^{\pi_t}(x_h,\cdot)\right\rangle\right]}_{\textsc{Bias-2}}\\
&\quad+\underbrace{\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{x_h\sim\pi^\star}\left[\left\langle\pi_t(\cdot|x_h)-\pi^\star(\cdot|x_h),\widehat{Q}_t(x_h,\cdot)-B_t(x_h,\cdot)\right\rangle\right]}_{\textsc{Reg-Term}}.
\end{aligned}$$

We then bound $\mathbb{E}[\textsc{Bias-1}+\textsc{Bias-2}]$ and $\mathbb{E}[\textsc{Reg-Term}]$ in Lemma H.1 and Lemma H.2 respectively.

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

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

Proof of Lemma H.1. Consider a specific (t,x,a). Let h be such that $x\in X_h$. Then we proceed as

$$\begin{aligned}
&\mathbb{E}_t\left[Q_t^{\pi_t}(x,a)-\widehat{Q}_t(x,a)\right]\\
&=\phi(x,a)^\top\left(\theta_{t,h}^{\pi_t}-\mathbb{E}_t\left[\widehat{\theta}_{t,h}\right]\right)\\
&=\phi(x,a)^\top\left(\theta_{t,h}^{\pi_t}-\mathbb{E}_t\left[\widehat{\Sigma}^+_{t,h}\right]\mathbb{E}_t\left[\phi(x_{t,h},a_{t,h})L_{t,h}\right]\right) && \text{(definition of } \widehat{\theta}_{t,h})
\end{aligned}$$

$$
\begin{aligned}
&= \phi(x,a)^\top \left(\theta^{\pi_t}_{t,h} - (\gamma I + \Sigma_{t,h})^{-1}\,\mathbb{E}_t\left[\phi(x_{t,h},a_{t,h})L_{t,h}\right]\right) + \mathcal{O}(\epsilon H) \\
&\qquad\qquad \text{(by Eq. (29) of Lemma G.1 and that } \|\phi(x,a)\| \le 1 \text{ for all } x,a \text{ and } L_{t,h} \le H) \\
&= \phi(x,a)^\top \left(\theta^{\pi_t}_{t,h} - (\gamma I + \Sigma_{t,h})^{-1}\,\Sigma_{t,h}\theta^{\pi_t}_{t,h}\right) + \mathcal{O}(\epsilon H) && (\mathbb{E}[L_{t,h}] = \phi(x_{t,h},a_{t,h})^\top \theta^{\pi_t}_{t,h}) \\
&= \gamma\phi(x,a)^\top (\gamma I + \Sigma_{t,h})^{-1}\theta^{\pi_t}_{t,h} + \mathcal{O}(\epsilon H) && (\theta^{\pi_t}_{t,h} = (\gamma I + \Sigma_{t,h})^{-1}(\gamma I + \Sigma_{t,h})\,\theta^{\pi_t}_{t,h}) \\
&\le \gamma \|\phi(x,a)\|^2_{(\gamma I + \Sigma_{t,h})^{-1}} \|\theta^{\pi_t}_{t,h}\|^2_{(\gamma I + \Sigma_{t,h})^{-1}} + \mathcal{O}(\epsilon H) && \text{(Cauchy-Schwarz inequality)} \\
&\le \frac{\beta}{4}\|\phi(x,a)\|^2_{(\gamma I + \Sigma_{t,h})^{-1}} + \frac{\gamma^2}{\beta}\|\theta^{\pi_t}_{t,h}\|^2_{(\gamma I + \Sigma_{t,h})^{-1}} + \mathcal{O}(\epsilon H) && \text{(AM-GM inequality)} \\
&\le \frac{\beta}{4}\mathbb{E}_t\left[\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}}\right] + \frac{\gamma d H^2}{\beta} + \mathcal{O}\left(\epsilon(H+\beta)\right) && (37)
\end{aligned}
$$

where in the last inequality we use Eq. (29) again and also $\|\theta^{\pi}_{t,h}\|^2 \le dH^2$ according to Assumption 1. Taking expectation over x and summing over t,a with weights $\pi_t(a|x)$, we get

$$
\mathbb{E}[\text{BIAS-1}] \le \frac{\beta}{4}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{x_h\sim\pi^\star}\left[\sum_a \pi_t(a|x_h)\|\phi(x_h,a)\|^2_{\widehat{\Sigma}^+_{t,h}}\right]\right] + \mathcal{O}\left(\frac{\gamma d H^3 T}{\beta} + \epsilon H^2 T\right). \qquad (\text{using } \beta \le H)
$$

By the same argument, we can show that $\mathbb{E}_t[\widehat{Q}_t(x,a) - Q^{\pi_t}_t(x,a)]$ is also upper bounded by the right-hand side of Eq. (37), and thus

$$
\mathbb{E}[\text{BIAS-2}] \le \frac{\beta}{4}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{x_h\sim\pi^\star}\left[\sum_a \pi^\star(a|x_h)\|\phi(x_h,a)\|^2_{\widehat{\Sigma}^+_{t,h}}\right]\right] + \mathcal{O}\left(\frac{\gamma d H^3 T}{\beta} + \epsilon H^2 T\right).
$$

Summing them up finishes the proof. $\square$

Lemma H.2. *If $\eta\beta \le \frac{\gamma}{12H^2}$ and $\eta \le \frac{\gamma}{2H}$, then $\mathbb{E}[\text{REG-TERM}]$ is upper bounded by*

$$
\begin{aligned}
&\frac{H\ln|A|}{\eta} + 2\eta H^2 \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{x_h\sim\pi^\star}\left[\sum_a \pi_t(a|x_h)\|\phi(x_h,a)\|^2_{\widehat{\Sigma}^+_{t,h}}\right]\right] \\
&\qquad + \frac{1}{H}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{x_h\sim\pi^\star}\left[\sum_a \pi_t(a|x_h) B_t(x,a)\right]\right] + \mathcal{O}\left(\eta\epsilon H^3 T + \frac{\eta H^3}{\gamma^2 T^2}\right).
\end{aligned}
$$

Proof of Lemma H.2. Again, we will apply the regret bound of the exponential weight algorithm Lemma D.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

$$
\begin{aligned}
\eta\left|\phi(x,a)^\top\widehat{\theta}_{\tau,h}\right| &= \eta\left|\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\phi(x_{t,h},a_{t,h})L_{t,h}\right| \\
&\le \eta \times \left\|\widehat{\Sigma}^+_{t,h}\right\|_{\text{op}} \times L_{t,h} \le \frac{\eta H}{\gamma} \le \frac{1}{2}, && \left(\text{Eq. (28) and the condition } \eta \le \tfrac{\gamma}{2H}\right)
\end{aligned}
$$

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

$$
\eta B_t(x,a) \le \eta \times H\left(1+\frac{1}{H}\right)^H \times 2\beta \sup_{x,a,h}\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}} \le \frac{6\eta\beta H}{\gamma} \le \frac{1}{2H}, \qquad (38)
$$

where the last inequality is by Eq. (28) again and the condition $\eta\beta \le \frac{\gamma}{12H^2}$.

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

$$
\mathbb{E}\left[\sum_{t=1}^{T}\sum_a \left(\pi_t(a|x) - \pi^\star(a|x)\right)\widehat{Q}_t(x,a)\right]
$$

$$\leq \frac{\ln|A|}{\eta} + 2\eta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\pi_t(a|x)\widehat{Q}_t(x,a)^2\right] + 2\eta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\pi_t(a|x)B_t(x,a)^2\right]. \tag{39}$$

The last term in Eq. (39) 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. (38). To bound the second term in Eq. (39), we use the following: for $(x,a) \in X_h \times A$,

$$\begin{aligned}
\mathbb{E}_t\left[\widehat{Q}_t(x,a)^2\right] &\leq H^2\mathbb{E}_t\left[\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\phi(x_{t,h},a_{t,h})\phi(x_{t,h},a_{t,h})^\top\widehat{\Sigma}^+_{t,h}\phi(x,a)\right] \\
&= H^2\mathbb{E}_t\left[\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\Sigma_{t,h}\widehat{\Sigma}^+_{t,h}\phi(x,a)\right] \\
&\leq H^2\mathbb{E}_t\left[\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\Sigma_{t,h}\left(\gamma I+\Sigma_{t,h}\right)^{-1}\phi(x,a)\right] + \mathcal{O}\left(\epsilon H^2 + \frac{H^2}{\gamma^2T^3}\right) && (*) \\
&\leq H^2\phi(x,a)^\top\left(\gamma I+\Sigma_{t,h}\right)^{-1}\Sigma_{t,h}\left(\gamma I+\Sigma_{t,h}\right)^{-1}\phi(x,a) + \mathcal{O}\left(\epsilon H^2 + \frac{H^2}{\gamma^2T^3}\right) && \text{(by Eq. (29))} \\
&\leq H^2\phi(x,a)^\top\left(\gamma I+\Sigma_{t,h}\right)^{-1}\phi(x,a) + \mathcal{O}\left(\epsilon H^2 + \frac{H^2}{\gamma^2T^3}\right) \\
&\leq H^2\mathbb{E}_t\left[\phi(x,a)^\top\widehat{\Sigma}^+_{t,h}\phi(x,a)\right] + \mathcal{O}\left(\epsilon H^2 + \frac{H^2}{\gamma^2T^3}\right) && \text{(by Eq. (29) again)} \\
&= H^2\mathbb{E}_t\left[\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}}\right] + \mathcal{O}\left(\epsilon H^2 + \frac{H^2}{\gamma^2T^3}\right)
\end{aligned}$$

where $(*)$ is because by Eq. (30) and Eq. (31), $\|(\gamma I+\Sigma_{t,h})^{-1} - \widehat{\Sigma}^+_{t,h}\|_{\text{op}} \leq 2\epsilon$ and $\|\widehat{\Sigma}^+_{t,h}\Sigma_{t,h}\|_{\text{op}} \leq 1+2\epsilon$ 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. (39) and taking expectation over states finishes the proof. $\square$

With Lemma H.1 and Lemma H.2, we can now prove Theorem B.1.

Proof of Theorem B.1. Combining Lemma H.1 and Lemma H.2, we get (under the required conditions of the parameters):

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

We see that Eq. (21) is satisfied in expectation as long as we have $2\eta H^2 + \frac{\beta}{4} \leq \beta$ 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. (20) is also satisfied with this choice of $b_t(x,a)$. Therefore, we can apply Lemma E.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

$$\begin{aligned}
&\mathbb{E}[\text{Reg}] \\
&= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\gamma dH^3T}{\beta} + \frac{\eta H^3}{\gamma^2T^2} + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{(x_h,a)\sim\pi_t}\left[b_t(x,a)\right]\right]\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\gamma dH^3T}{\beta} + \frac{\eta H^3}{\gamma^2T^2} + \beta\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{(x_h,a)\sim\pi_t}\left[\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{t,h}}\right]\right]\right)
\end{aligned}$$

$$= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\gamma d H^3 T}{\beta} + \frac{\eta H^3}{\gamma^2 T^2} + \beta \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1} \mathbb{E}_{(x_h,a)\sim\pi_t}\left[\|\phi(x,a)\|^2_{(\gamma I + \Sigma_{t,h})^{-1}}\right]\right]\right) \qquad \text{(Eq. (29) and } \beta \le H)$$
$$= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\gamma d H^3 T}{\beta} + \frac{\eta H^3}{\gamma^2 T^2} + \beta d H T\right),$$

where the last step uses the fact

$$\mathbb{E}_t\left[\sum_h \mathbb{E}_{(x_h,a)\sim\pi_t}\left[\|\phi(x,a)\|^2_{(\gamma I + \Sigma_{t,h})^{-1}}\right]\right] \le \mathbb{E}_t\left[\sum_h \mathbb{E}_{(x_h,a)\sim\pi_t}\left[\|\phi(x,a)\|^2_{\Sigma_{t,h}^{-1}}\right]\right]$$
$$= \sum_h \left\langle \Sigma_{t,h}, \Sigma_{t,h}^{-1}\right\rangle = dH. \tag{40}$$

Finally, choosing the parameters under the specified constraints as:

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

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

I. Details Omitted in Section C

In this section, we analyze our algorithm for linear MDPs. First, we show the main benefit of exploring with the policy cover, that is, it ensures a small magnitude for $b_t(x,a)$, as shown below.

Lemma I.1. *If $\gamma \ge \frac{36\beta^2}{\delta_e}$ and $\beta\epsilon \le \frac{1}{8}$, then $b_k(x,a) \le 1$ for all (x,a) and all k (with high probability).*

Proof. According to the definition of $b_k(x,a)$ (in Algorithm 5), it suffices to show that for $x \in \mathcal{K}$, $\beta\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}} \le \frac{1}{2}$ for any a. To do so, note that the GEOMETRICRESAMPLING procedure ensures that $\widehat{\Sigma}^+_{k,h}$ is an estimation of the inverse of $\gamma I + \Sigma^{\text{mix}}_{k,h}$, where

$$\Sigma^{\text{mix}}_{k,h} = \delta_e \Sigma^{\text{cov}}_h + (1-\delta_e)\mathbb{E}_{(x_h,a)\sim\pi_k}[\phi(x_h,a)\phi(x_h,a)^\top] \tag{41}$$

and $\Sigma^{\text{cov}}_h = \frac{1}{M_0}\sum_{m=1}^{M_0} \mathbb{E}_{(x_h,a)\sim\pi_m}\left[\phi(x_h,a)\phi(x_h,a)^\top\right]$ is the covariance matrix of the policy cover π_{cov}. By Eq. (30), we have with probability at least $1 - 1/T^3$,

$$\beta\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}} \le \beta\|\phi(x,a)\|^2_{(\gamma I + \Sigma^{\text{mix}}_{k,h})^{-1}} + 2\beta\epsilon \le \beta\|\phi(x,a)\|^2_{(\gamma I + \Sigma^{\text{mix}}_{k,h})^{-1}} + \frac{1}{4}.$$

The first term can be further bounded as $\frac{\beta}{\delta_e}\|\phi(x,a)\|^2_{(\frac{\gamma}{\delta_e} I + \Sigma^{\text{cov}}_h)^{-1}} \le \frac{\beta}{\delta_e}\|\phi(x,a)\|^2_{(\frac{1}{M_0} I + \Sigma^{\text{cov}}_h)^{-1}}$, where the last step is because $\frac{\gamma}{\delta_e} M_0 \ge \frac{\gamma}{\delta_e} \times \frac{\delta_e^2}{36\beta^2} \ge 1$ by our condition. Finally, we show that $\frac{1}{M_0} I + \Sigma^{\text{cov}}_h$ and $\widehat{\Sigma}^{\text{cov}}_h$ are close. Recall the definition of the latter:

$$\widehat{\Sigma}^{\text{cov}}_h = \frac{1}{M_0} I + \frac{1}{M_0 N_0}\sum_{m=1}^{M_0}\sum_{t=(m-1)N_0+1}^{mN_0} \phi(x_{t,h}, a_{t,h})\phi(x_{t,h}, a_{t,h})^\top.$$

We now apply Lemma D.3 with $n = N_0$ and

$$X_k = \frac{1}{M_0}\sum_{m=1}^{M_0} \phi(x_{\tau(m,k),h}, a_{\tau(m,k),h})\phi(x_{\tau(m,k),h}, a_{\tau(m,k),h})^\top - \frac{1}{M_0}\sum_{m=1}^{M_0} \mathbb{E}_{(x_h,a)\sim\pi_m}\left[\phi(x_h,a)\phi(x_h,a)^\top\right],$$

for $k = 1, \dots, N_0$, where $\tau(m,k) \triangleq (m-1)N_0 + k$. Note that $X_k^2 \preceq I$. Therefore, we can pick $A_k = I$ and $\sigma = 1$. By Lemma D.3, we have with probability at least $1-\delta$,

$$\left\| \widehat{\Sigma}_h^{\text{cov}} - \frac{1}{M_0} I - \Sigma_h^{\text{cov}} \right\|_{\text{op}} \leq \sqrt{\frac{8\log(d/\delta)}{N_0}}.$$

Following the same proof as (Meng & Zheng, 2010, Theorem 2.1), we have

$$\left\| \left(\frac{1}{M_0} I + \Sigma_h^{\text{cov}} \right)^{-1} - \left(\widehat{\Sigma}_h^{\text{cov}} \right)^{-1} \right\|_{\text{op}} \leq M_0^2 \left\| \widehat{\Sigma}_h^{\text{cov}} - \frac{1}{M_0} I - \Sigma_h^{\text{cov}} \right\|_{\text{op}} \leq M_0^2 \sqrt{\frac{8\log(d/\delta)}{N_0}} \leq \frac{\alpha}{2}.$$

(by our choice of N_0 and M_0)

Consequently, for any vector ϕ with $\|\phi\| \leq 1$, we have

$$\left| \|\phi\|^2_{\left(\frac{1}{M_0} I + \Sigma_h^{\text{cov}}\right)^{-1}} - \|\phi\|^2_{\left(\widehat{\Sigma}_h^{\text{cov}}\right)^{-1}} \right| \leq \left\| \left(\frac{1}{M_0} I + \Sigma_h^{\text{cov}} \right)^{-1} - \left(\widehat{\Sigma}_h^{\text{cov}} \right)^{-1} \right\|_{\text{op}} \leq \frac{\alpha}{2}.$$

Therefore, combining everything we have

$$\beta \|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}} \leq \frac{\beta}{\delta_e} \left(\|\phi(x,a)\|^2_{\left(\widehat{\Sigma}_h^{\text{cov}}\right)^{-1}} + \frac{\alpha}{2} \right) + \frac{1}{4} \leq \frac{3\beta\alpha}{2\delta_e} + \frac{1}{4} \leq \frac{1}{2},$$

where the last two steps use the fact $x \in \mathcal{K}$ and the value of α. This finishes the proof. □

Next, we define the following notations for convenience due to the epoch schedule of our algorithm, and then proceed to prove the main theorem.

Definition 2.

$$\begin{aligned}
\overline{\ell}_k(x,a) &= \frac{1}{W} \sum_{t=T_0+(k-1)W+1}^{T_0+kW} \ell_t(x,a) \\
\overline{Q}_k^{\pi}(x,a) &= Q^{\pi}(x,a;\overline{\ell}_k) \\
\overline{\theta}_{k,h}^{\pi} &\ \textit{is such that}\ \overline{Q}_k^{\pi}(x,a) = \phi(x,a)^\top \overline{\theta}_{k,h}^{\pi} \\
B_k(x,a) &= b_k(x,a) + \left(1 + \frac{1}{H}\right) \mathbb{E}_{x' \sim P(\cdot|x,a)} \mathbb{E}_{a' \sim \pi_k(\cdot|x')} [B_k(x',a')] \\
\widehat{B}_k(x,a) &= b_k(x,a) + \phi(x,a)^\top \widehat{\Lambda}_{k,h} \qquad \text{(for } x \in X_h)
\end{aligned}$$

Proof of Theorem C.1. We first analyze the regret of policy optimization after the first T_0 rounds. Our goal is again to prove Eq. (21) which in this case bounds

$$\sum_{k=1}^{(T-T_0)/W} \sum_h \mathbb{E}_{X_h \ni x \sim \pi^\star} \left[\sum_a \left(\pi_k(a|x) - \pi^\star(a|x) \right) \left(\overline{Q}_k^{\pi_k}(x,a) - B_k(x,a) \right) \right].$$

The first step is to separate known states and unknown states. For unknown states, we have

$$\begin{aligned}
&\sum_{k=1}^{(T-T_0)/W} \sum_h \mathbb{E}_{X_h \ni x \sim \pi^\star} \left[\mathbf{1}[x \notin \mathcal{K}] \sum_a \left(\pi_k(a|x) - \pi^\star(a|x) \right) \left(\overline{Q}_k^{\pi_k}(x,a) - B_k(x,a) \right) \right] \\
&\leq \frac{(T-T_0)He}{W} \sum_h \mathbb{E}_{X_h \ni x \sim \pi^\star} \left[\mathbf{1}[x \notin \mathcal{K}] \right] = \widetilde{\mathcal{O}}\left(\frac{dH^3 T}{\alpha W} \right),
\end{aligned}$$

where the first step is by the facts $0 \leq \overline{Q}_k^{\pi_k}(x,a) \leq H$ and $0 \leq B_k(x,a) \leq (1+\frac{1}{H})^H \times H \leq He$ (Lemma I.1), and the second step applies Lemma G.4. For known states, we apply a similar decomposition as previous analysis, but since we also use function approximation for bonus $B_t(x,a)$, we need to account for its estimation error, which results in two extra bias terms:

$$
\begin{aligned}
&\sum_{k=1}^{(T-T_0)/W}\sum_h \mathbb{E}_{X_h\ni x\sim\pi^\star}\left[\mathbf{1}[x\in\mathcal{K}]\sum_a\Big(\pi_k(a|x)-\pi^\star(a|x)\Big)\Big(\overline{Q}_k^{\pi_k}(x,a)-B_k(x,a)\Big)\right]\\
&=\underbrace{\sum_{k=1}^{(T-T_0)/W}\sum_h \mathbb{E}_{X_h\ni x\sim\pi^\star}\left[\mathbf{1}[x\in\mathcal{K}]\sum_a\pi_k(a|x)\Big(\overline{Q}_k^{\pi_k}(x,a)-\widehat{Q}_k(x,a)\Big)\right]}_{\textsc{Bias-1}}\\
&\quad+\underbrace{\sum_{k=1}^{(T-T_0)/W}\sum_h \mathbb{E}_{X_h\ni x\sim\pi^\star}\left[\mathbf{1}[x\in\mathcal{K}]\sum_a\pi^\star(a|x)\Big(\widehat{Q}_k(x,a)-\overline{Q}_k^{\pi_k}(x,a)\Big)\right]}_{\textsc{Bias-2}}\\
&\quad+\underbrace{\sum_{k=1}^{(T-T_0)/W}\sum_h \mathbb{E}_{X_h\ni x\sim\pi^\star}\left[\mathbf{1}[x\in\mathcal{K}]\sum_a\pi_k(a|x)\Big(\widehat{B}_k(x,\cdot)-B_k(x,\cdot)\Big)\right]}_{\textsc{Bias-3}}\\
&\quad+\underbrace{\sum_{k=1}^{(T-T_0)/W}\sum_h \mathbb{E}_{X_h\ni x\sim\pi^\star}\left[\mathbf{1}[x\in\mathcal{K}]\sum_a\pi^\star(a|x)\Big(B_k(x,a)-\widehat{B}_k(x,a)\Big)\right]}_{\textsc{Bias-4}}\\
&\quad+\underbrace{\sum_{k=1}^{(T-T_0)/W}\sum_h \mathbb{E}_{X_h\ni x\sim\pi^\star}\left[\mathbf{1}[x\in\mathcal{K}]\sum_a\Big(\pi_k(\cdot|x)-\pi^\star(\cdot|x)\Big)\Big(\widehat{Q}_k(x,\cdot)-\widehat{B}_k(x,\cdot)\Big)\right]}_{\textsc{Reg-Term}}.
\end{aligned}
$$

Now we combine the bounds in Lemma I.3, Lemma I.3, and Lemma I.4 (included after this proof). Suppose that the conditions on the parameters specified in Lemma I.4 hold. We get

$$
\begin{aligned}
&\mathbb{E}[\textsc{Bias-1}+\textsc{Bias-2}+\textsc{Bias-3}+\textsc{Bias-4}+\textsc{Reg-Term}]\\
&=\widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta\epsilon H^4T}{W}+\frac{\eta H^4}{\gamma^2T^2W}+\frac{\gamma dH^3T}{\beta W}+\frac{\epsilon H^3T}{W}\right)\\
&\qquad+\left(\frac{\beta}{2}+2\eta H^3\right)\sum_k\sum_h\mathbb{E}_{X_h\ni x\sim\pi^\star}\left[\mathbf{1}[x\in\mathcal{K}]\sum_a(\pi^\star(a|x)+\pi_k(a|x))\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right]\\
&\qquad+\frac{1}{H}\sum_k\sum_h\mathbb{E}_{X_h\ni x\sim\pi^\star}\left[\sum_a\pi_k(a|x)B_k(x,a)\right]\\
&\le\widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta\epsilon H^4T}{W}+\frac{\eta H^4}{\gamma^2T^2W}+\frac{\gamma dH^3T}{\beta W}+\frac{\epsilon H^3T}{W}\right)\\
&\qquad+\sum_k V^{\pi^\star}(x_0;b_k)+\frac{1}{H}\sum_k\sum_h\mathbb{E}_{X_h\ni x\sim\pi^\star}\left[\sum_a\pi_k(a|x)B_k(x,a)\right]
\end{aligned}
$$

where the last inequality is because $\frac{\beta}{2}+2\eta H^3\le\beta$ (implied by $\frac{\eta}{\beta}\le\frac{1}{20H^4}$, a condition specified in Lemma I.4).

Combining two cases and applying Lemma E.2, we thus have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{(T-T_0)/W}V^{\pi_k}(x_0;\overline{\ell}_k)\right]-\sum_{k=1}^{(T-T_0)/W}V^{\pi^\star}(x_0;\overline{\ell}_k)\\
&\le\widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta\epsilon H^4T}{W}+\frac{\eta H^4}{\gamma^2T^2W}+\frac{\gamma dH^3T}{\beta W}+\frac{\epsilon H^3T}{W}+\beta\mathbb{E}\left[\sum_{k=1}^{(T-T_0)/W}\sum_h\mathbb{E}_{(x_h,a)\sim\pi_k}\left[\|\phi(x_h,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right]\right]+\frac{dH^3T}{\alpha W}\right)
\end{aligned}
$$

$$= \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\eta\epsilon H^4 T}{W} + \frac{\eta H^4}{\gamma^2 T^2 W} + \frac{\gamma d H^3 T}{\beta W} + \frac{\epsilon H^3 T}{W} + \frac{\beta d H T}{W} + \frac{d H^3 T}{\alpha W}\right). \qquad \text{(by similar calculation as Eq. (40))}$$

Finally, to get the overall regret, it remains to multiply the bound above by W, add the trivial bound $HT_0 = 2HM_0N_0 = \mathcal{O}\left(\frac{\delta_e^8 d^{10} H^{11}}{\beta^8}\right)$ for the initial T_0 rounds, and consider the exploration probability δ_e, which leads to

$$\begin{aligned}
\mathbb{E}[\text{Reg}] &= \widetilde{\mathcal{O}}\left(\frac{HW}{\eta} + \eta\epsilon H^4 T + \frac{\eta H^4}{\gamma^2 T^2} + \frac{\gamma d H^3 T}{\beta} + \epsilon H^3 T + \beta d H T + \frac{d H^3 T}{\alpha} + \frac{d^{10} H^{11}\delta_e^8}{\beta^8} + \delta_e H T\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{H}{\eta\epsilon^2\gamma^3} + \frac{\eta H^4}{\gamma^2 T^2} + \frac{\gamma d H^3 T}{\beta} + \epsilon H^3 T + \beta d H T + \frac{d H^3 T \beta}{\delta_e} + \frac{d^{10} H^{11}\delta_e^8}{\beta^8} + \delta_e H T\right)
\end{aligned}$$

where we use the specified value of M, N, $W = 2MN$, α, and that $\eta H \le 1$ (so that the second term $\eta\epsilon H^4 T$ is absorbed by the fifth term $\epsilon H^3 T$).

Considering the constraints in Lemma I.4 and Lemma I.1, we choose $\gamma = \max\left\{16\eta H^4, \frac{4\beta^2}{\delta_e}\right\}$. This gives the following simplified regret

$$\widetilde{\mathcal{O}}\left(\frac{1}{\epsilon^2\eta^4 H^{11}} + \frac{1}{\eta H^4 T^2} + \frac{\eta d H^7 T}{\beta} + \epsilon H^3 T + \beta d H T + \frac{d H^3 T \beta}{\delta_e} + \frac{d^{10} H^{11}\delta_e^8}{\beta^8} + \delta_e H T\right).$$

Choosing δ_e optimally, and supposing $\eta \ge \frac{1}{T}$, the above is simplified to

$$\begin{aligned}
&\widetilde{\mathcal{O}}\left(\frac{1}{\epsilon^2\eta^4 H^{11}} + \frac{\eta d H^7 T}{\beta} + \epsilon H^3 T + \beta d H T + H^2\sqrt{d\beta}T + d^2 H^{35/9} T^{8/9}\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{1}{\epsilon^2\eta^4 H^{11}} + \frac{\eta d H^7 T}{\beta} + \epsilon H^3 T + H^2\sqrt{d\beta}T + d^2 H^{35/9} T^{8/9}\right). \qquad \text{(choosing } \beta \le \tfrac{H^2}{d}\text{)}
\end{aligned}$$

Picking optimal parameters in the last expression, we get $\widetilde{\mathcal{O}}\left(d^2 H^4 T^{14/15}\right)$. □

Lemma I.2.

$$\begin{aligned}
&\mathbb{E}[\textsc{Bias-1} + \textsc{Bias-2}] \\
&\le \frac{\beta}{4}\mathbb{E}\left[\sum_{k=1}^{(T-T_0)/W}\sum_h \mathbb{E}_{X_h \ni x \sim \pi^\star}\left[\mathbf{1}[x \in \mathcal{K}]\sum_a \left(\pi^\star(a|x) + \pi_k(a|x)\right)\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right]\right] + \mathcal{O}\left(\frac{\gamma d H^3 T}{\beta W} + \frac{\epsilon H^3 T}{W}\right).
\end{aligned}$$

Proof. The proof of this lemma is similar to that of Lemma H.1, except that we replace T by $(T - T_0)/W$, and consider the averaged loss $\overline{\ell}_k$ in an epoch instead of the single episode loss ℓ_t:

$$\begin{aligned}
&\mathbb{E}_k\left[\overline{Q}_k^{\pi_k}(x,a) - \widehat{Q}_k(x,a)\right] \\
&= \phi(x,a)^\top\left(\overline{\theta}^{\pi_k}_{k,h} - \mathbb{E}_k\left[\widehat{\theta}_{k,h}\right]\right) \\
&= \phi(x,a)^\top\left(\overline{\theta}^{\pi_k}_{k,h} - \mathbb{E}_k\left[\widehat{\Sigma}^+_{k,h}\right]\mathbb{E}_k\left[\frac{1}{|S_k'|}\sum_{t\in S_k'}((1-Y_t) + Y_t H\mathbb{1}[h = h_t^*])\phi(x_{t,h}, a_{t,h})L_{t,h}\right]\right) \\
&\hspace{20em} (S_k' \text{ is the } S' \text{ in Algorithm 5 within epoch } k) \\
&= \phi(x,a)^\top\left(\overline{\theta}^{\pi_k}_{k,h} - \left(\gamma I + \Sigma^{\text{mix}}_{k,h}\right)^{-1}\mathbb{E}_k\left[\frac{1}{|S_k'|}\sum_{t\in S_k'}((1-Y_t) + Y_t H\mathbb{1}[h = h_t^*])\phi(x_{t,h}, a_{t,h})L_{t,h}\right]\right) + \mathcal{O}(\epsilon H^2) \\
&\hspace{6em} \text{(by Lemma G.1 and that } \|\phi(x,a)\| \le 1 \text{ for all } x, a \text{ and } L_{t,h} \le H;\ \Sigma^{\text{mix}}_{k,h} \text{ is defined in Eq. (41))} \\
&= \phi(x,a)^\top\left(\overline{\theta}^{\pi_k}_{k,h} - \left(\gamma I + \Sigma^{\text{mix}}_{k,h}\right)^{-1}\mathbb{E}_k\left[\frac{1}{|S_k'|}\sum_{t\in S_k'}\Sigma^{\text{mix}}_{k,h}\theta^{\pi_k}_{t,h}\right]\right) + \mathcal{O}(\epsilon H^2)
\end{aligned}$$

$$= \phi(x,a)^\top \left(\overline{\theta}_{k,h}^{\pi_k} - \left(\gamma I + \Sigma_{k,h}^{\text{mix}}\right)^{-1} \mathbb{E}_k \left[\frac{1}{W} \sum_{t=(k-1)W+1}^{kW} \Sigma_{k,h}^{\text{mix}} \theta_{t,h}^{\pi_k} \right] \right) + \mathcal{O}(\epsilon H^2)$$

(S'_k is randomly chosen from epoch k)

$$= \phi(x,a)^\top \left(\overline{\theta}_{k,h}^{\pi_k} - \left(\gamma I + \Sigma_{k,h}^{\text{mix}}\right)^{-1} \Sigma_{k,h}^{\text{mix}} \overline{\theta}_{k,h}^{\pi_k} \right) + \mathcal{O}(\epsilon H^2)$$
$$= \gamma \phi(x,a)^\top \left(\gamma I + \Sigma_{k,h}^{\text{mix}}\right)^{-1} \overline{\theta}_{k,h}^{\pi_k} + \mathcal{O}\left(\epsilon H^2\right)$$
$$\leq \frac{\beta}{4} \|\phi(x,a)\|^2_{(\gamma I + \Sigma_{t,h}^{\text{mix}})^{-1}} + \frac{\gamma^2}{\beta} \left\| \overline{\theta}_{k,h}^{\pi_k} \right\|^2_{(\gamma I + \Sigma_{k,h}^{\text{mix}})^{-1}} + \mathcal{O}(\epsilon H^2) \qquad \text{(AM-GM inequality)}$$
$$\leq \frac{\beta}{4} \mathbb{E}_k \left[\|\phi(x,a)\|^2_{\widehat{\Sigma}_{k,h}^+} \right] + \frac{\gamma d H^2}{\beta} + \mathcal{O}\left(\epsilon H^2\right).$$

The same bound also holds for $\mathbb{E}_k \left[\widehat{Q}_k(x,a) - \overline{Q}_k^{\pi_k}(x,a) \right]$ by the same reasoning. Taking expectation over x, summing over k, h and a (with weights $\pi_k(a|x)$ and $\pi^\star(a|x)$ respectively) finishes the proof. $\square$

Lemma I.3.

$$\mathbb{E}[\text{BIAS-3} + \text{BIAS-4}]$$
$$\leq \frac{\beta}{4} \mathbb{E} \left[\sum_{k=1}^{(T-T_0)/W} \sum_h \mathbb{E}_{X_h \ni x \sim \pi^\star} \left[\mathbf{1}[x \in \mathcal{K}] \sum_a \left(\pi^\star(a|x) + \pi_k(a|x) \right) \|\phi(x,a)\|^2_{\widehat{\Sigma}_{k,h}^+} \right] \right] + \mathcal{O}\left(\frac{\gamma d H^3 T}{\beta W} + \frac{\epsilon H^3 T}{W} \right).$$

Proof. The proof is almost identical to that of the previous lemma. The only difference is that $L_{t,h}$ is replaced by $D_{t,h}$ and $\overline{\theta}_{t,h}^{\pi_k}$ is replaced by $\Lambda_{k,h}^{\pi_k}$ (recall the definition of $\Lambda_{k,h}^{\pi_k}$ in Section C). Note that $b_t(x,a) \in [0,1]$ (Lemma I.1), so $D_{t,h} \in [0, He]$, which is also the same order for $L_{t,h}$. Therefore, we get the same bound as in the previous lemma. $\square$

Lemma I.4. *Let $\frac{\eta}{\gamma} \leq \frac{1}{16H^4}$ and $\frac{\eta}{\beta} \leq \frac{1}{40H^4}$. Then*

$$\mathbb{E}[\text{REG-TERM}] = \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\eta \epsilon H^4 T}{W} + \frac{\eta H^4}{\gamma^2 T^2 W} \right)$$
$$+ 2\eta H^3 \mathbb{E} \left[\sum_{k,h} \mathbb{E}_{X_h \ni x \sim \pi^\star} \left[\mathbf{1}[x \in \mathcal{K}] \sum_a \pi_k(x,a) \|\phi(x,a)\|^2_{\widehat{\Sigma}_{k,h}^+} \right] \right] + \frac{1}{H} \mathbb{E} \left[\sum_{k,h} \mathbb{E}_{X_h \ni x \sim \pi^\star} \left[\sum_a \pi_k(x,a) B_k(x,a) \right] \right].$$

Proof. We first check the condition for Lemma D.4: $\eta \left| \widehat{Q}_k(x,a) - \widehat{B}_t(x,a) \right| \leq 1$. In our case,

$$\eta \left| \widehat{Q}_k(x,a) \right| = \eta \left| \phi(x,a)^\top \widehat{\Sigma}_{k,h}^+ \left(\frac{1}{|S'|} \sum_{t \in S'} ((1-Y_t) + Y_t H \mathbb{1}[h = h_t^*]) \phi(x_{t,h}, a_{t,h}) L_{t,h} \right) \right|$$
$$\leq \eta \times \|\widehat{\Sigma}_{k,h}^+\|_{\text{op}} \times H \times \sup_{t \in S'} L_{t,h}$$
$$\leq \eta \times \frac{1}{\gamma} \times H^2 \qquad \text{(by Lemma G.1)}$$
$$\leq \frac{1}{2} \qquad \text{(by the condition specified in the lemma)}$$

and

$$\eta \left| \widehat{B}_k(x,a) \right| \leq \eta \left| b_k(x,a) \right| + \eta \left| \phi(x,a)^\top \widehat{\Sigma}_{k,h}^+ \left(\frac{1}{|S'|} \sum_{t \in S'} ((1-Y_t) + Y_t H \mathbb{1}[h = h_t^*]) \phi(x_{t,h}, a_{t,h}) D_{t,h} \right) \right|$$
$$\leq \eta + \eta \times \|\widehat{\Sigma}_{k,h}^+\|_{\text{op}} \times H \times \sup_{t \in S'} D_{t,h}$$

$$
\begin{aligned}
&\leq \eta + \eta \times \|\widehat{\Sigma}^+_{k,h}\|_{\mathrm{op}} \times H \times (H-1)\left(1+\frac{1}{H}\right)^H && \text{(Lemma I.1)}\\
&\leq \eta + \frac{3\eta H^2}{\gamma} \leq \frac{4\eta H^2}{\gamma}\\
&\leq \frac{1}{2H}. && \text{(by the condition specified in the lemma)}
\end{aligned}
$$

Now we derive an upper bound for $\mathbb{E}_k\left[\widehat{Q}_k(x,a)^2\right]$:

$$
\begin{aligned}
&\mathbb{E}_k\left[\widehat{Q}_k(x,a)^2\right]\\
&\leq \mathbb{E}_k\left[\frac{1}{|S'_k|}\sum_{t\in S'_k} H^2\phi(x,a)^\top\widehat{\Sigma}^+_{k,h}\Big(((1-Y_t)+Y_tH\mathbb{1}[h=h_t^*])^2\phi(x_{t,h},a_{t,h})\phi(x_{t,h},a_{t,h})^\top\Big)\widehat{\Sigma}^+_{k,h}\phi(x,a)\right] && (*)\\
&= \mathbb{E}_k\left[H^2\phi(x,a)^\top\widehat{\Sigma}^+_{k,h}\Big((1-\delta_e)\Sigma_{k,h}+\delta_e H\Sigma_h^{\mathrm{cov}}\Big)\widehat{\Sigma}^+_{k,h}\phi(x,a)\right]\\
&\leq H^3\mathbb{E}_k\left[\phi(x,a)^\top\widehat{\Sigma}^+_{k,h}\Sigma^{\mathrm{mix}}_{k,h}\widehat{\Sigma}^+_{k,h}\phi(x,a)\right]\\
&\leq H^3\mathbb{E}_k\left[\phi(x,a)^\top\widehat{\Sigma}^+_{k,h}\Sigma^{\mathrm{mix}}_{k,h}(\gamma I+\Sigma^{\mathrm{mix}}_{k,h})^{-1}\phi(x,a)\right] + \widetilde{\mathcal{O}}\left(\epsilon H^3+\frac{H^3}{\gamma^2T^3}\right) && \text{(Lemma G.1)}\\
&\leq H^3\phi(x,a)^\top(\gamma I+\Sigma^{\mathrm{mix}}_{k,h})^{-1}\Sigma^{\mathrm{mix}}_{k,h}(\gamma I+\Sigma^{\mathrm{mix}}_{k,h})^{-1}\phi(x,a) + \widetilde{\mathcal{O}}\left(\epsilon H^3+\frac{H^3}{\gamma^2T^3}\right) && \text{(Lemma G.1)}\\
&\leq H^3\phi(x,a)^\top(\gamma I+\Sigma^{\mathrm{mix}}_{k,h})^{-1}\phi(x,a) + \widetilde{\mathcal{O}}\left(\epsilon H^3+\frac{H^3}{\gamma^2T^3}\right)\\
&= H^3\mathbb{E}_k\left[\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right] + \widetilde{\mathcal{O}}\left(\epsilon H^3+\frac{H^3}{\gamma^2T^3}\right), && (42)
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathbb{E}_k\left[\widehat{B}_k(x,a)^2\right]\\
&\leq 2\mathbb{E}_k\left[b_k(x,a)^2\right] + 2\mathbb{E}_k\left[(\phi(x,a)^\top\widehat{\Lambda}_{k,h})^2\right]\\
&\leq 2\mathbb{E}_k[b_k(x,a)] + 18H^3\mathbb{E}_k\left[\|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right] + \widetilde{\mathcal{O}}\left(\epsilon H^3+\frac{H^3}{\gamma^2T^3}\right)\\
&\leq \frac{20H^3}{\beta}b_k(x,a) + \widetilde{\mathcal{O}}\left(\epsilon H^3+\frac{H^3}{\gamma^2T^3}\right),
\end{aligned}
$$

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. (42), except that we replace the upper bound H for $L_{t,h}$ by the upper bound for $D_{t,h}$: $H\left(1+\frac{1}{H}\right)^H\sup_{t,x,a} b_t(x,a) \leq 3H$ (since $b_t(x,a)\leq 1$ by Lemma I.1).

Thus, by Lemma D.4, we have

$$
\begin{aligned}
&\mathbb{E}[\textsc{Reg-Term}]\\
&\leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta}\right) + 2\eta\sum_{k,h}\mathbb{E}_{X_h\ni x\sim\pi^\star}\left[\mathbf{1}[x\in\mathcal{K}]\sum_a \pi_k(a|x)(\widehat{Q}_k(x,a)^2+\widehat{B}_k(x,a)^2)\right]\\
&\leq \widetilde{\mathcal{O}}\left(\frac{H}{\eta}+\frac{\eta\epsilon H^4T}{W}+\frac{\eta H^4}{\gamma^2T^2W}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\eta H^3 \mathbb{E}\left[\sum_{k,h} \mathbb{E}_{X_h \ni x \sim \pi^\star}\left[\mathbf{1}[x \in \mathcal{K}] \sum_a \pi_k(x,a) \|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right]\right] + \frac{40\eta H^3}{\beta} \mathbb{E}\left[\sum_{k,h} \mathbb{E}_{X_h \ni x \sim \pi^\star}\left[\sum_a \pi_k(a|x) b_k(x,a)\right]\right] \\
&\le \widetilde{\mathcal{O}}\left(\frac{H}{\eta} + \frac{\eta \epsilon H^4 T}{W} + \frac{\eta H^4}{\gamma^2 T^2 W}\right) \\
&+ 2\eta H^3 \mathbb{E}\left[\sum_{k,h} \mathbb{E}_{X_h \ni x \sim \pi^\star}\left[\mathbf{1}[x \in \mathcal{K}] \sum_a \pi_k(x,a) \|\phi(x,a)\|^2_{\widehat{\Sigma}^+_{k,h}}\right]\right] + \frac{1}{H} \mathbb{E}\left[\sum_{k,h} \mathbb{E}_{X_h \ni x \sim \pi^\star}\left[\sum_a \pi_k(a|x) \textsc{Bonus}_k(x,a)\right]\right]
\end{aligned}
$$

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