Improved Regret for Efficient Online Reinforcement Learning with Linear Function Approximation

Uri Sherman [1] Tomer Koren [1 2] Yishay Mansour [1 2]

Abstract

We study reinforcement learning with linear function approximation and adversarially changing cost functions, a setup that has mostly been considered under simplifying assumptions such as full information feedback or exploratory conditions. We present a computationally efficient policy optimization algorithm for the challenging general setting of unknown dynamics and bandit feedback, featuring a combination of mirror-descent and least squares policy evaluation in an auxiliary MDP used to compute exploration bonuses. Our algorithm obtains an $\widetilde{O}(K^{6/7})$ regret bound, improving significantly over previous state-of-the-art of $\widetilde{O}(K^{14/15})$ in this setting. In addition, we present a version of the same algorithm under the assumption a simulator of the environment is available to the learner (but otherwise no exploratory assumptions are made), and prove it obtains state-of-the-art regret of $\widetilde{O}(K^{2/3})$.

1 Introduction

Reinforcement Learning (RL; Sutton & Barto, 2018; Mannor et al., 2022) studies online decision making problems in which an agent learns through experience within a dynamic environment, with the goal to minimize a loss function associated with the agent-environment interaction. Modern applications of RL such as robotics (Schulman et al., 2015; Lillicrap et al., 2015; Akkaya et al., 2019), game playing (Mnih et al., 2013; Silver et al., 2018) and autonomous driving (Kiran et al., 2021), almost invariably consist of large scale environments where function approximation techniques are necessary to allow the agent to generalize across different states. Furthermore, some form of agent robustness is usually required to cope with environment irregularities that cannot be faithfully represented by stochasticity assumptions (see e.g., Dulac-Arnold et al., 2021).

Theoretical foundations for RL with function approximation (e.g., Jiang et al., 2017; Yang & Wang, 2019; Jin et al., 2020b; Agarwal et al., 2020) have been steadily coming into fruition. The influential work of Jin et al. (2020b) has set the ground for the de facto standard of linearly realizable RL; the linear Markov Decision Process (linear MDP), and has lead to a range of algorithmic approaches in this setting or variants thereof (e.g., Zanette et al., 2020a; Agarwal et al., 2020; Wagenmaker et al., 2022b, see also Agarwal et al., 2019). Likewise, a growing line of work studies RL with adversarial interventions, such as non-stationary dynamics (Mao et al., 2021), adversarial corruptions (Lykouris et al., 2021), delayed feedback (Lancewicki et al., 2022; Jin et al., 2022), and adversarial costs (Even-Dar et al., 2009; Neu et al., 2012; Rosenberg & Mansour, 2019; 2020; Jin et al., 2020a). The latter is, arguably, the more fundamental and well studied setting in the scope of adversarial RL.

The present paper aims at advancing state-of-the-art algorithmic methods for computationally and statistically efficient RL in the linear MDP setup, under the challenging setting of *adversarially changing costs, unknown dynamics, and bandit feedback*. At this time, there exist only a handful of papers that consider RL in a setup that combines function approximation and adversarial costs, with most prior works adopting one or more assumptions that alleviate the challenge of exploration. Cai et al. (2020) was the first work to establish $\widetilde{O}(\sqrt{K})$ regret over K episodes in the related model of linear mixture MDP, yet considered full information feedback. Later, Neu & Olkhovskaya (2021) obtain the same minimax optimal rates in terms of K for linear MDPs and bandit feedback, but with full knowledge of the environment dynamics, and an additional factor depending on the coverage of the initial state-action distribution. Finally, the recent work of Luo et al. (2021) establishes an $\widetilde{O}(K^{14/15})$ guarantee in the linear MDP setup without any simplifying assumptions, and an $\widetilde{O}(K^{2/3})$ regret bound in the more general linear-Q setting but with simulator access (albeit with a computationally inefficient algorithm). Notably, to the best of our knowledge, Luo et al. (2021) is the only prior work to consider the adversarial linear MDP with bandit feedback

[1]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel [2]Google Research, Tel Aviv, Israel. Correspondence to: Uri Sherman <urisherman@mail.tau.ac.il>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

in its full generality.

Contributions. Our main contribution significantly improves over the existing prior art (Luo et al., 2021) in a number of respects. We present a computationally efficient algorithm for the most general setup without any exploratory assumptions, and prove a regret bound of $\widetilde{O}(K^{6/7})$ establishing a substantial advancement with respect to the previous $\widetilde{O}(K^{14/15})$. In addition, we present a version of the same algorithm under the assumption a simulator is available to the learner, and prove it obtains an $\widetilde{O}(K^{2/3})$ bound matching the state-of-the-art in this setup given by the linear-Q algorithm of Luo et al. (2021) (which, notably, also applies in a more general setup). However, our algorithm improves upon that of Luo et al. (2021) in being computationally efficient, [1] and in requiring a weaker simulator, which we use only to generate agent policy rollouts from the initial state. Also noteworthy in this context is the algorithm of Neu & Olkhovskaya (2021), which obtains an $\widetilde{O}(\sqrt{K})$ regret bound, though requires not only a simulator but also perfect knowledge of the transition function.

Overview of techniques. Our work combines elements from Jin et al. (2020b); Shani et al. (2020); Neu & Olkhovskaya (2021); Luo et al. (2021) with a novel algorithmic approach towards exploration bonuses in linear MDPs. We follow the insightful work of Luo et al. (2021) and consider a regret decomposition and bonus design that at a high level are similar to those presented in their work, but reframed and extended to incorporate optimistic approximations of the *bonus-to-go*; the bonus function that drives exploration. Our central observation is that the bonus-to-go may be optimistically approximated using least squares regression in the auxiliary full information *bonus MDP*, in a manner that is efficient, and to an extent decoupled from estimation of the cost function. The (non-linear) reward function in this MDP is the immediate bonus function that compensates for uncertainty in the instantaneous Q-estimates; importantly, while this is not a linear MDP, it is still amenable to least squares value backups (e.g., Jin et al., 2020b) owed to the linear structure in the dynamics.

During value backups in the bonus MDP, we incorporate an additional bonus in order to maintain (w.h.p.) Bellman consistency errors that are positive across the entire state action space. This is a form of optimism employed in policy optimization algorithms (e.g., Cai et al., 2020; Shani et al., 2020), where the long term reward of the policy in each value backup step is overestimated (as opposed to optimizing a value function that is an overestimate of the reward of a benchmark policy). Unlike previous approaches that apply this directly towards the loss (or reward) optimization, here we utilize it solely for bonus calculation. Finally, through a refined analysis, we simplify the framework of Luo et al. (2021), remove the necessity of the dilation component, and show we can use an immediate bonus function that is significantly smaller than that used in Luo et al. (2021). In particular, we keep the immediate bonus bounded (almost surely) by a constant across the entire state-action space, a property that is essential to arrive at a tighter bound for the least squares estimation procedure.

1.1 Additional Related Work

Tabular RL with stationary and adversarial losses. Tabular RL with stationary losses is perhaps the most fundamental and well studied framework, beginning with the works of Auer & Ortner (2006); Tewari & Bartlett (2007); Jaksch et al. (2010), and with many important advances more recently (Dann & Brunskill, 2015; Azar et al., 2017; Dann et al., 2017; Fruit et al., 2018; Jin et al., 2018). In the context of policy optimization methods in particular, most of the recent works consider the pure optimization perspective or under simplifying exploratory assumptions (e.g., Bhandari & Russo, 2019; Agarwal et al., 2021; Zhan et al., 2021; Lan, 2022), with the exception of Shani et al. (2020) that study the exploration setting and will be discussed momentarily.

The study of adversarially changing costs was initiated in the works of Even-Dar et al. (2009); Yu et al. (2009), and can be largely divided into policy optimization (PO) based methods (Neu et al., 2010; Shani et al., 2020) and algorithms that optimize over the set of occupancy measures (Zimin & Neu, 2013; Rosenberg & Mansour, 2019; Jin et al., 2020a), where both approaches ultimately involve a mirror descent (Nemirovskij & Yudin, 1983; Beck & Teboulle, 2003) optimization component with online guarantees. In the context of PO methods, which are more relevant to our work, Neu et al. (2010) initially achieve $\widetilde{O}(K^{2/3})$ regret for the known dynamics setup with bandit feedback. In a later paper, Shani et al. (2020) present PO algorithms based on value backups for the stochastic and adversarial settings with *unknown* dynamics and bandit feedback, establishing an $\widetilde{O}(\sqrt{K})$ bound in the stochastic case and $\widetilde{O}(K^{2/3})$ in the adversarial case. The recent work of Luo et al. (2021) presents, for the tabular case, a PO algorithm and prove it obtains the optimal $\widetilde{O}(\sqrt{K})$ bound. Their algorithm, as opposed to that of Shani et al. (2020), is not based on value backups but rather stochastic estimates of the cumulative cost. The algorithm we present here combines both approaches.

RL with function approximation. The study of function approximation in RL goes back a long way (e.g., Schweitzer & Seidmann, 1985; Barto, 1990; Bradtke & Barto, 1996; see also Sutton & Barto, 2018 and references therein), although these earlier works did not provide polynomial sample effi-

[1] In order to compute a single action probability of the agent policy, the algorithm of Luo et al. (2021) requires exponentially many simulator samples, generated by traversing the tree structure implicitly defined by the recursive bonus-policy-bonus relation.

ciency. More recently, a line of work initiated by Yang & Wang (2019; 2020); Jin et al. (2020b), studies MDPs with linear structure and focuses on computationally and statistically efficient algorithms (e.g., Zanette et al., 2020b; Modi et al., 2020; Wei et al., 2021; Wagenmaker et al., 2022a). The linear MDP model we adopt here was introduced by Jin et al. (2020b). Also noteworthy is the linear mixture MDP (Modi et al., 2020; Ayoub et al., 2020; Zhou et al., 2021a;b), which is a different model that in general is incomparable with the linear MDP (Zhou et al., 2021b). Finally, there is a rich line of works studying statistical properties of RL with more general function approximation (e.g., Munos, 2005; Jiang et al., 2017; Dong et al., 2020; Jin et al., 2021; Du et al., 2021), although these usually do not provide computationally efficient algorithms.

Policy optimization with function approximation. Formulation of policy optimization methods that incorporate function approximation was given in classical works such as Sutton et al. (1999); Kakade (2001), although these did not study convergence rates nor learning in the exploration setting. More recently, several papers (e.g., Agarwal et al., 2021; Liu et al., 2019) consider convergence properties of policy optimization approaches from a pure optimization perspective, or subject to exploratory assumptions such as bounded concentrability coefficient (Munos, 2003; 2005; Chen & Jiang, 2019), distribution mismatch coefficient or a relative eigenvalue condition (Agarwal et al., 2021). More relevant to our paper are works that consider policy optimization in a setup that requires exploration be handled algorithmically, such as Zanette et al. (2021) who improve upon the prior work of Agarwal et al. (2020), both of which consider stationary losses. The work of Cai et al. (2020) that was mentioned earlier studies the adversarial setting, but in the linear mixture MDP model and with full information feedback.

Concurrent works on adversarial linear MDPs. Two concurrent works study the same setup as ours; online linear MDPs with adversarial costs, unknown dynamics, and bandit feedback. Dai et al. (2023) propose a computationally efficient algorithm that obtains an $\widetilde{O}(K^{8/9})$ regret bound using an algorithmic approach similar to that of (Luo et al., 2021) but with a different OMD regularizer, which contributes as the main source of improvement. Dai et al. (2023) also present a computationally inefficient algorithm for the more general linear-Q setup that obtains a $\widetilde{O}(\sqrt{K})$ bound when given simulator access. The work of Kong et al. (2023) takes a different approach; namely, a linear bandits blackbox algorithm, and obtains a $\widetilde{O}(K^{4/5} + 1/\lambda^\star_{\min})$[2] regret bound, albeit with a computationally inefficient algorithm. At a high level, the algorithm proposed by Kong et al. (2023) first learns approximations to the expected feature occupancies of a class of sufficiently expressive policies, then runs a variant of geometric-hedge (Dani et al., 2007) for the rest of the game.

[2]Here, $\lambda^\star_{\min}$ denotes the minimum eigenvalue of the best exploratory policy's 2nd moment matrix.

2 Problem Setup

Episodic MDPs. A finite horizon episodic MDP is defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \ell, s_1)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action set, $H \in \mathbb{Z}_+$ the length of the horizon, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^{H-1}$ the time dependent transition function, $\ell = \{\ell_h\}_{h=1}^{H}$ a sequence of loss functions, and $s_1 \in \mathcal{S}$ the initial state that we assume to be fixed w.l.o.g. The transition density given the agent is at state $s \in \mathcal{S}$ at time h and takes action a is given by $\mathbb{P}_h(\cdot|s,a) \in \Delta(\mathcal{S})$. After the agent takes an action on the last time step H, the episode terminates immediately. We assume the state space $\mathcal{S}$ is measurable space (which may contain uncountably many states) and the action set $\mathcal{A}$ is finite with $A := |\mathcal{A}|$. A policy is defined by a mapping $\pi\colon \mathcal{S}\times[H] \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the probability simplex over the action set $\mathcal{A}$. We let $\pi_h(\cdot|s) \in \Delta(\mathcal{A})$ denote the distribution over actions given by π at s, h. Finally, we use the convention that for any function $V\colon \mathcal{S} \to \mathbb{R}$, we interpret $\mathbb{P}_h V\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ as the result of applying the conditional expectation operator $\mathbb{P}_h$; $\mathbb{P}_h V(s,a) := \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} V(s')$ (see Appendix A for comments regarding this notation).

Episodic Linear MDPs with adversarial costs. We consider the adversarial online learning setup, with unknown dynamics and bandit feedback. In this setup, the agent interacts with the MDP over the course of $K \ge 1$ episodes, where in each episode, the loss function associated with the MDP changes as chosen by an adversary that observes the current and past player policies. The feedback provided to the learner consists of the instantaneous scalar loss associated with the state-action pairs she has visited during episode rollout. Our central structural assumption is that the combination of transition function and adversarial losses form a *linear MDP* (Jin et al., 2020b) in each episode.

Assumption 1 (Linear MDP with changing costs)**.** The learner interacts with a sequence of MDPs $\{\mathcal{M}^k\}_{k=1}^K$, $\mathcal{M}^k = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \ell^k, s_1)$ that share all elements other than the loss functions, such that the following holds. There is a feature mapping $\phi\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ that is **known** to the learner, and for every $h \in [H]$, d **unknown** signed measures $\psi_{h,1}, \dots, \psi_{h,d} \in \mathcal{S} \to \mathbb{R}$ forming $\psi_h(\cdot) := (\psi_{h,1}(\cdot), \dots, \psi_{h,d}(\cdot)) \in \mathcal{S} \to \mathbb{R}^d$, such that for all $h, s, a, s' \in [H-1]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}$:

$$\mathbb{P}_h(s'|s,a) = \phi(s,a)^\top \psi_h(s'). \quad (1)$$

W.l.o.g., we assume $\|\phi(s,a)\| \le 1$ for all s, a, and that for any measurable function $f\colon \mathcal{S} \to \mathbb{R}$ with $\|f\|_\infty \le 1$, it holds that $\left\|\int \psi_h(s')f(s')\mathrm{d}s'\right\| \le \sqrt{d}$ for all $h \in [H]$. In addition, for all k;

$$\ell_h^k(s,a) = \phi(s,a)^\top \mathbf{c}_h^k, \tag{2}$$

where $\left\{\mathbf{c}_h^k\right\}$ are adversarially chosen cost vectors. W.l.o.g., we assume $\left|\phi(s,a)^\top \mathbf{c}_h^k\right| \le 1$ for all s, a, h, k, and $\left\|\mathbf{c}_h^k\right\| \le \sqrt{d}$ for all h, k.

The pseudocode for learner environment interaction is provided below in Protocol 1.

Protocol 1 Learner-Environment Interaction

```
parameters: (S, A, H, P, φ, s_1; K)
for k = 1, ..., K do
  agent decides on a policy π^k
  adversary chooses H cost vectors {c_h^k} ∈ R^d
  define ℓ_h^k: S × A → R by ℓ_h^k(s,a) = φ(s,a)^⊤ c_h^k.
  environment resets to s_1^k = s_1
  for h = 1, ..., H do
    agent observes s_h^k ∈ S
    agent chooses a_h^k ~ π_h^k(·|s_h^k)
    agent observes and incurs loss ℓ_h^k = ℓ_h^k(s_h^k, a_h^k)
    if h < H:
      environment transitions to s_{h+1}^k ~ P_h(·|s, a)
  end for
end for
```

We make the following additional notes with regards to the model we consider: (1) for any $s, a \in \mathcal{S} \times \mathcal{A}$, the agent may evaluate $\phi(s,a)$ in $O(1)$ time; (2) we assume an oblivious and deterministic adversary (but in fact our results hold more generally for the case that the adversary is random, and observes the agent's policies, *but not* trajectory realizations); (3) with slight overloading of notation, we let $\ell_h^k = \ell_h^k(s_h^k, a_h^k)$ denote the random loss incurred by the agent on episode k time step h.

Learning objective. The expected loss of a policy π when starting from state $s \in \mathcal{S}$ at time step $h \in [H]$ is given by the value function;

$$V_h^\pi(s;\ell) := \mathbb{E}\left[\sum_{t=h}^{H} \ell_t(s_t, a_t) \mid s_h = s, \pi, \ell\right], \tag{3}$$

where we use the extra $(;\ell)$ notation to emphasize the specific loss function considered. The expected loss conditioned on the agent taking action $a \in \mathcal{A}$ on time step h at s and then continuing with π is given by the action-value function;

$$Q_h^\pi(s,a;\ell) := \mathbb{E}\left[\sum_{t=h}^{H} \ell_t(s_t, a_t) \mid s_h = s, a_h = a, \pi, \ell\right]. \tag{4}$$

The value and action-value functions of a policy π in the MDP $\left(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \ell^k, s_1\right)$ associated with episode $k \in [K]$ are denoted by, respectively;

$$V_h^{k,\pi}(s) := V_h^\pi(s;\ell^k);\ Q_h^{k,\pi}(s,a) := Q_h^\pi(s,a;\ell^k),$$

where $V_h^\pi(s;\ell^k)$ and $Q_h^\pi(s,a;\ell^k)$ have been defined in Equations (3) and (4).

We let $\pi^\star$ denote the best policy in hindsight;

$$\pi^\star := \arg\min_\pi \left\{\sum_{k=1}^{K} V_1^{k,\pi}(s_1)\right\},$$

and seek to minimize the *pseudo regret* of the agent policy sequence $\pi^1, \ldots, \pi^K$;

$$\text{Regret} := \sum_{k=1}^{K} V_1^{k,\pi^k}(s_1) - V_1^{k,\pi^\star}(s_1). \tag{5}$$

Finally, we note that $\pi^\star$ may depend on player decisions, as the adversary is adaptive.

Additional notation and definitions. We let $\|\cdot\| = \|\cdot\|_2$ denote the standard Euclidean norm, and for a positive definite matrix $\Lambda \in \mathbb{R}^{d\times d}$, we let $\|v\|_\Lambda = \sqrt{v^\top \Lambda v}$ denote the weighted norm induced by Λ. Further, we let $\|\Lambda\| = \|\Lambda\|_{\text{op}} = \max_{v,\|v\|=1} v^\top \Lambda v$ denote the operator norm of Λ. Finally, we use $\text{clip}\,[\mathrm{x}]_{\mathrm{b}}^{\mathrm{a}} := \max\{\min\{\mathrm{x},\mathrm{a}\},\mathrm{b}\}$ to denote clipping of a real scalar x between $a \in \mathbb{R}$ and $b \in \mathbb{R}$.

3 Algorithm and Main Result

The pseudocode for our main algorithm; **P**olicy **O**ptimization with **L**east **S**quares **B**onus **E**xploration, is provided in Algorithm 1. The high level algorithmic template is relatively simple; (1) Rollout π^k in the environment; (2) Obtain a (nearly) unbiased estimate $\widehat{Q}^k$ of Q^k; (3) Construct a bonus-to-go estimate $\widetilde{B}^k$ through least squares policy evaluation in an auxiliary bonus MDP; (4) Perform a mirror-descent update step using the optimistic Q^k function estimate given by $\widehat{Q}^k - \widetilde{B}^k$.

The bonus-to-go estimate is obtained by the least squares policy evaluation subroutine Algorithm 2 (discussed in Section 3.1), which outputs an approximation that is optimistic and with bias that can be controlled efficiently. This provides for the major contributing factor in the final regret guarantee; specifically, this approach along with a refined instantaneous Q-bonus design allows us to avoid the policy cover used in Luo et al. (2021), and leads to a simpler algorithm that explores more efficiently. The final bonus function $\widetilde{B}^k$ encompasses two bonus types; one to compensate for uncertainty in the Q^k estimates (b^k in Equation (9)), and the other ($b^{\mathbb{P},k}$ in Algorithm 2) to compensate for uncertainty in the estimation of the dynamics in the policy evaluation procedure. Intuitively speaking, given the agent is at state s_h, her bonus for taking action a_h will be high when the expected rollout following a_h traverses state action pairs (s_t, a_t) for which (1) we have poor next state information s_{t+1}, and (2) their feature vector $\phi(s_t, a_t)$ points in a direction in the state-action space for which we have poor knowledge of past Q-cost vectors $\mathbf{q}_t^1, \ldots, \mathbf{q}_t^k$ (these are the low dimensional representations of the Q functions; see Lemma 1). On a conceptual level, $b^{\mathbb{P},k}$ drives exploration for the purpose of learning the dynamics, and b^k for the sake of cost function information.

Algorithm parameters. Algorithm 1 takes as arguments the mirror descent step-size η, the 2nd-moment regularization parameter γ, the bonus coefficients β (for instantaneous loss estimation) and $\beta^{\mathbb{P}}$ (for value backup estimation), and finally ϵ, σ^2 which control respectively, the bias and variance of the inverse 2nd-moment estimation procedure.

Two-way partitioned blocking. In order to estimate feature occupancy covariance matrices and Bellman backup operators, Algorithm 1 plays each policy multiple times. For a given parameter $\tau \geq 1$, we divide episodes $k \in [K]$ into $\lceil K/(2\tau) \rceil$ blocks, and assume for simplicity of exposition that $K/(2\tau)$ is an integer. We define for all $j \in [K/(2\tau)]$;

$$T_{j,1} \coloneqq \{(j-1)\tau + 1, \ldots, j\tau\}, \tag{6}$$
$$T_{j,2} \coloneqq \{j\tau + 1, \ldots, (j+1)\tau\}, \tag{7}$$
$$T_j \coloneqq T_{j,1} \cup T_{j,2}. \tag{8}$$

For all episodes $k \in T_j$ (which we call block j), the policy is held fixed and denoted $\pi^{(j)}$. We let π^k denote the policy played on episode k throughout, thus $\pi^k = \pi^{(j)}$ for all $k \in T_j$. The two-way partitioning of each block described above is motivated by the need to keep Q-function estimates (nearly) unbiased.

These estimates build on multiplying the inverse covariance estimator $\widehat{\Sigma}^+_{kh\gamma}$ with the state-action feature vector $\phi(s_h^k, a_h^k)$ and the scalar cumulative loss $\sum_{t=h}^H \ell_t^k$. In order to show (near) unbiasedness, we need to argue the inverse covariance estimator is independent of the other factors. The two way partitioning is a simple mechanism to ensure the estimator $\widehat{\Sigma}^+_{kh\gamma}$ does not involve samples from the k'th episode (to which $\phi(s_h^k, a_h^k)$ and the cumulative loss belong). Thus, the desired independence follows immediately (see Lemma 2). Specifically, for $k \in T_{j,1}$, we construct the inverse-covariance estimator from the collection of samples in $T_{j,2}$, which does not contain k (note the estimation procedure takes place at the end of the block, hence this is possible). Similarly, for for $k \in T_{j,2}$, we use samples from $T_{j,1}$. Throughout, we let $\mathcal{D}^k = \{\mathcal{D}_h^k\}_{h=1}^H$ denote the dataset used for estimations of episode k, and slightly abuse notation by referring to it as either containing episode indices, or transition tuples $(s_h^i, a_h^i, s_{h+1}^i)$.

Algorithm 1 PO-LSBE

input: $(\eta, \gamma, \beta, \beta^{\mathbb{P}}, \epsilon, \sigma^2)$
Set $M = \frac{48d}{\gamma\sigma} \log \frac{72d}{\gamma^2\sigma}$, $N = \frac{2}{\gamma} \log \frac{1}{\gamma\epsilon}$, $\tau = MN$.
Initialize $\pi^{(1)}$ to take actions uniformly at random.
for $j = 1, \ldots, \lceil K/(2\tau) \rceil$ **do**
  Play $\pi^k = \pi^{(j)}$ for the 2τ episodes $k \in T_j$ (defined in Equation (8)), and collect $(s_h^k, a_h^k, \ell_h^k)_{h\in[H],k\in T_j}$
  for $k \in T_j$ **do**
    if $k \in T_{j,1}$ populate $\mathcal{D}^k$ with $T_{j,2}$ rollouts
    otherwise ($k \in T_{j,2}$) populate $\mathcal{D}^k$ with $T_{j,1}$ rollouts

    $\widehat{\Sigma}^+_{kh\gamma} \leftarrow \mathrm{MGR}(\mathcal{D}_h^k; N, M, \gamma)$ (see Algorithm 3)
    $\widehat{\mathbf{q}}_h^k \leftarrow \widehat{\Sigma}^+_{kh\gamma} \phi(s_h^k, a_h^k) \sum_{t=h}^H \ell_t^k$
    $\widehat{Q}_h^k(s,a) = \phi(s,a)^\top \widehat{\mathbf{q}}_h^k$

    Define the Q-bonus by

$$b_h^k(s,a) = \beta\Big(\|\phi(s,a)\|_{\widehat{\Sigma}^+_{kh\gamma}} + \big\langle \pi_h^k(\cdot|s), \|\phi(s,\cdot)\|_{\widehat{\Sigma}^+_{kh\gamma}} \big\rangle \Big) \tag{9}$$

    Compute the bonus-to-go with Algorithm 2;

$$\widetilde{B}^k \leftarrow \mathrm{OLSPE}(\mathcal{D}^k, b^k; \beta^{\mathbb{P}}, \beta, \gamma)$$

  end for
  Policy improvement step:

$$\pi_h^{(j+1)}(a|s) \propto \exp\left(-\eta \sum_{i=1}^{j} \mathcal{L}_h^{(i)}(s,a)\right)$$

$$\text{where } \mathcal{L}_h^{(j)}(s,a) = \frac{1}{\tau} \sum_{k\in T_j} \widehat{Q}_h^k(s,a) - \widetilde{B}_h^k(s,a)$$

end for

Our main result stated below establishes the regret bound for Algorithm 1.

Theorem 1. *With an appropriate choice of parameters (see Appendix B for details) and assuming $K = \Omega((d \log d)^2)$, Algorithm 1 obtains an expected regret guarantee of*

$$\mathbb{E}\,[\mathrm{Regret}] = \widetilde{O}\left(dH^2K^{6/7} + d^{3/2}H^4K^{5/7}\right),$$

where big-$\widetilde{O}$ hides constant and logarithmic factors.

3.1 Least Squares Policy Evaluation in Bonus MDPs

The **O**ptimistic-**L**east-**S**quares-**P**olicy-**E**valuation (OLSPE) procedure given in Algorithm 2 is a variant of LSVI-UCB (Jin et al., 2020b, see also Agarwal et al., 2019) that is aimed at policy evaluation, and tasked with the computation of the bonus-to-go estimates $\widetilde{B}^k$. Unlike prior works, we evaluate the policy's bonus (i.e., exploration) coverage, rather than its loss performance (which is estimated separately, in Algorithm 1) in an auxiliary *full information* bonus MDP. Given the immediate Q-bonus b^k of episode k, we consider the bonus MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, b^k, s_1)$, which should be interpreted as a *reward* MDP, as the agent will be trying to collect *higher* bonus values. It is immediate to see that this *is not* a linear MDP, as the reward function b^k is non-linear. Nonetheless, the dynamics do admit a linear factorization (as per Assumption 1), which allows the use of least squares regression to approximate the value and action-value functions in this MDP.

For any policy π, we denote the true value and action-value functions in the bonus MDP of episode k, respectively, by

$$B_h^{k,\pi}(s,a) := Q_h^\pi(s,a;b^k), \tag{10}$$

$$W_h^{k,\pi}(s) := V_h^\pi(s;b^k). \tag{11}$$

Algorithm 2 computes optimistic versions of the above functions for the policy passed as input, which on episode k is always the agent's policy π^k. These are denoted by $\widetilde{B}^k$ and $\widetilde{W}^k$, and defined by the algorithm in Equations (14) and (15). In accordance, we let $\widetilde{\mathbb{P}}_h^k$ defined in Equation (13) denote the optimistic estimate of the conditional expectation operator given by the dataset $\mathcal{D}_h^k$. Our notation here is motivated by the true conditional expectation operator $\mathbb{P}_h$; recall we adopt the convention that $\mathbb{P}_h W(s,a) = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} W(s')$ for any function $W \colon \mathcal{S} \to \mathbb{R}$.

Algorithm 2 OLSPE($\mathcal{D}^k, b^k; \beta^{\mathbb{P}}, \beta, \gamma$)

Set $\lambda = 1$
$\widetilde{W}_{H+1}^k(\cdot) = 0$
for $h = H, \ldots, 1$ **do**

$$\Lambda_h^k \leftarrow \lambda I + \sum\nolimits_{i \in \mathcal{D}_h^k} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top$$

$$\widehat{\mathbf{w}}_h^k \leftarrow \left(\Lambda_h^k\right)^{-1} \sum\nolimits_{i \in \mathcal{D}_h^k} \phi(s_h^i, a_h^i) \widetilde{W}_{h+1}^k(s_{h+1}^i)$$

$$b_h^{\mathbb{P},k}(s,a) = \beta^{\mathbb{P}} \left\|\phi(s,a)\right\|_{(\Lambda_h^k)^{-1}} \tag{12}$$

$$\widetilde{\mathbb{P}}_h^k \widetilde{W}_{h+1}^k(s,a) = \phi(s,a)^\top \widehat{\mathbf{w}}_h^k + b_h^{\mathbb{P},k}(s,a) \tag{13}$$

$$B_h^{\max} = 2\beta(H-h+1)/\sqrt{\gamma}$$

$$\widetilde{B}_h^k(s,a) = \text{clip}\left[b_h^k(s,a) + \widetilde{\mathbb{P}}_h^k \widetilde{W}_{h+1}^k(s,a)\right]_0^{B_h^{\max}} \tag{14}$$

$$\widetilde{W}_h^k(s) = \left\langle \pi^k(\cdot|s), \widetilde{B}_h^k(s,\cdot) \right\rangle \tag{15}$$

end for
return $\widetilde{B}^k = \{\widetilde{B}_h^k\}_{h \in [H]}$

3.2 Obtaining unbiased Q estimates

In order to construct estimates of the loss vector associated with the action-value function of episode k time step h, Q_h^{k,π^k}, we follow prior works and use a linear bandit type estimation procedure (e.g., Dani et al., 2007). Unlike the linear bandit setting, here we do not know the feature occupancy covariance matrix, and moreover it may not be well conditioned. We address both of these issues in the same natural manner as did Luo et al. (2021); we estimate a γ-regularized version of the inverse covariance using the Matrix Geometric Resampling (MGR) procedure of Neu & Olkhovskaya (2020a) (see also Neu & Olkhovskaya, 2021). Like Luo et al. (2021), we employ a version of MGR given in Algorithm 3 that averages over multiple estimators to get better control of the variance of the final output, however we obtain tighter bounds owed to a refined analysis (see Lemma 3).

Algorithm 3 MGR ($\mathcal{D}, N, M, \gamma$)

Set $c = 1/2$
Enumerate samples in $\mathcal{D}$ by $\{\phi_{m,n}\}_{m \in [M], n \in [N]}$
Let $A_{m,n} = \gamma I + \phi_{m,n}\phi_{m,n}^\top \quad \forall m, n$
for $m = 1, \ldots, M$ **do**
  for $n = 1, \ldots, N$ **do**
    $\widehat{\Sigma}_{m,\gamma}^{(n)} \leftarrow \prod_{i=1}^n (I - cA_{m,i})$
  end for
  $\widehat{\Sigma}_{m,\gamma}^+ \leftarrow cI + c\sum_{n=1}^N \widehat{\Sigma}_{m,\gamma}^{(n)}$
end for
return $\widehat{\Sigma}_\gamma^+ = \frac{1}{M}\sum_{m=1}^M \widehat{\Sigma}_{m,\gamma}^+$

4 The Simulator Setting

The pseudocode for the simulator version of our method is given in Algorithm 4 below. It has the same structure as the simulator based algorithm proposed by Luo et al. (2021) for the linear-Q setting, only that our bonus-to-go is computed using optimistic approximations via Algorithm 2. Notably, the simulator required by our algorithm is weaker than that of Luo et al. (2021); we only need to execute agent policies from the initial state s_1, but do not require next state samples from arbitrarily chosen state action pairs. Formally, we make the following assumption in this section.

Assumption 2 (simulator access)**.** The learner has access to a simulator, which takes a policy π as input and returns a trajectory $(s_h, a_h)_{h=1}^H$ sampled from the MDP using π; $a_h \sim \pi(\cdot|s_h)$, and $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h)$.

We note that Algorithm 4 follows the exact same algorithmic design as Algorithm 1; only that instead of blocking, the version presented here executes simulator rollouts. The significance of the result presented next is two-fold. First, it establishes the state-of-the-art regret bound for the simulator setting with a computationally efficient algorithm. Second, it demonstrates the guarantee our approach would yield without the limiting factor of the number of online samples; specifically, that given $\widetilde{O}(K^{4/3})$ additional samples per episode, we arrive at a $\widetilde{O}(K^{2/3})$ regret bound.

Theorem 2. *With an appropriate choice of parameters and assuming $K = \Omega((d \log d)^2)$, under Assumption 2, Algorithm 4 obtains an expected regret guarantee of*

$$\mathbb{E}\left[\text{Regret}\right] = \widetilde{O}\left(H^2(dK)^{2/3} + H^4(dK)^{1/3}\right),$$

where big-$\widetilde{O}$ hides constant and logarithmic factors. Furthermore, the number of simulator rollouts required per episode is $\widetilde{O}(K^{4/3})$.

Algorithm 4 PO-LSBE (simulator version)

input: $(\eta, \gamma, \beta, \beta^{\mathbb{P}}, \epsilon, \sigma^2)$, and a simulator
Set $M = \frac{48d}{\gamma\sigma} \log \frac{72d}{\gamma^2\sigma}$, $N = \frac{2}{\gamma} \log \frac{1}{\gamma\epsilon}$, $\tau = d^2MN$.
Initialize π^1 to take actions uniformly at random.
for $k = 1, \ldots, K$ **do**
  Rollout π^k in and collect $\left\{(s_h^k, a_h^k, \ell_h^k)\right\}_{h=1}^H$
  Populate $\mathcal{D}^k$ with τ *simulator rollouts* of π^k

  $\widehat{\Sigma}^+_{kh\gamma} \leftarrow \text{MGR}(\mathcal{D}_h^k; N, M, \gamma)$ (see Algorithm 3)
  $\widehat{\mathbf{q}}_h^k \leftarrow \widehat{\Sigma}^+_{kh\gamma}\phi(s_h^k, a_h^k)\sum_{t=h}^H \ell_t^k$
  $\widehat{Q}_h^k(s,a) = \phi(s,a)^\top \widehat{\mathbf{q}}_h^k$

  Define the Q-bonus as in Equation (9)
  Compute the bonus-to-go with Algorithm 2;

$$\widetilde{B}^k \leftarrow \text{OLSPE}(\mathcal{D}^k, b^k; \beta^{\mathbb{P}}, \beta, \gamma)$$

  Policy improvement step:

$$\pi_h^{k+1}(a|s) \propto \exp\left(-\eta \sum_{i=1}^k \widehat{Q}_h^i(s,a) - \widetilde{B}_h^i(s,a)\right)$$

end for

5 Analysis of Main Algorithm

In this section, we present an overview of the proof of Theorem 1. We will make use of some additional notation described next. The state-action occupancy measure induced by a policy π on time step h is denoted $d_h^\pi(s,a) = \Pr(s_h = s, a_h = a \mid \pi)$, and with slight overloading $d_h^\pi(s) = \sum_a d_h^\pi(s,a)$ denotes the state occupancy measure. In sake of conciseness, we let

$$d_h^k := d_h^{\pi^k}, \quad d_h^\star := d_h^{\pi^\star}, \tag{16}$$

denote the occupancy measures of, respectively, the agent's policy on episode k and the benchmark policy $\pi^\star$. We let $\mathbb{E}_k\left[\cdot\right] = \mathbb{E}\left[\cdot | \pi^k, \ldots, \pi^1\right]$ denote the expected value of random variables conditioned on the sequence of agent policies up to and including episode k; and note this only indicates conditioning on policies and not trajectory rollouts. Finally, we may also use the more compact notation

$$Q_h^k := Q_h^{k,\pi^k}, \tag{17}$$

to refer to the true action-value function of the agent's policy π^k in the MDP of episode k.

In what follows, we present the high level components of the analysis and provide a proof sketch for Theorem 1; for the full technical details, see Appendix B. Our high level proof structure is an extended (and slightly reframed) version of the one proposed by Luo et al. (2021). We consider the following regret decomposition;

$$\underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{s\sim d_h^\star}\left[\left\langle Q_h^k(s,\cdot) - \widehat{Q}_h^k(s,\cdot), \pi_h^k(\cdot|s)\right\rangle\right]}_{\text{Bias1}}$$
$$+ \underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{s\sim d_h^\star}\left[\left\langle \widehat{Q}_h^k(s,\cdot) - Q_h^k(s,\cdot), \pi_h^\star(\cdot|s)\right\rangle\right]}_{\text{Bias2}}$$
$$+ \underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{s\sim d_h^\star}\left[\left\langle \widetilde{Q}_h^k(s,\cdot), \pi_h^k(\cdot|s) - \pi_h^\star(\cdot|s)\right\rangle\right]}_{\text{OMD}}$$
$$+ \underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{s\sim d_h^\star}\left[\left\langle \widetilde{B}_h^k(s,\cdot), \pi_h^k(\cdot|s) - \pi_h^\star(\cdot|s)\right\rangle\right]}_{\text{Exploration}},$$

where $\widetilde{Q}_h^k(s,a) := \widehat{Q}_h^k(s,a) - \widetilde{B}_h^k(s,a)$. An important observation made in Luo et al. (2021) was that with an appropriate bonus design, the bias and OMD terms contribute $\sum_k V^{\pi^\star}(s_1; b^k)$, while the exploration term contributes the *exact negative* of this quantity. Fortunately, what we will pay for exploration (with a positive term), are the bonuses collected along trajectories of the agent's policy, which may be bounded efficiently.

Bounding the exploration term. We begin by establishing confidence bounds on the bonus-to-go estimations computed by Algorithm 2 and defined in Equations (14) and (15).

Lemma (simplified statement of Lemma 4). For any $\delta > 0$, an appropriate choice of parameters ensures that w.p. $\geq 1-\delta$ the following holds for all k, h, s, a;

$$\widetilde{B}_h^k(s,a) \geq b_h^k(s,a) + \mathbb{P}_h\widetilde{W}_{h+1}^k(s,a) \tag{18}$$
$$\widetilde{B}_h^k(s,a) \leq b_h^k(s,a) + \mathbb{P}_h\widetilde{W}_{h+1}^k(s,a) + 2b_h^{\mathbb{P},k}(s,a) \tag{19}$$

The proof follows from uniform concentration of the least squares estimates over the class of bonus value functions explored by the algorithm; the arguments are similar in spirit to those made in the work of Jin et al. (2020b). Next, we use the confidence bounds to deduce a bound on the exploration term. The lemma below contains a part that is implicit in Luo et al. (2021) Lemma B.1, and an extension to incorporate the effect of the bonus-to-go approximations. We note our proof below provides a simpler argument than the original of Luo et al. (2021), by offloading most of the technicalities to the extended value difference Lemma 26.

Lemma (compact restatement of Lemma 10). Assume that both Equations (18) and (19) hold. Then, we have that EXPLORATION $\leq$

$$2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a)+b_h^k(s,a)\right] - \sum_{k=1}^{K}V_1^{k,\pi^\star}(s_1;b^k). \tag{20}$$

Proof sketch. By the lower bound on $\widetilde{B}_h^k(s,a)$ Equation (18), we have EXPLORATION $\leq$

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\left\langle \widetilde{B}_h^k(s,\cdot),\pi_h^k(\cdot|s)-\pi_h^\star(\cdot|s)\right\rangle\right] + \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^\star}\left[\widetilde{B}_h^k(s,a)-b_h^k(s,a)-\mathbb{P}_h\widetilde{W}_{h+1}^k(s,a)\right] = \sum_{k=1}^{K}\widetilde{W}_1^k - W_1^{k,\pi^\star},$$

where the inequality is since we only add non-negative terms, and the equality follows from the extended value difference Lemma 26 with $\widehat{V}_1^\pi = \widetilde{W}_1^k = \widetilde{W}_1^{k,\pi^k}$ and $V_1^{\pi'} = W_1^{k,\pi^\star}$ (and we recall definitions in Equations (11) and (15)). Next, using Lemma 26 again and our upper bound on $\widetilde{B}_h^k(s,a)$ given by Equation (19), establishes that $\sum_{k=1}^K \widetilde{W}_1^k - W_1^{k,\pi^k} \leq 2\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a)\right]$. Therefore,

$$\sum_{k=1}^{K}\widetilde{W}_1^k - W_1^{k,\pi^\star} = \sum_{k=1}^{K}\widetilde{W}_1^k - W_1^{k,\pi^k} + \sum_{k=1}^{K}W_1^{k,\pi^k} - W_1^{k,\pi^\star} \leq 2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a)\right] + \sum_{k=1}^{K}W_1^{k,\pi^k} - W_1^{k,\pi^\star},$$

which completes the proof after substituting for the definition of true bonus value functions Equation (11). □

From this point, it is not hard to obtain an in expectation bound;

$$\mathbb{E}\left[\text{EXPLORATION}\right] \lesssim \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{d_h^k}\left[b_h^{\mathbb{P},k}(s,a)+b_h^k(s,a)\right]\right] - \mathbb{E}\left[\sum_{k=1}^{K}V_1^{k,\pi^\star}(s_1;b^k)\right]. \tag{21}$$

Notably, the arguments thus far do not depend on the particular form of the immediate bonuses b^k, suggesting we would like to choose the bonus so that as much of BIAS1, BIAS2 and OMD can be expressed as $V_1^{k,\pi^\star}(s_1;b^k)$.

Bounding BIAS1 + BIAS2. To bound these terms, we employ relatively standard arguments in similar nature to those of Luo et al. (2021). However, we aim for a different immediate bonus function, earning important savings in the policy evaluation procedure. Henceforth, we let

$$\Sigma_{kh} := \mathbb{E}_{s,a\sim d_h^k}\left[\phi(s,a)\phi(s,a)^\top\right] \tag{22}$$

denote the true covariance matrix of the feature occupancy induced by π^k on time step h, and denote by $\Sigma_{kh\gamma} := \gamma I + \Sigma_{kh}$ the γ-regularized version of it.

Lemma (simplified restatement of Lemma 7). For the immediate bonus function b^k defined in Equation (9) and an appropriate choice of parameters, we have that the expected bias terms are bounded as

$$\mathbb{E}\left[\text{BIAS1}+\text{BIAS2}\right] \leq \left(\sqrt{\gamma d H^2}\right)\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\sum_a u_h^k(s,a)\right]\right] + 4\epsilon H^2 K,$$

where $u_h^k(s,a) := \left(\pi_h^k(a|s)+\pi_h^\star(a|s)\right)\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^+}$.

Proof sketch. Since the MDPs on each episode are linear, we have $Q_h^{k,\pi^k}(s,a) = \phi(s,a)^\top \mathbf{q}_h^k$ for some $\mathbf{q}_h^k \in \mathbb{R}^d$ of bounded norm. In addition,

$$\mathbb{E}_k\left[\widehat{\mathbf{q}}_h^k\right] = \mathbb{E}_k\left[\widehat{\Sigma}_{kh\gamma}^+\right]\Sigma_{kh}\mathbf{q}_h^k,$$

and with an appropriate choice of parameters, our inverse covariance estimator is only ϵ-biased (see Lemma 3), which can be used to show that

$$\mathbb{E}_k\left[Q_h^k(s,a)-\widehat{Q}_h^k(s,a)\right] = \mathbb{E}_k\left[\phi(s,a)^\top\left(\mathbf{q}_h^k-\widehat{\mathbf{q}}_h^k\right)\right] \leq \gamma\phi(s,a)^\top\Sigma_{kh\gamma}^{-1}\mathbf{q}_h^k + \epsilon H.$$

Using standard algebraic manipulations, we can further bound the first term on the RHS by $\sqrt{\gamma d}H\mathbb{E}_k\left[\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^+}\right] + \epsilon H$, which leads to

$$\mathbb{E}_k\left[Q_h^k(s,a)-\widehat{Q}_h^k(s,a)\right] \leq \sqrt{\gamma d}H\mathbb{E}_k\left[\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^+}\right] + 2\epsilon H.$$

The proof is complete by summing the appropriate terms in BIAS1 and BIAS2, and adding them together. □

From this point, it is not hard to show that owed to our choice of bonus function b^k, the result of the above lemma becomes;

$$\mathbb{E}\left[\text{BIAS1} + \text{BIAS2}\right] \lesssim \frac{1}{2}\mathbb{E}\left[\sum_{k=1}^{K} V_1^{\pi^\star}(s_1; b^k)\right] + \epsilon H^2 K. \tag{23}$$

Bounding OMD term. The variance of our estimators $\widehat{\Sigma}^+_{kh\gamma}$ comes into play in the second moment bound derived on the basic mirror-descent guarantee. Using a refined analysis, we show in Lemma 3 that $\tau = O(1/\gamma^2)$ samples are sufficient to ensure, for $\sigma = 1/4$;

$$\mathbb{E}\left[\widehat{\Sigma}^+_{kh\gamma}\Sigma_{kh\gamma}\widehat{\Sigma}^+_{kh\gamma}\right] \preceq 2\mathbb{E}\left[\widehat{\Sigma}^+_{kh\gamma}\right] + \sigma I,$$

Using the above, we prove;

Lemma (simplified restatement of Lemma 6). Upon executing Algorithm 1 with an appropriate choice of parameters, we have for any s, h;

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\left\langle \widetilde{Q}_h^k(s,\cdot), \pi_h^k(\cdot|s) - \pi_h^\star(\cdot|s)\right\rangle\right] \\ &\lesssim \frac{\eta H^2}{\sqrt{\gamma}}\mathbb{E}\left[\sum_{k=1}^{K}\sum_a \pi_h^k(a|s)\,\|\phi(s,a)\|_{\widehat{\Sigma}^+_{kh\gamma}}\right] \\ &\quad + \frac{\tau}{\eta} + \frac{\eta\beta^2 H^2 K}{\gamma} + \eta(1+\sigma)H^2 K.\end{aligned}$$

Taken together, these, along with our choice of bonus function b^k, establish that

$$\mathbb{E}\left[\text{OMD}\right] \lesssim \frac{1}{2}\mathbb{E}\left[\sum_{k=1}^{K} V_1^{\pi^\star}(s_1; b^k)\right] + \frac{H}{\eta\gamma^2} + \eta H^3 K. \tag{24}$$

Concluding the proof. Combining Equations (21), (23) and (24), and focusing on dependence on K, we obtain

$$\begin{aligned}\mathbb{E}\left[\text{Regret}\right] \lesssim\ &\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{d_h^k}\left[b_h^{\mathbb{P},k}(s,a) + b_h^k(s,a)\right]\right] \\ &+ \frac{1}{\eta\gamma^2} + \eta K + \epsilon K.\end{aligned}$$

We bound the bonus terms collected along the agent's trajectories above using standard arguments in Lemmas 11 and 13, arriving at

$$\mathbb{E}\left[\text{Regret}\right] \lesssim \sqrt{\gamma}K + \frac{1}{\eta\gamma^2} + \eta K + \epsilon K.$$

We can easily rid of the bias term ϵK as τ depends on it only logarithmically. Finally, the first two terms dominate the regret at $\widetilde{O}(K^{6/7})$ for the setting of $\eta = \gamma/(2H)$, and $\gamma = K^{-2/7}$, and the proof is complete.

Acknowledgements

The authors would like to thank Asaf Cassel for many helpful discussions. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grants 993/17, 2549/19), by the Len Blavatnik and the Blavatnik Family foundation, by the Yandex Initiative in Machine Learning at Tel Aviv University, and by a grant from the Tel Aviv University Center for AI and Data Science (TAD).

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pp. 10–4, 2019.

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

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76, 2021.

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Auer, P. and Ortner, R. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19, 2006.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

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.

Barto, A. G. Connectionist learning for control. *Neural networks for control*, pp. 5–58, 1990.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

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, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Dai, Y., Luo, H., Wei, C.-Y., and Zimmert, J. Refined regret for adversarial mdps with linear function approximation. *arXiv preprint arXiv:2301.12942*, 2023.

Dani, V., Kakade, S. M., and Hayes, T. The price of bandit information for online optimization. *Advances in Neural Information Processing Systems*, 20, 2007.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Dong, K., Peng, J., Wang, Y., and Zhou, Y. Root-n-regret for learning in markov decision processes with function approximation and low bellman rank. In *Conference on Learning Theory*, pp. 1554–1557. PMLR, 2020.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pp. 2826–2836. PMLR, 2021.

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., and Hester, T. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

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.

Hazan, E. et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 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*, pp. 4860–4869. PMLR, 2020a.

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

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Jin, T., Lancewicki, T., Luo, H., Mansour, Y., and Rosenberg, A. Near-optimal regret for adversarial mdp with delayed bandit feedback. *arXiv preprint arXiv:2201.13172*, 2022.

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

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

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

Lan, G. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pp. 1–48, 2022.

Lancewicki, T., Rosenberg, A., and Mansour, Y. Learning adversarial markov decision processes with delayed feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7281–7289, 2022.

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

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural trust region/proximal policy optimization attains globally optimal policy. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pp. 3242–3245. PMLR, 2021.

Mannor, S., Mansour, Y., and Tamar, A. *Reinforcement Learning: Foundations*. -, 2022. URL https://sites.google.com/view/rlfoundations/home.

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., and Basar, T. Near-optimal model-free reinforcement learning in non-stationary episodic mdps. In *International Conference on Machine Learning*, pp. 7447–7458. PMLR, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

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

Munos, R. Error bounds for approximate policy iteration. In *ICML*, volume 3, pp. 560–567. Citeseer, 2003.

Munos, R. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, pp. 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization, 1983.

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

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

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

Neu, G., Antos, A., György, A., and Szepesvári, C. Online markov decision processes under bandit feedback. *Advances in Neural Information Processing Systems*, 23, 2010.

Neu, G., Gyorgy, A., and Szepesvári, C. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pp. 805–813. PMLR, 2012.

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

Rosenberg, A. and Mansour, Y. Stochastic shortest path with adversarially changing costs. *arXiv preprint arXiv:2006.11561*, 2020.

Rosenberg, A., Cohen, A., Mansour, Y., and Kaplan, H. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pp. 8210–8219. PMLR, 2020.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Schweitzer, P. J. and Seidmann, A. Generalized polynomial approximations in markovian decision processes. *Journal of mathematical analysis and applications*, 110(2):568–582, 1985.

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.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Tewari, A. and Bartlett, P. Optimistic linear programming gives logarithmic regret for irreducible mdps. *Advances in Neural Information Processing Systems*, 20, 2007.

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pp. 22384–22429. PMLR, 2022a.

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

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. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019.

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.

Yu, J. Y., Mannor, S., and Shimkin, N. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

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, 2020a.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020b.

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

Zhan, W., Cen, S., Huang, B., Chen, Y., Lee, J. D., and Chi, Y. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *arXiv preprint arXiv:2105.11066*, 2021.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021a.

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

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

A Analysis Preliminaries

For convenience, the table below summarizes most of the notation used throughout the analysis.

$\mathbb{P}_h(\cdot\|s,a)$	The probability density function of the next state given the agent is at s and takes action a
$\mathbb{P}_h V : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$	For any function $V : \mathcal{S}\to\mathbb{R}$, defined by $\mathbb{P}_h V(s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot\|s,a)} V(s')$
$\mathbb{E}_k[\cdot]$	Expectation conditioned on past policies; $\mathbb{E}_k[\cdot] \coloneqq \mathbb{E}\left[\cdot\|\pi^1,\ldots,\pi^k\right]$
$d_h^k, d_h^\star$	State and state-action occupancy measures of $\pi^k, \pi^\star$.
$D^k = \{\mathcal{D}_h^k\}$	The dataset used to compute $\widetilde{B}^k$ and $\widehat{\Sigma}^+_{kh\gamma}$. Contains episode indices / tuples $\left(s_h^i, a_h^i, s_{h+1}^i\right)$
$\mathbf{c}_h^k \in \mathbb{R}^d$	The adversarially chosen cost vector of episode k
$\ell_h^k(s,a)$	The loss function of episode k applied to s,a; $\ell_h^k(s,a) = \phi(s,a)^\top \mathbf{c}_h^k$
$\ell_h^k \in \mathbb{R}$	Loss of the agent on episode k time h; $\ell_h^k = \ell_h^k(s_h^k, a_h^k)$ (slight notation overloading here)
$Q^{k,\pi}$	The Q function of policy π in the MDP of episode k
Q^k	The true Q function of policy π^k in the MDP of episode k
$\mathbf{q}_h^k \in \mathbb{R}^d$	The low dimensional representation of Q_h^{k,π^k}
$\widehat{\mathbf{q}}_h^k \in \mathbb{R}^d$	(nearly) unbiased estimate of $\mathbf{q}_h^k$, see Algorithm 1
$\widehat{Q}_h^k$	(nearly) unbiased estimate of Q_h^{k,π^k}; $\widehat{Q}_h^k(s,a) = \phi(s,a)^\top \widehat{\mathbf{q}}_h^k$; see Algorithm 1
b_h^k	Immediate bonus (also referred to as Q-bonus) function; see Algorithm 1
$b_h^{\mathbb{P},k}$	Dynamics bonus function, used for bonus-to-go optimism; see Algorithm 2
$B_h^{k,\pi}$	True action-value function in the bonus MDP $B_h^{k,\pi}(s,a) = Q_h^\pi(s,a;b^k)$ (aka bonus-to-go)
$W_h^{k,\pi}$	True value function in the bonus MDP; $W_h^{k,\pi} = V_h^\pi(s;b^k)$
$\widetilde{B}_h^k$	The optimistic approximation of B_h^{k,π^k} (aka bonus-to-go approximation) ; see Algorithm 2
$\widetilde{W}_h^k$	The optimistic approximation of W_h^{k,π^k}; see Algorithm 2
$\widetilde{P}_h^k \widetilde{W}_{h+1}^k : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$	The optimistic approximation of $\mathbb{P}_h \widetilde{W}_{h+1}^k$; see Algorithm 2
$\Lambda_h^k \in \mathbb{R}^{d\times d}$	Empirical non-normalized covariance of d_h^k; see Algorithm 2
$\widehat{\mathbf{w}}_h^k \in \mathbb{R}^d$	Estimate of the low dimensional representation of $\mathbb{P}_h \widetilde{W}_{h+1}^k : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ see Algorithm 2
$\Sigma_{kh} \in \mathbb{R}^{d\times d}$	Feature occupancy covariance; $\Sigma_{kh} = \mathbb{E}_{s,a\sim d_h^k}\left[\phi(s,a)\phi(s,a)^\top\right]$
$\Sigma_{kh\gamma} \in \mathbb{R}^{d\times d}$	γ-regularized feature occupancy covariance; $\Sigma_{kh\gamma} = \gamma I + \Sigma_{kh}$
$\widehat{\Sigma}^+_{kh\gamma} \in \mathbb{R}^{d\times d}$	(nearly) unbiased estimate of $\Sigma^{-1}_{kh\gamma}$, computed by Algorithm 3
λ	Regularization parameter for least squares backups in Algorithm 2, it is set to $\lambda = 1$ throughout.
γ	Regularization parameter for inverse covariance estimation, used in Algorithm 3
β	Q-bonus function factor (see Equation (9))
$\beta^{\mathbb{P}}$	Dynamics bonus function factor (see Equation (12))

Notation for conditional expectation operators. We use the convention that for any function $V : \mathcal{S}\to\mathbb{R}$, the conditional expectation operator is denoted by $\mathbb{P}_h$;

$$\mathbb{P}_h V(s,a) \coloneqq \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} V(s'). \tag{25}$$

We note the motivation for this notation comes from considering (when the state space is finite) the matrix $\mathbb{P}_h \in \mathbb{R}^{SA\times S}$ where $S = |\mathcal{S}|$, and the vector $V \in \mathbb{R}^S$. Then the result of multiplying them is indeed a vector $\mathbb{P}_h V \in \mathbb{R}^{SA}$ with $\mathbb{P}_h V(s,a) = \sum_{s'} \mathbb{P}_h(s'|s,a)V(s') = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} V(s')$. In similar spirit and with slight abuse of notation, we let $\widetilde{\mathbb{P}}_h^k \colon \mathbb{R}^S \to \mathbb{R}^{SA}$ denote an optimistic conditional expectation that *is not* a linear operator, but rather defined by;

$$\widetilde{\mathbb{P}}_h^k W(s,a) \coloneqq (\widetilde{\mathbb{P}}_h^k W)(s,a) \coloneqq \widehat{\mathbb{P}}_h^k W(s,a) + b_h^{\mathbb{P},k}(s,a), \quad \text{where } \widehat{\mathbb{P}}_h^k \coloneqq \left(\Lambda_h^k\right)^{-1} \sum_{i\in\mathcal{D}_h^k} \phi(s_h^i, a_h^i)\mathbf{e}[s_{h+1}^i]^\top,$$

where $\mathbf{e}[s]$ denotes the s'th standard basis vector in $\mathbb{R}^S$. Thus, the $\widetilde{\mathbb{P}}_h^k$ operator is composed from a linear one $\widehat{\mathbb{P}}_h^k$ plus a bonus term. The above decomposition is discussed to motivate our notation, but otherwise is not needed anywhere in our proofs as we always apply $\widetilde{\mathbb{P}}_h^k$ to $\widetilde{W}_{h+1}^k$.

Bellman consistency equations. The value and action-value functions, in any MDP, satisfy;

$$Q_h^\pi = \ell_h + \mathbb{P}_h V_{h+1}^\pi \tag{26}$$

$$V_h^\pi(s) = \left\langle \pi(\cdot|s), Q_h^\pi(s,\cdot)\right\rangle \tag{27}$$

Preliminary lemmas.

Lemma 1. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \ell)$ be any linear MDP (see Assumption 1) with $\ell_h(s,a) = \phi(s,a)^\top \mathbf{c}_h$ for cost vectors $\{\mathbf{c}_h\}_{h=1}^H \subset \mathbb{R}^d$. Then, for any policy π and time step h, there exists $\mathbf{q}_h^\pi \in \mathbb{R}^d$ such that $Q_h^\pi(s,a) = \phi(s,a)^\top \mathbf{q}_h^\pi$. Furthermore, $\|\mathbf{q}_h^\pi\| \le H\sqrt{d}$.*

Proof. Observe;

$$Q_h^\pi(s,a) = \ell_h(s,a) + \mathbb{E}\left[V_{h+1}^\pi(s_{h+1}) \mid s_h = s, a_h = a\right] = \phi(s,a)^\top \left(\mathbf{c}_h + \int \psi_h(s')V_{h+1}^\pi(s')\mathrm{d}s'\right),$$

thus the first claim follows with $\mathbf{q}_h^\pi \coloneqq \mathbf{c}_h + \int \psi_h(s')V_{h+1}^\pi(s')\mathrm{d}s'$. For the second part, note that

$$\|\mathbf{q}_h^\pi\| = \left\|\mathbf{c}_h + \int \psi_h(s')V_{h+1}^\pi(s')\mathrm{d}s'\right\| \le \sqrt{d} + \sqrt{d}\left\|V_{h+1}^\pi\right\|_\infty \le \sqrt{d} + \sqrt{d}(H-1) = H\sqrt{d},$$

where the first inequality follows by assumption (see Assumption 1). □

In what follows we will refer to the true low dimensional Q-vector on episode k time step h;

$$\mathbf{q}_h^k \coloneqq \mathbf{q}_h^{k,\pi^k} \coloneqq \mathbf{c}_h^k + \int \psi_h(s')V_{h+1}^{k,\pi^k}(s')\mathrm{d}s'. \tag{28}$$

By Lemma 1, we have that $\|\mathbf{q}_h^k\| \le H\sqrt{d}$, and

$$Q_h^{k,\pi^k}(s,a) = \phi(s,a)^\top \mathbf{q}_h^k,$$

for all s, a, h, k.

Lemma 2. *In both Algorithms 1 and 4, it holds that for all $h \in [H], k \in [K]$, conditioned on $\pi^1, \ldots, \pi^k$, we have that $\mathbf{q}_h^k$ is fixed, and that $\widehat{\Sigma}_{kh\gamma}^+$ and $\left(s_t^k, a_t^k, \ell_t^k\right)_{t=1}^H$ are independent.*

Proof. First note that a-priori $\mathbf{q}_h^k$ is a random variable determined by the adversary's choice of cost vectors on episode k, which may depend on $\pi^1, \ldots, \pi^k$. However, when conditioning on $\pi^1, \ldots, \pi^k$ the adversary's (which we assume is deterministic) is clearly fixed.

For the second part in the claim, consider first Algorithm 4, where $\widehat{\Sigma}_{kh\gamma}^+$ is computed from samples generated by the simulator. Thus it immediately follows that $\widehat{\Sigma}_{kh\gamma}^+$ and $\left(s_t^k, a_t^k, \ell_t^k\right)_{t=1}^H$ are indeed independent conditioned on π^k, for all h, k.

For Algorithm 1, let k, h, such that $k \in T_j$, and note that $\left\{\pi^1, \ldots, \pi^k\right\}$ are in fact just $\left\{\pi^{(1)}, \ldots, \pi^{(j)}\right\}$. Conditioning on π^k, all rollouts in block j are independent. In addition, transitions of episode k *are not* contained in $\mathcal{D}_h^k$ (by the two-way block partitioning Equation (8)). Thus, conditioning on $\pi^k = \pi^{(j)}$, this immediately implies $\widehat{\Sigma}_{kh\gamma}^+$ (which is computed only from samples in $\mathcal{D}_h^k$) and $\left(s_t^k, a_t^k, \ell_t^k\right)_{t=1}^H$ are indeed independent, and completes the proof. □

B Theorem Proofs

The analysis begins by considering a slightly reframed version of the regret decomposition proposed by (Luo et al., 2021);

$$
\begin{aligned}
\text{Regret} = \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\left\langle Q_h^k(s,\cdot),\pi_h^k(\cdot|s)-\pi_h^\star(\cdot|s)\right\rangle\right] \\
&= \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\left\langle Q_h^k(s,\cdot)-\widehat{Q}_h^k(s,\cdot),\pi_h^k(\cdot|s)\right\rangle\right]}_{\text{Bias1}} \\
&+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\left\langle \widehat{Q}_h^k(s,\cdot)-Q_h^k(s,\cdot),\pi_h^\star(\cdot|s)\right\rangle\right]}_{\text{Bias2}} \\
&+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\left\langle \widehat{Q}_h^k(s,\cdot)-\widetilde{B}_h^k(s,\cdot),\pi_h^k(\cdot|s)-\pi_h^\star(\cdot|s)\right\rangle\right]}_{\text{OMD}} \\
&+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\left\langle \widetilde{B}_h^k(s,\cdot),\pi_h^k(\cdot|s)-\pi_h^\star(\cdot|s)\right\rangle\right]}_{\text{Exploration}} \qquad (29)
\end{aligned}
$$

Next, we will state the relevant lemmas used to bound each of the terms, and then proceed to the main proof. All subsequent arguments hinge on peroperties of our inverse covariance estimators, which are stated in the below lemma, and proved in Appendix E.

Lemma 3 (MGR). *Let $\epsilon,\sigma,\gamma>0$ be three parameters and assume also $\sigma\le 1/4$, $\epsilon\le\sigma/6$ and that $\gamma<1/2$. Assume $\mathcal{D}$ contains MN i.i.d. samples $\{\phi\}\subset\mathbb{R}^d$, $\|\phi\|\le 1$, from some distribution p, and let $\Sigma_\gamma := \mathbb{E}_{\phi\sim p}\left[\phi\phi^\top\right]+\gamma I$. Then invoking Algorithm 3 with arguments $(\mathcal{D},M,N,\gamma)$, for $M=\frac{48d}{\gamma\sigma}\log\frac{72d}{\gamma^2\sigma}$ and $N=\frac{2}{\gamma}\log\frac{1}{\gamma\epsilon}$, we have*

$$\left\|\widehat{\Sigma}_\gamma^+\right\|\le\frac{1}{\gamma}\ \text{almost surely,} \qquad (30)$$

$$\left\|\mathbb{E}\left[\widehat{\Sigma}_\gamma^+\right]-\Sigma_\gamma^{-1}\right\|\le\epsilon, \qquad (31)$$

$$\mathbb{E}\left[\widehat{\Sigma}_\gamma^+\Sigma_\gamma\widehat{\Sigma}_\gamma^+\right]\le 2\mathbb{E}\left[\widehat{\Sigma}_\gamma^+\right]+\sigma I. \qquad (32)$$

To bound the exploration term, we intially establish confidence bounds on our approximate bouns-to-go functions.

Lemma 4 (Bonus backup confidence bounds). *Assume $\beta=2H\sqrt{\gamma d}$, $\lambda\ge 1,\gamma\ge 1/K$, $|\mathcal{D}_h^k|=\widetilde{O}((dHK)^4)$, and $\left\|\widehat{\Sigma}_{kh\gamma}^+\right\|\le 1/\gamma$ for all k,h. Then, there exists a universal constant C_1, such that for any $\delta>0$, setting $\beta^{\mathbb{P}}\ge C_1H^2d^{3/2}\log(d\beta KH/\delta)$ ensures that w.p. $\ge 1-\delta$ the following holds for all k,h,s,a;*

$$b_h^k(s,a)+\mathbb{P}_h\widetilde{W}_{h+1}^k(s,a)\le\widetilde{B}_h^k(s,a)\le b_h^k(s,a)+\mathbb{P}_h\widetilde{W}_{h+1}^k(s,a)+2b_h^{\mathbb{P},k}(s,a), \qquad (33)$$

where $\widetilde{B}_h^k,\widetilde{W}_h^k$ are defined in Equations (14) and (15).

The proof of Lemma 4 follows from uniform concentration over the class of bonus value functions explored by our algorithm. The arguments are in the spirit of those given in (Jin et al., 2020b), and is deferred to Appendix D. With the above confidence bounds in place, the exploration term bound follows from the next lemma (for proof see Appendix C.2).

Lemma 5. *Assume the backup confidence bounds Equation (33) hold with probability at least $1-\delta$, where $\delta\le$*

$(7KH^2(\beta/\sqrt{\gamma} + \beta^{\mathbb{P}}/\sqrt{\lambda}))^{-1}$. *Then expected exploration term is bounded as*

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\left\langle \widetilde{B}_h^k(s,a), \pi_h^k(\cdot|s) - \pi_h^\star(\cdot|s)\right\rangle\right]\right]$$
$$\leq 2\mathbb{E}\left[\sum_{k}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a) + b_h^k(s,a)\right]\right] - \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^\star}\left[b_h^k(s,a)\right]\right] + 1.$$

The two final important lemmas we state before turning to the proof are given next; these bound, respectively, the OMD and bias terms. We defer proofs of both to Appendix C.1.

Lemma 6 (Algorithm 1 OMD term bound)**.** *Assume that Algorithm 1 is executed with $\eta \leq \gamma/(2H)$, $\beta \leq 1/2\sqrt{\gamma}$, and $\gamma \leq 1$. Further, assume that for all k, h; $\mathbb{E}\left[\widehat{\Sigma}_{kh\gamma}^{+}\Sigma_{kh\gamma}\widehat{\Sigma}_{kh\gamma}^{+}\right] \leq 2\mathbb{E}\left[\widehat{\Sigma}_{kh\gamma}^{+}\right] + \sigma I$, and $\|\widehat{\Sigma}_{kh\gamma}^{+}\| \leq 1/\gamma$ almost surely.*

Then, we have for any s, h;

$$\mathbb{E}\left[\sum_{k=1}^{K}\left\langle \widehat{Q}_h^k(s,\cdot) - \widetilde{B}_h^k(s,\cdot), \pi_h^k(\cdot|s) - \pi_h^\star(\cdot|s)\right\rangle\right]$$
$$\leq \frac{\tau\log A}{\eta} + \frac{8\eta\beta^2H^2K}{\gamma} + \frac{2\eta H^2}{\sqrt{\gamma}}\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\pi_h^k(a|s)\,\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^{+}}\right] + 2\eta(1+\sigma)H^2K + \frac{2\tau H}{\gamma}.$$

Lemma 7 (Bias bound)**.** *Assuming $\left\|\mathbb{E}_k\left[\widehat{\Sigma}_{kh\gamma}^{+}\right] - \Sigma_{kh\gamma}^{-1}\right\| \leq \epsilon$, $\left\|\widehat{\Sigma}_{kh\gamma}^{+}\right\| \leq 1/\gamma$ for all h, k, and $\gamma \leq 1/\sqrt{d}$, we have*

$$\mathbb{E}\left[\textsc{Bias}1 + \textsc{Bias}2\right] \leq \left(\sqrt{\gamma d H^2}\right)\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\sum_{a}\left(\pi_h^k(a|s) + \pi_h^\star(a|s)\right)\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^{+}}\right]\right] + 4\epsilon H^2K.$$

B.1 Theorem 1 proof

Proof of Theorem 1. We shall use the following set of parameters; $\sigma = 1/4$, $\beta = 2H\sqrt{d\gamma}$, $\epsilon = 1/K$, $\eta = \gamma/(2H)$, $\gamma = K^{-2/7}$, and $\beta^{\mathbb{P}} = 10C_1H^2d^{3/2}\log(28C_1d\beta KH)$, where the constant C_1 is that specified by Lemma 4.

By our setting of $\tau = MN$ in the algorithm, each estimation dataset is of size $|\mathcal{D}_h^k| = \frac{48d}{\gamma\sigma}\log\frac{72d}{\gamma^2\sigma} \times \frac{2}{\gamma}\log\frac{1}{\gamma\epsilon}$. This, as well as Lemma 2 and our parameter choices imply the conditions for Lemma 3 are met, thus it follows that for all h, k Equations (30) to (32) hold for $\widehat{\Sigma}_{\gamma}^{+} = \widehat{\Sigma}_{kh\gamma}^{+}, \Sigma_\gamma = \Sigma_{kh\gamma}$.

Proceeding, we begin by bounding the bias and OMD terms of Equation (29). From Lemma 6, we immediately get that

$$\mathbb{E}\left[\text{OMD}\right] \leq \frac{\tau H\log A}{\eta} + \frac{8\eta\beta^2H^3K}{\gamma} + \frac{2\eta H^2}{\sqrt{\gamma}}\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\sum_{a}\pi_h^k(a|s)\,\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^{+}}\right]\right] + 2\eta(1+\sigma)H^3K + \frac{2\tau H^2}{\gamma}.$$

Combining the above with Lemma 7 and setting

$$\mathcal{E} := \frac{\tau H\log A}{\eta} + 4\epsilon H^2K + \frac{8\eta\beta^2H^3K}{\gamma} + 2\eta H^3K(1+\sigma) + \frac{2\tau H^2}{\gamma}, \tag{34}$$

we have,

$$
\begin{aligned}
\mathbb{E}\left[\textsc{Bias1}+\textsc{Bias2}+\textsc{OMD}\right] &\le \left(\sqrt{\gamma dH^2}\right)\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\sum_a\left(\pi_h^k(a|s)+\pi_h^\star(a|s)\right)\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^+}\right]\right] \\
&\quad + \frac{2\eta H^2}{\sqrt{\gamma}}\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\sum_a \pi_h^k(a|s)\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^+}\right]\right]+\mathcal{E} \\
&\le \left(\sqrt{\gamma dH^2}+\frac{2\eta H^2}{\sqrt{\gamma}}\right)\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\sum_a\left(\pi_h^k(a|s)+\pi_h^\star(a|s)\right)\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^+}\right]\right]+\mathcal{E} \\
&\le \beta\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\sum_a\left(\pi_h^k(a|s)+\pi_h^\star(a|s)\right)\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^+}\right]\right]+\mathcal{E},
\end{aligned}
$$

where the last inequality follows from our setting of η and β;

$$
\sqrt{\gamma dH^2}+\frac{2\eta H^2}{\sqrt{\gamma}} = \sqrt{\gamma dH^2}+\sqrt{\gamma}H \le 2H\sqrt{\gamma d}=\beta.
$$

Further, note that by our Q-bonus definition (see Equation (9)),

$$
\begin{aligned}
\beta\sum_a\left(\pi_h^k(a|s)+\pi_h^\star(a|s)\right)\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^+} &= \beta\sum_a\pi_h^\star(a|s)\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^+}+\beta\sum_a\pi_h^k(a|s)\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^+} \\
&= \beta\sum_a\pi_h^\star(a|s)\left(\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^+}+\sum_{a'}\pi_h^k(a'|s)\|\phi(s,a')\|_{\widehat{\Sigma}_{kh\gamma}^+}\right) \\
&= \sum_a\pi_h^\star(a|s)b_h^k(s,a),
\end{aligned}
$$

therefore,

$$
\mathbb{E}\left[\textsc{Bias1}+\textsc{Bias2}+\textsc{OMD}\right] \le \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d_h^\star}\left[b_h^k(s,a)\right]\right]+\mathcal{E}.
$$

Next, for the exploration term in Equation (29), first observe our choice of $\beta^{\mathbb{P}}$ is such that $\beta^{\mathbb{P}} \ge C_1H^2d^{3/2}\log(d\beta KH/\delta)$ for $\delta=(28C_1KHd)^{-9}$. In addition, our choice of parameters is such that $\delta \le \left(7KH^2(\beta/\sqrt{\gamma}+\beta^{\mathbb{P}}/\sqrt{\lambda})\right)^{-1}$, and for all h,k; $|\mathcal{D}_h^k|=\widetilde{O}(dK)$. Thus, we may invoke Lemma 4 which ensures the backup confidence bounds Equation (33) hold w.p. $\ge 1-\delta$, and by Lemma 5, this now implies that

$$
\mathbb{E}\left[\textsc{Exploration}\right] \le 2\mathbb{E}\left[\sum_k\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a)+b_h^k(s,a)\right]\right]-\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^\star}\left[b_h^k(s,a)\right]\right]+1. \tag{35}
$$

Combining the the last two displays, we obtain;

$$
\mathbb{E}\left[\text{Regret}\right] \le 2\mathbb{E}\left[\sum_k\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a)+b_h^k(s,a)\right]\right]+\mathcal{E}+1.
$$

To finish the proof, by Lemma 12, and that $|\mathcal{D}_h^k|=\tau$;

$$
\mathbb{E}\left[\sum_k\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a)\right]\right] \le \frac{20\beta^{\mathbb{P}}\sqrt{d}\log(\tau)}{\sqrt{\tau}} \lesssim \frac{H^3d^2\log(dHK)}{\sqrt{\tau}}.
$$

Combining this with the bound on the Q-bonus given by Lemma 13 and replacing $\mathcal{E}$ for its definition Equation (34), we finally get

$$\mathbb{E}\left[\text{Regret}\right] \lesssim \frac{H^3 d^2 K \log(dHK)}{\sqrt{\tau}} + \beta(\sqrt{d} + \sqrt{\epsilon})HK + \frac{\tau H \log A}{\eta} + \epsilon H^2 K + \frac{\eta \beta^2 H^3 K}{\gamma} + \eta(1+\sigma)H^3 K + \frac{\tau H^2}{\gamma}$$
$$\lesssim \gamma H^3 d^{3/2} K + \sqrt{\gamma} d H^2 K + \frac{dH^2}{\gamma^3} + \gamma d H^4 K,$$

where the second relation follows from $\sigma = 1/4, \beta = 2H\sqrt{d\gamma}, \epsilon = 1/K, \eta = \gamma/(2H)$, and $\tau \approx d/(\sigma\gamma^2)$. Balancing the two middle terms by setting $\gamma = K^{-2/7}$ leads to,

$$\mathbb{E}\left[\text{Regret}\right] \lesssim dH^2 K^{6/7} + d^{3/2} H^4 K^{5/7},$$

which concludes the proof. ◻

B.2 Theorem 2 proof

Most of the proof below follows the exact same steps as that of Theorem 1. We avoid repeating arguments that are completely identical, and refer the reader to the proof of Theorem 1 for the full details.

Proof of Theorem 2. We shall use the following parameter settings; $\eta = \gamma/(2H)$, $\sigma = 1/4, \epsilon = K^{-1}$, $\beta = 2H\sqrt{\gamma d}$, $\gamma = \frac{2}{(dK)^{2/3}}$, and $\beta^{\mathbb{P}} = 10 C_1 H^2 d^{3/2} \log\left(28 C_1 d\beta K H\right)$, where the constant C_1 is that specified by Lemma 4.

Similarly to the beginning of Theorem 1 we observe that Lemma 2, our parameter choices and the setting of τ imply the conditions for Lemma 3 are met, thus it follows that for all h, k Equations (30) to (32) hold for $\widehat{\Sigma}_\gamma^+ = \widehat{\Sigma}_{kh\gamma}^+, \Sigma_\gamma = \Sigma_{kh\gamma}$. We note that we use here slightly larger datasets $|\mathcal{D}_h^k| = \tau = d^2 MN$ than needed for Lemma 3; this is done in order to obtain sharper bounds for the dynamics estimation which enter later in the proof.

Proceeding, we combine Lemmas 7 and 9 and set

$$\mathcal{E} := \frac{H \log A}{\eta} + 4\epsilon H^2 K + \frac{8\eta\beta^2 H^3 K}{\gamma} + 2\eta(1+\sigma)H^3 K, \tag{36}$$

to obtain

$$\mathbb{E}\left[\textsc{Bias}1 + \textsc{Bias}2 + \text{OMD}\right] \le \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_h^\star}\left[b_h^k(s,a)\right]\right] + \mathcal{E}.$$

This last argument followed in exactly the same manner as in the proof of Theorem 1, with the only difference being the improved bound of the OMD term $H \log A/\eta$, that does not have τ in the numerator (and without the extra $\tau H/\gamma$ term introduced by the last block).

Next, again in the same manner of Theorem 1, we claim our choice of parameters are such that conditions of Lemma 4 are satisfied (in particular, we have for all h, k; $|\mathcal{D}_h^k| = \widetilde{O}((dHK)^4)$) with a $\delta > 0$ sufficiently small so that Lemma 5 gives;

$$\mathbb{E}\left[\textsc{Exploration}\right] \le 2\mathbb{E}\left[\sum_{k}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a) + b_h^k(s,a)\right]\right] - \mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^\star}\left[b_h^k(s,a)\right]\right] + 1. \tag{37}$$

Adding together our two bounds on the regret terms, we get

$$\mathbb{E}\left[\text{Regret}\right] \le 2\mathbb{E}\left[\sum_{k}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a) + b_h^k(s,a)\right]\right] + \mathcal{E} + 1.$$

Bounding the first term using Lemmas 11 and 13, and replacing $\mathcal{E}$ for its definition, leads to

$$\begin{aligned}
\mathbb{E}\left[\text{Regret}\right] &\leq 40H\beta^{\mathbb{P}}\frac{\sqrt{d}\log\left(2|\mathcal{D}_h^k|\right)}{|\mathcal{D}_h^k|}K + 5H\sqrt{d}\beta K + \frac{H\log A}{\eta} + 4\epsilon H^2 K + \frac{8\eta\beta^2 H^3 K}{\gamma} + 2\eta(\sigma+1)H^3K + 1 \\
&\lesssim H^3 d^{3/2}\frac{\sqrt{d}}{\sqrt{d^3/\gamma^2}}K + H\sqrt{d}\beta K + \frac{H\log A}{\eta} + \frac{\eta\beta^2 H^3 K}{\gamma} + \eta H^3 K && (\sigma, \epsilon, |D_h^k|, \beta^{\mathbb{P}}) \\
&\lesssim \sqrt{\gamma}dH^2K + \frac{H^2}{\gamma} + \gamma d H^4 K + \gamma H^2 K, && (\eta, \beta) \\
&\lesssim H^2(dK)^{2/3} + H^4(dK)^{1/3}. && (\gamma)
\end{aligned}$$

In the second relation above, we replace $\sigma = 1/4, \epsilon = 1/K, |D_h^k| = \widetilde{\Theta}(d^3/\gamma^2), \beta^{\mathbb{P}} = \widetilde{O}(H^2 d^{3/2})$, and in the third $\eta = \gamma/(2H), \beta = 2H\sqrt{\gamma d}$, simplify and absorb the first term $\gamma\sqrt{d}H^3K$ in the $\gamma d H^4 K$ term. Finally, we replace $\gamma = \frac{2}{(dK)^{2/3}}$, which completes the proof.

□

C Regret Terms Proofs

C.1 Bias and OMD Terms

Proof of Lemma 7. Recall the low dimensional representation $\mathbf{q}_h^k \in \mathbb{R}^d$ of $Q_h^k = Q_h^{k,\pi^k}$ defined in Equation (28) , and note that

$$\mathbb{E}_k\left[\sum_{t=h}^{H}\ell_t^k\right] = \mathbb{E}_k\left[Q_h^k(s_h^k, a_h^k)\right] = \mathbb{E}_k\left[\phi(s_h^k, a_h^k)^\top \mathbf{q}_h^k\right].$$

Therefore,

$$\begin{aligned}
\mathbb{E}_k\left[\widehat{\mathbf{q}}_h^k\right] &= \mathbb{E}_k\left[\widehat{\Sigma}_{kh\gamma}^{+}\phi(s_h^k, a_h^k)\phi(s_h^k, a_h^k)^\top \mathbf{q}_h^k\right] \\
&= \mathbb{E}_k\left[\widehat{\Sigma}_{kh\gamma}^{+}\right]\Sigma_{kh}\mathbf{q}_h^k && \text{(independence, Lemma 2)} \\
&= \Sigma_{kh\gamma}^{-1}\Sigma_{kh}\mathbf{q}_h^k + \left(\mathbb{E}_k\left[\widehat{\Sigma}_{kh\gamma}^{+}\right] - \Sigma_{kh\gamma}^{-1}\right)\Sigma_{kh}\mathbf{q}_h^k \\
&= \mathbf{q}_h^k - \gamma\Sigma_{kh\gamma}^{-1}\mathbf{q}_h^k + \left(\mathbb{E}_k\left[\widehat{\Sigma}_{kh\gamma}^{+}\right] - \Sigma_{kh\gamma}^{-1}\right)\Sigma_{kh}\mathbf{q}_h^k,
\end{aligned}$$

so for any s, a;

$$\mathbb{E}_k\left[\phi(s,a)^\top\widehat{\mathbf{q}}_h^k\right] = \phi(s,a)^\top\mathbf{q}_h^k - \gamma\phi(s,a)^\top\Sigma_{kh\gamma}^{-1}\mathbf{q}_h^k + \phi(s,a)^\top\left(\mathbb{E}\left[\widehat{\Sigma}_{kh\gamma}^{+}\right] - \Sigma_{kh\gamma}^{-1}\right)\Sigma_{kh}\mathbf{q}_h^k.$$

To bound the contribution of the third term above, observe that;

$$\begin{aligned}
\left\|\phi(s,a)^\top\left(\mathbb{E}\left[\widehat{\Sigma}_{kh\gamma}^{+}\right] - \Sigma_{kh\gamma}^{-1}\right)\Sigma_{kh}\mathbf{q}_h^k\right\| &\leq \left\|\mathbb{E}\left[\widehat{\Sigma}_{kh\gamma}^{+}\right] - \Sigma_{kh\gamma}^{-1}\right\|\left\|\Sigma_{kh}\mathbf{q}_h^k\right\| \\
&\leq \epsilon\left\|\Sigma_{kh}\mathbf{q}_h^k\right\| \\
&= \epsilon\left\|\mathbb{E}_{d_h^k}\left[\phi(s_h^k, a_h^k)\phi(s_h^k, a_h^k)^\top\mathbf{q}_h^k\right]\right\| \\
&\leq \epsilon H\mathbb{E}_{d_h^k}\left[\left\|\phi(s_h^k, a_h^k)\right\|\right] \\
&\leq \epsilon H.
\end{aligned}$$

Therefore, using Lemma 1;

$$\begin{aligned}
\mathbb{E}_k\left[Q_h^k(s,a) - \widehat{Q}_h^k(s,a)\right] &= \mathbb{E}_k\left[\phi(s,a)^\top\left(\mathbf{q}_h^k - \widehat{\mathbf{q}}_h^k\right)\right] \\
&\le \gamma\phi(s,a)^\top\Sigma_{kh\gamma}^{-1}\mathbf{q}_h^k + \epsilon H \\
&= \gamma\phi(s,a)^\top\mathbb{E}_k\left[\widehat{\Sigma}_{kh\gamma}^{+}\right]\mathbf{q}_h^k + \gamma\phi(s,a)^\top\left(\Sigma_{kh\gamma}^{-1} - \mathbb{E}\left[\widehat{\Sigma}_{kh\gamma}^{+}\right]\right)\mathbf{q}_h^k + \epsilon H \\
&\le \gamma\phi(s,a)^\top\mathbb{E}_k\left[\widehat{\Sigma}_{kh\gamma}^{+}\right]\mathbf{q}_h^k + \gamma\epsilon\sqrt{d}H + \epsilon H && (\|\mathbf{q}_h^k\| \le H\sqrt{d}) \\
&\le \gamma\mathbb{E}_k\left[\phi(s,a)^\top\widehat{\Sigma}_{kh\gamma}^{+}\mathbf{q}_h^k\right] + 2\epsilon H && (\gamma \le 1/\sqrt{d}) \\
&\le \gamma\mathbb{E}_k\left[\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^{+}}\|\mathbf{q}_h^k\|_{\widehat{\Sigma}_{kh\gamma}^{+}}\right] + 2\epsilon H \\
&\le \sqrt{\gamma d}H\mathbb{E}_k\left[\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^{+}}\right] + 2\epsilon H. && (\|\widehat{\Sigma}_{kh\gamma}^{+}\| \le 1/\gamma, \|\mathbf{q}_h^k\| \le H\sqrt{d})
\end{aligned}$$

Now,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^H\mathbb{E}_{s\sim d_h^\star}\left[\sum_a\pi_h^k(a|s)\left(Q_h^k(s,a) - \widehat{Q}_h^k(s,a)\right)\right]\right] \\
&= \mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^H\mathbb{E}_{s\sim d_h^\star}\left[\sum_a\pi_h^k(a|s)\mathbb{E}_k\left[Q_h^k(s,a) - \widehat{Q}_h^k(s,a)\right]\right]\right] \\
&\le \sqrt{\gamma d}H\mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^H\mathbb{E}_{s\sim d_h^\star}\left[\sum_a\pi_h^k(a|s)\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^{+}}\right]\right] + 2\epsilon H^2K.
\end{aligned}$$

The argument for BIAS2 is identical, apart from summing in the last step over probabilities given by $\pi_h^\star(a|s)$. The result follows by summing the two bounds. ◻

Lemma 8 (OMD term bound base)**.** *Assume that $\eta \le \gamma/(2H)$, $\beta \le 1/2\sqrt{\gamma}$, and $\gamma \le 1$. Further, assume that for all k,h; $\mathbb{E}\left[\widehat{\Sigma}_{kh\gamma}^{+}\Sigma_{kh\gamma}\widehat{\Sigma}_{kh\gamma}^{+}\right] \le 2\mathbb{E}\left[\widehat{\Sigma}_{kh\gamma}^{+}\right] + \sigma I$, and $\|\widehat{\Sigma}_{kh\gamma}^{+}\| \le 1/\gamma$ almost surely. Then, for both Algorithms 1 and 4, it holds that;*

$$\forall s,a;\quad \left|\widehat{Q}_h^k(s,a) - \widetilde{B}_h^k(s,a)\right| \le \frac{2H}{\gamma} \tag{38}$$

$$\begin{aligned}
\forall s,h;\quad &\mathbb{E}\left[\sum_{k=1}^K\sum_a\pi_h^k(a|s)\left(\widehat{Q}_h^k(s,a) - \widetilde{B}_h^k(s,a)\right)^2\right] \le \\
&\frac{2H^2}{\sqrt{\gamma}}\mathbb{E}\left[\sum_{k=1}^K\sum_a\pi_h^k(a|s)\|\phi(s,a)\|_{\widehat{\Sigma}_{kh\gamma}^{+}}\right] + 2(\sigma+1)H^2K + \frac{8\beta^2H^2K}{\gamma}.
\end{aligned} \tag{39}$$

Proof. Note that for any s,a, by definition, we have $\widehat{Q}_h^k(s,a) = \phi(s,a)^\top\widehat{\mathbf{q}}_h^k$ and $\left|\widetilde{B}_h^k(s,a)\right| \le B_1^{\max}$ (by the clipping in Equation (14)). Thus;

$$\begin{aligned}
\left|\widehat{Q}_h^k(s,a) - \widetilde{B}_h^k(s,a)\right| \le \left\|\phi(s,a)^\top\widehat{\mathbf{q}}_h^k\right\| + B_1^{\max} &\le \left\|\widehat{\Sigma}_{kh\gamma}^{+}\phi(s_h^k,a_h^k)\sum_{t=h}^H\ell_t^k\right\| + \frac{2\beta H}{\sqrt{\gamma}} \\
&\le H\left\|\widehat{\Sigma}_{kh\gamma}^{+}\right\| + \frac{2\beta H}{\sqrt{\gamma}} \\
&\le H\left(\frac{1}{\gamma} + \frac{2\beta}{\sqrt{\gamma}}\right) \le \frac{2H}{\gamma},
\end{aligned}$$

where the second to last and last inequalities follow from our assumptions $\|\widehat{\Sigma}_{kh\gamma}^{+}\| \le 1/\gamma$ and $\beta \le 1/(2\sqrt{\gamma})$. For the second part, observe that for all s, h;

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\pi_h^k(a|s)\left(\widehat{Q}_h^k(s,a)-\widetilde{B}_h^k(s,a)\right)^2\right] &\leq 2\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\pi_h^k(a|s)\widehat{Q}_h^k(s,a)^2\right]+2\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\pi_h^k(a|s)\widetilde{B}_h^k(s,a)^2\right]\\ &\leq 2\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\pi_h^k(a|s)\widehat{Q}_h^k(s,a)^2\right]+\frac{8\beta^2H^2K}{\gamma}, \end{aligned} \tag{40}$$

where the last transition uses again our bound on $\widetilde{B}_h^k(s,a)$. Further, for any s, a, h, k, using independence of $\widehat{\Sigma}^+_{kh\gamma}$ and $(s_h^k, a_h^k, \ell_h^k)_{h=1}^H$ conditioned on $\pi^1, \ldots, \pi^k$ (Lemma 2), we have;

$$\begin{aligned}\mathbb{E}_k\left[\widehat{Q}_h^k(s,a)^2\right] &= \mathbb{E}_k\left[\phi(s,a)^\top\widehat{\mathbf{q}}_h^k\left(\widehat{\mathbf{q}}_h^k\right)^\top\phi(s,a)\right]\\ &= \mathbb{E}_k\left[\phi(s,a)^\top\left(\widehat{\Sigma}^+_{kh\gamma}\phi(s_h^k,a_h^k)L_h^k\right)\left(\widehat{\Sigma}^+_{kh\gamma}\phi(s_h^k,a_h^k)L_h^k\right)^\top\phi(s,a)\right] && (L_h^k \coloneqq \textstyle\sum_{t=h}^H \ell_t^k)\\ &= \mathbb{E}_k\left[\left(L_h^k\right)^2\phi(s,a)^\top\left(\widehat{\Sigma}^+_{kh\gamma}\phi(s_h^k,a_h^k)\phi(s_h^k,a_h^k)^\top\widehat{\Sigma}^+_{kh\gamma}\right)\phi(s,a)\right]\\ &\leq H^2\mathbb{E}_k\left[\phi(s,a)^\top\widehat{\Sigma}^+_{kh\gamma}\phi(s_h^k,a_h^k)\phi(s_h^k,a_h^k)^\top\widehat{\Sigma}^+_{kh\gamma}\phi(s,a)\right]\\ &= H^2\mathbb{E}_k\left[\phi(s,a)^\top\widehat{\Sigma}^+_{kh\gamma}\mathbb{E}_{(s_h^k,a_h^k)\sim\text{Alg}}\left[\phi(s_h^k,a_h^k)\phi(s_h^k,a_h^k)^\top\right]\widehat{\Sigma}^+_{kh\gamma}\phi(s,a)\right] && \text{(independence)}\\ &= H^2\mathbb{E}_k\left[\phi(s,a)^\top\widehat{\Sigma}^+_{kh\gamma}\Sigma_{kh}\widehat{\Sigma}^+_{kh\gamma}\phi(s,a)\right]\\ &\leq H^2\mathbb{E}_k\left[\phi(s,a)^\top\widehat{\Sigma}^+_{kh\gamma}\left(\gamma I+\Sigma_{kh}\right)\widehat{\Sigma}^+_{kh\gamma}\phi(s,a)\right]\\ &\leq 2H^2\mathbb{E}_k\left[\phi(s,a)^\top\widehat{\Sigma}^+_{kh\gamma}\phi(s,a)\right]+\sigma H^2 && \left(\mathbb{E}\left[\widehat{\Sigma}^+_{kh\gamma}\Sigma_{kh\gamma}\widehat{\Sigma}^+_{kh\gamma}\right]\leq 2\mathbb{E}\left[\widehat{\Sigma}^+_{kh\gamma}\right]+\sigma I\right)\\ &\leq \frac{2H^2}{\sqrt{\gamma}}\mathbb{E}_k\left[\|\phi(s,a)\|_{\widehat{\Sigma}^+_{kh\gamma}}\right]+\sigma H^2. && \left(\|\phi(s,a)\|_{\widehat{\Sigma}^+_{kh\gamma}}\leq\tfrac{1}{\sqrt{\gamma}}\right)\end{aligned}$$

Now,

$$2\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\pi_h^k(a|s)\widehat{Q}_h^k(s,a)^2\right]\leq\frac{2H^2}{\sqrt{\gamma}}\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\pi_h^k(a|s)\,\|\phi(s,a)\|_{\widehat{\Sigma}^+_{kh\gamma}}\right]+2\sigma H^2K+2H^2K,$$

and the result follows by plugging the above back into Equation (40). □

Lemma 9 (Algorithm 4 OMD term bound)**.** *Assume that Algorithm 4 is executed with $\eta\leq\gamma/(2H)$, $\beta\leq 1/2\sqrt{\gamma}$, and $\gamma\leq 1$. Further, assume that for all k, h; $\mathbb{E}\left[\widehat{\Sigma}^+_{kh\gamma}\Sigma_{kh\gamma}\widehat{\Sigma}^+_{kh\gamma}\right]\leq 2\mathbb{E}\left[\widehat{\Sigma}^+_{kh\gamma}\right]+\sigma I$, and $\left\|\widehat{\Sigma}^+_{kh\gamma}\right\|\leq 1/\gamma$ almost surely. Then for any s, h, we have;*

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\left\langle\widehat{Q}_h^k(s,\cdot)-\widetilde{B}_h^k(s,\cdot),\pi_h^k(\cdot|s)-\pi_h^\star(\cdot|s)\right\rangle\right]\\ &\leq\frac{\log A}{\eta}+\frac{8\eta\beta^2H^2K}{\gamma}+\frac{2\eta H^2}{\sqrt{\gamma}}\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\pi_h^k(a|s)\,\|\phi(s,a)\|_{\widehat{\Sigma}^+_{kh\gamma}}\right]+2\eta H^2K(1+\sigma).\end{aligned}$$

Proof. By Equation (38) of Lemma 8 and our condition of $\eta\leq\frac{\gamma}{2H}$, we may apply the OMD bound Lemma 27, which gives for all s, h;

$$\mathbb{E}\left[\sum_{k=1}^{K}\left\langle\widehat{Q}_h^k(s,\cdot)-\widetilde{B}_h^k(s,\cdot),\pi_h^k(\cdot|s)-\pi_h^\star(\cdot|s)\right\rangle\right]\leq\frac{\log A}{\eta}+\eta\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\pi_h^k(a|s)\left(\widehat{Q}_h^k(s,a)-\widetilde{B}_h^k(s,a)\right)^2\right].$$

The result now follows by bounding the secon term above with Equation (39) given by Lemma 8. □

Next, we give the proof of Lemma 6 that combines the blocking OMD regret bound Lemma 28 with Lemma 8.

Proof of Lemma 6. By Equation (38) of Lemma 8 and our assumption that $\eta \le \frac{\gamma}{2H}$, the conditions for blocking OMD regret bound Lemma 28 are met. Thus, for all s, h;

$$\mathbb{E}\left[\sum_{k=1}^{K}\left\langle \widehat{Q}_h^k(s,\cdot) - \widetilde{B}_h^k(s,\cdot), \pi_h^k(\cdot|s) - \pi_h^\star(\cdot|s)\right\rangle\right] \le \frac{\tau\log A}{\eta} + \frac{2\tau H}{\gamma} + \eta\mathbb{E}\left[\sum_{k=1}^{K}\sum_a \pi_h^k(a|s)\left(\widehat{Q}_h^k(s,a) - \widetilde{B}_h^k(s,a)\right)^2\right].$$

The result now follows by bounding the second term above with Equation (39) given by Lemma 8. □

C.2 Exploration Terms

Proof of Lemma 5. By our assumption and Lemma 10, the random variable

$$\begin{aligned} Z := &-\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\left\langle \widetilde{B}_h^k(s,a), \pi_h^k(\cdot|s) - \pi_h^\star(\cdot|s)\right\rangle\right] + 2\sum_{k}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a) + b_h^k(s,a)\right] \\ &-\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^\star}\left[b_h^k(s,a)\right]. \end{aligned}$$

is non-negative w.p. $\ge 1-\delta$. In addition, it is not hard to verify that

$$|Z| \le 2KH(2\beta H/\sqrt{\gamma}) + 2KH(\beta/\sqrt{\gamma} + \beta^{\mathbb{P}}/\sqrt{\lambda}) + KH\beta/\sqrt{\gamma} \le 7KH^2(\beta/\sqrt{\gamma} + \beta^{\mathbb{P}}/\sqrt{\lambda}) \le \delta^{-1}.$$

Thus, Z is supported on $[-D, D]$ for $D := 7KH^2(\beta/\sqrt{\gamma} + \beta^{\mathbb{P}}/\sqrt{\lambda})$, which implies

$$\mathbb{E}Z \ge -\delta D = -\delta 7KH^2(\beta/\sqrt{\gamma} + \beta^{\mathbb{P}}/\sqrt{\lambda}) \ge -1,$$

which completes the proof after rearranging the terms. □

The next lemma is partially implicit in (Luo et al., 2021) Lemma B.1, but extends it to incorporate the affect of the bonus-to-go approximations. In addition, we provide a simpler argument owed to the removal of the dilation term, and by letting the extended value difference Lemma 26 handle most of the technicalities.

Lemma 10. *Assume that the approximate bonus-to-go functions $\widetilde{B}_{h+1}^k : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ computed by the algorithm satisfy for all s, a, h, k;*

$$b_h^k(s,a) + \mathbb{P}_h\widetilde{W}_{h+1}^k(s,a) \le \widetilde{B}_h^k(s,a) \le b_h^k(s,a) + \mathbb{P}_h\widetilde{W}_{h+1}^k(s,a) + 2b_h^{\mathbb{P},k}(s,a)$$

Then the exploration term is bounded as

$$\begin{aligned} &\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\left\langle \widetilde{B}_h^k(s,a), \pi_h^k(\cdot|s) - \pi_h^\star(\cdot|s)\right\rangle\right] \\ &\qquad \le 2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a) + b_h^k(s,a)\right] - \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^\star}\left[b_h^k(s,a)\right]. \end{aligned}$$

Proof. By assumption, for any s, a, h, k; $0 \le \widetilde{B}_h^k(s,a) - b_h^k(s,a) - \mathbb{P}_h\widetilde{W}_{h+1}^k(s,a)$, thus,

$$\begin{aligned} &\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\left\langle \widetilde{B}_h^k(s,\cdot), \pi_h^k(\cdot|s) - \pi_h^\star(\cdot|s)\right\rangle\right] \\ &\quad\le \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^\star}\left[\left\langle \widetilde{B}_h^k(s,a), \pi_h^k(\cdot|s) - \pi_h^\star(\cdot|s)\right\rangle\right] + \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s,a\sim d_h^\star}\left[\widetilde{B}_h^k(s,a) - b_h^k(s,a) - \mathbb{P}_h\widetilde{W}_{h+1}^k(s,a)\right] \\ &\quad= \sum_{k=1}^{K}\widetilde{W}_1^k - W_1^{k,\pi^\star}, \end{aligned}$$

where the equality follows from the extended value difference Lemma 26 with $\widehat{V}_1^{\pi} = \widetilde{W}_1^k$ and $V_1^{\pi'} = W_1^{k,\pi^\star}$ (and we recall definitions in Equations (11) and (15)). Further, again by Lemma 26 and our upper bound on $\widetilde{B}_h^k$;

$$\begin{aligned}\sum_{k=1}^{K} \widetilde{W}_1^k - W_1^{k,\pi^k} &= \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^k}\left[\widetilde{B}_h^k(s,a) - b_h^k(s,a) - \mathbb{P}_h \widetilde{W}_{h+1}^k(s,a)\right] \\ &\le 2\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a)\right].\end{aligned}$$

In addition, by definition of the true bonus value functions,

$$\sum_{k=1}^{K} W_1^{k,\pi^k} - W_1^{k,\pi^\star} = \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^k}\left[b_h^k(s,a)\right] - \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^\star}\left[b_h^k(s,a)\right],$$

thus we see that,

$$\begin{aligned}\sum_{k=1}^{K} \widetilde{W}_1^k - W_1^{k,\pi^\star} &= \sum_{k=1}^{K} \widetilde{W}_1^k - W_1^{k,\pi^k} + \sum_{k=1}^{K} W_1^{k,\pi^k} - W_1^{k,\pi^\star} \\ &\le 2\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a)\right] + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^k}\left[b_h^k(s,a)\right] - \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^\star}\left[b_h^k(s,a)\right] \\ &\le 2\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a) + b_h^k(s,a)\right] - \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim d_h^\star}\left[b_h^k(s,a)\right],\end{aligned}$$

which completes the proof. □

C.3 Bonus Terms

Lemma 11. *The dynamics bonus functions $b_h^{\mathbb{P},k}$ samples in $\mathcal{D}_h^k$, satisfy for all episodes k and all time steps h;*

$$\mathbb{E}\left[\mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a)\right]\right] \le \frac{10\beta^{\mathbb{P}}\sqrt{d}\log\left(2|\mathcal{D}_h^k|\right)}{\sqrt{|\mathcal{D}_h^k|}}.$$

Proof. Follows immediately by Lemma 12 with $\delta = |\mathcal{D}_h^k|^{-2}$, and noting that $|b_h^{\mathbb{P},k}(s,a)| \le \beta^{\mathbb{P}}$ almost surely. □

Lemma 12. *Assume $\lambda \ge 1$, and let h, k. For all $\delta > 0$, we have that the following holds w.p. $\ge 1-\delta$:*

$$\mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a)\right] \le \frac{5\beta^{\mathbb{P}}\sqrt{d}\log\left(2|\mathcal{D}_h^k|/\delta\right)}{\sqrt{|\mathcal{D}_h^k|}}.$$

Proof. Let $N := |\mathcal{D}_h^k|$, and observe;

$$\begin{aligned}\mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a)\right] &= \beta^{\mathbb{P}}\mathbb{E}_{\tilde{s}_h,\tilde{a}_h\sim d_h^k}\left[\|\phi(\tilde{s},\tilde{a})\|_{(\Lambda_h^k)^{-1}}\right] \\ &= \frac{\beta^{\mathbb{P}}}{N}\mathbb{E}_{(\tilde{s}_h^1,\tilde{a}_h^1),\ldots,(\tilde{s}_h^N,\tilde{a}_h^N)\sim d_h^k}\left[\sum_{i=1}^{N}\left\|\phi(\tilde{s}_h^i,\tilde{a}_h^i)\right\|_{(\Lambda_h^k)^{-1}}\right].\end{aligned} \tag{41}$$

Further, let $\Lambda_h^{k,i} = \lambda I + \sum_{t=1}^{i-1}\phi(s_h^t,a_h^t)\phi(s_h^t,a_h^t)^\top$ for some arbitrary ordering $(s_h^i,a_h^i)_{i=1}^N$ of the elements in $\mathcal{D}_h^k$. Then,

$$\mathbb{E}_{d_h^k}\left[\sum_{i=1}^{N}\left\|\phi(\tilde{s}_h^i,\tilde{a}_h^i)\right\|_{(\Lambda_h^k)^{-1}}\right] \le \mathbb{E}_{d_h^k}\left[\sum_{i=1}^{N}\left\|\phi(\tilde{s}_h^i,\tilde{a}_h^i)\right\|_{\left(\Lambda_h^{k,i}\right)^{-1}}\right].$$

Now, by Lemma 24 with $X_i := \left\|\phi(\tilde{s}_h^i, \tilde{a}_h^i)\right\|_{\left(\Lambda_h^{k,i}\right)^{-1}}$;

$$\begin{aligned}
\mathbb{E}_{d_h^k}\left[\sum_{i=1}^N \left\|\phi(\tilde{s}_h^i, \tilde{a}_h^i)\right\|_{\left(\Lambda_h^{k,i}\right)^{-1}}\right] &\le 2\sum_{i=1}^N \left\|\phi(s_h^i, a_h^i)\right\|_{\left(\Lambda_h^{k,i}\right)^{-1}} + \frac{4}{\sqrt{\lambda}}\log\frac{2N}{\delta} \\
&\le 2\sum_{i=1}^N \left\|\phi(s_h^i, a_h^i)\right\|_{\left(\Lambda_h^{k,i}\right)^{-1}} + 4\log\frac{2N}{\delta} \\
&\le 2\sqrt{N\sum_{i=1}^N \left\|\phi(s_h^i, a_h^i)\right\|^2_{\left(\Lambda_h^{k,i}\right)^{-1}}} + 4\log\frac{2N}{\delta}.
\end{aligned}$$

By Lemma 25, we can further bound this by

$$2\sqrt{2Nd\log\left(1+\frac{N}{d\lambda}\right)} + 4\log\frac{2N}{\delta} \le 5\sqrt{Nd}\log\frac{2N}{\delta}.$$

Combining the derived inequality with Equation (41), we get

$$\mathbb{E}_{s,a\sim d_h^k}\left[b_h^{\mathbb{P},k}(s,a)\right] \le \frac{5\beta^{\mathbb{P}}\sqrt{d}\log(2N/\delta)}{\sqrt{N}},$$

which completes the proof. □

Lemma 13. *Assuming* $\left\|\mathbb{E}_k\left[\widehat{\Sigma}^+_{kh\gamma}\right] - \Sigma^{-1}_{kh\gamma}\right\| \le \epsilon$, *it holds that*

$$\mathbb{E}_k\left[\mathbb{E}_{s,a\sim d_h^k}\left[b_h^k(s,a)\right]\right] \le 2\beta(\sqrt{d}+\sqrt{\epsilon})$$

Proof. Note that

$$\begin{aligned}
\mathbb{E}_{s,a\sim d_h^k}\left[b_h^k(s,a)\right] &= \beta\mathbb{E}_{s,a\sim d_h^k}\left[\|\phi(s,a)\|_{\widehat{\Sigma}^+_{kh\gamma}} + \sum_a \pi_h^k(a'|s)\,\|\phi(s,a')\|_{\widehat{\Sigma}^+_{kh\gamma}}\right] \\
&= 2\beta\mathbb{E}_{s,a\sim d_h^k}\left[\|\phi(s,a)\|_{\widehat{\Sigma}^+_{kh\gamma}}\right]
\end{aligned} \tag{42}$$

Further, for any s, a,

$$\begin{aligned}
&\mathbb{E}_k\left[\|\phi(s,a)\|_{\widehat{\Sigma}^+_{kh\gamma}}\right] \\
&\quad= \mathbb{E}_k\left[\sqrt{\phi(s,a)^\top\Sigma^{-1}_{kh\gamma}\phi(s,a) + \phi(s,a)^\top\left(\widehat{\Sigma}^+_{kh\gamma} - \Sigma^{-1}_{kh\gamma}\right)\phi(s,a)}\right] \\
&\quad\le \sqrt{\phi(s,a)^\top\Sigma^{-1}_{kh\gamma}\phi(s,a) + \phi(s,a)^\top\left(\mathbb{E}_k\left[\widehat{\Sigma}^+_{kh\gamma}\right] - \Sigma^{-1}_{kh\gamma}\right)\phi(s,a)} && \text{(Jensen's inequality)} \\
&\quad\le \sqrt{\phi(s,a)^\top\Sigma^{-1}_{kh\gamma}\phi(s,a)} + \sqrt{\phi(s,a)^\top\left(\mathbb{E}_k\left[\widehat{\Sigma}^+_{kh\gamma}\right] - \Sigma^{-1}_{kh\gamma}\right)\phi(s,a)} \\
&\quad\le \sqrt{\phi(s,a)^\top\Sigma^{-1}_{kh\gamma}\phi(s,a)} + \sqrt{\left\|\mathbb{E}_k\left[\widehat{\Sigma}^+_{kh\gamma}\right] - \Sigma^{-1}_{kh\gamma}\right\|_{\text{op}}} \\
&\quad\le \sqrt{\phi(s,a)^\top\Sigma^{-1}_{kh\gamma}\phi(s,a)} + \sqrt{\epsilon}.
\end{aligned}$$

Now, conditioning on π^k, we have;

$$\begin{aligned}
\mathbb{E}_k\left[\mathbb{E}_{s,a\sim d_h^k}\left[b_h^k(s,a)\right]\right] &= 2\beta\mathbb{E}_k\left[\mathbb{E}_{s,a\sim d_h^k}\left[\|\phi(s,a)\|_{\widehat{\Sigma}^+_{kh\gamma}}\right]\right] && \text{(Equation (42))}\\
&= 2\beta\mathbb{E}_{s,a\sim d_h^k}\left[\mathbb{E}_k\left[\|\phi(s,a)\|_{\widehat{\Sigma}^+_{kh\gamma}}\right]\right] && (d_h^k \perp \widehat{\Sigma}^+_{kh\gamma} \mid \pi^k)\\
&\le 2\beta\mathbb{E}_{s,a\sim d_h^k}\left[\sqrt{\phi(s,a)^\top\Sigma^{-1}_{kh\gamma}\phi(s,a)}\right] + 2\beta\sqrt{\epsilon} && \text{(previous inequality)}\\
&\le 2\beta\sqrt{\mathbb{E}_{s,a\sim d_h^k}\left[\phi(s,a)^\top\Sigma^{-1}_{kh\gamma}\phi(s,a)\right]} + 2\beta\sqrt{\epsilon} && \text{(Jensen)}\\
&\le 2\beta\sqrt{\mathbb{E}_{s,a\sim d_h^k}\left[\phi(s,a)^\top\Sigma^{-1}_{kh}\phi(s,a)\right]} + 2\beta\sqrt{\epsilon} && (\Sigma^{-1}_{kh\gamma} \preceq \Sigma^{-1}_{kh})\\
&= 2\beta\sqrt{\mathbb{E}_{s,a\sim d_h^k}\left[\mathrm{tr}\left(\Sigma^{-1}_{kh}\phi(s,a)\phi(s,a)^\top\right)\right]} + 2\beta\sqrt{\epsilon}\\
&= 2\beta\sqrt{\mathrm{tr}\left(\Sigma^{-1}_{kh}\mathbb{E}_{s,a\sim d_h^k}\left[\phi(s,a)\phi(s,a)^\top\right]\right)} + 2\beta\sqrt{\epsilon}\\
&= 2\beta\sqrt{\mathrm{tr}\left(\Sigma^{-1}_{kh}\Sigma_{kh}\right)} + 2\beta\sqrt{\epsilon}\\
&= 2\beta\left(\sqrt{d}+\sqrt{\epsilon}\right),
\end{aligned}$$

which completes the proof. □

D Approximate bonus-to-go confidence bounds

In this section, we establish optimism / bonus-bias confidence bounds on our approximate bonus action-value functions (aka bonus-to-go). These follow from uniform concentration over the estimated bonus value function backup operator which is computed by the least squares regression procedure in Algorithm 2. The arguments given here, at a conceptual level, follow those of (Jin et al., 2020b).

Bonus value functions explored by the algorithm. Define

$$\begin{aligned}
&B(s,a;\beta,\Sigma^+,\beta^{\mathbb{P}},\Lambda,w,B_{\max},\pi)\\
&\quad= \mathrm{clip}\left[\beta\Big(\|\phi(s,a)\|_{\Sigma^+} + \sum_a \pi(a'|s)\,\|\phi(s,a')\|_{\Sigma^+}\Big) + \phi(s,a)^\top w + \beta^{\mathbb{P}}\,\|\phi(s,a)\|_{\Lambda^{-1}}\right]_0^{B_{\max}}
\end{aligned}$$

and

$$\begin{aligned}
&\mathcal{B}(\beta,\lambda_{\Sigma^+},\beta^{\mathbb{P}},\lambda_\Lambda,L,B_{\max},\pi) && (43)\\
&\quad= \left\{B(s,a;\beta,\Sigma,\beta^{\mathbb{P}},\Lambda,w,\pi) \mid \lambda_{\max}(\Sigma^+)\le\lambda_{\Sigma^+},\lambda_{\min}(\Lambda)\ge\lambda_\Lambda,\|w\|\le L\right\}\\
&\mathcal{W}(\beta,\lambda_{\Sigma^+},\beta^{\mathbb{P}},\lambda_\Lambda,L,B_{\max},\pi) && (44)\\
&\quad= \left\{W\colon \mathcal{S}\to\mathbb{R}; W(s)=\langle\pi(\cdot|s),B(s,\cdot)\rangle \mid B\in\mathcal{B}(\beta,\lambda_{\Sigma^+},\beta^{\mathbb{P}},\lambda_\Lambda,L,\pi)\right\}.
\end{aligned}$$

We note that with appropriate parameter choices, $\widetilde{B}_h^k \in \mathcal{B}$ and $\widetilde{W}_h^k \in \mathcal{W}$ for $\widetilde{B}_h^k, \widetilde{W}_h^k$ computed by Algorithm 2 and defined in Equations (14) and (15). This will be made rigorous in the proof of Lemma 4 below.

Proof of Lemma 4. By Lemma 16, our choice of β, $\lambda \ge 1$, and that $|\mathcal{D}_h^k| = \widetilde{O}((HdK)^4)$, we have $\left\|\widehat{\mathbf{w}}_h^k\right\| \le aH^5d^4K^4\log(dK)$ for some constant a. Further, again by our choice of β, $2\beta H/\sqrt{\gamma} = 4H^2\sqrt{d}$, thus by algorithm definition and our assumptions, it is readily verified that;

$$\widetilde{W}_{h+1}^k \in \mathcal{W} := \mathcal{W}(\beta, 1/\gamma, \beta^{\mathbb{P}}, \lambda, \boldsymbol{L} = aH^5d^4K^4\log(dK), \boldsymbol{B}_{\mathbf{max}} = 4H^2\sqrt{d}, \pi^k)$$

Hence, by Lemma 19, there exist $c > 0$ such that for any $\epsilon > 0$,

$$\log\mathcal{N}_\epsilon(\mathcal{W}) \le cd^2\log\left(\frac{4\beta^{\mathbb{P}}\beta HKd}{\gamma\lambda\epsilon}\right) \le c_{\mathrm{cov}}d^2\log\left(\frac{d\beta^{\mathbb{P}}\beta HK}{\epsilon}\right),$$

where the second inequality follows from our assumptions that $\gamma \geq 1/K$, $\lambda \geq 1$ and the appropriate choice of constant c_{cov}. Thus we may apply Lemma 14, to obtain that for the constant C specified by the lemma, with $\beta^{\mathbb{P}} \geq 8CH^2 d^{3/2} \log(d\beta KH/\delta) \geq C(4H^2\sqrt{d})d \log(d\beta K^2 H/\delta)$, we have w.p. $\geq 1-\delta$ that for all s, a, h, k;

$$\left|\phi(s,a)^\top \widehat{\mathbf{w}}_h^k - \mathbb{P}_h \widetilde{W}_{h+1}^k(s,a)\right| \leq \beta^{\mathbb{P}} \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} = b_h^{\mathbb{P},k}(s,a). \tag{45}$$

This establishes that $0 \leq (\widetilde{\mathbb{P}}_h^k - \mathbb{P}_h)\widetilde{W}_{h+1}^k(s,a) \leq 2b_h^{\mathbb{P},k}(s,a)$ holds for all s, a, h, k, leaving us only with the task to verify the truncations defined in Equation (14) do not interfere with the desired conclusion. First, we show that;

$$\widetilde{B}_h^k(s,a) \leq b_h^k(s,a) + \mathbb{P}_h \widetilde{W}_{h+1}^k(s,a) + 2b_h^{\mathbb{P},k}(s,a). \tag{46}$$

Indeed, by definition Equation (14);

$$\begin{aligned}\widetilde{B}_h^k(s,a) &= \operatorname{clip}\left[b_h^k(s,a) + \widetilde{\mathbb{P}}_h^k \widetilde{W}_{h+1}^k(s,a)\right]_0^{2\beta(H-h+1)/\sqrt{\gamma}} \\ &\leq \operatorname{clip}\left[b_h^k(s,a) + \widetilde{\mathbb{P}}_h^k \widetilde{W}_{h+1}^k(s,a)\right]_0^{\infty},\end{aligned}$$

and when $\widetilde{B}_h^k(s,a) = 0$, Equation (46) holds trivially as all RHS terms are non-negative. Otherwise,

$$\begin{aligned}\widetilde{B}_h^k(s,a) \leq b_h^k(s,a) + \widetilde{\mathbb{P}}_h^k \widetilde{W}_{h+1}^k(s,a) &= b_h^k(s,a) + \phi(s,a)^\top \widehat{\mathbf{w}}_h^k + b_h^{\mathbb{P},k}(s,a) && \text{(def. in Equation (13))} \\ &\leq b_h^k(s,a) + \mathbb{P}_h \widetilde{W}_{h+1}^k(s,a) + 2b_h^{\mathbb{P},k}(s,a). && \text{(Equation (45))}\end{aligned}$$

Next, to verify

$$\widetilde{B}_h^k(s,a) \geq b_h^k(s,a) + \mathbb{P}_h \widetilde{W}_{h+1}^k(s,a), \tag{47}$$

note that

$$b_h^k(s,a) + \mathbb{P}_h \widetilde{W}_{h+1}^k(s,a) \leq \frac{2\beta}{\sqrt{\gamma}} + \frac{2\beta(H-h)}{\sqrt{\gamma}} = \frac{2\beta(H-h+1)}{\sqrt{\gamma}}.$$

Thus, when $\widetilde{B}_h^k(s,a) = 2\beta(H-h+1)/\sqrt{\gamma}$, Equation (47) holds trivially. Otherwise,

$$\begin{aligned}\widetilde{B}_h^k(s,a) &= \operatorname{clip}\left[b_h^k(s,a) + \widetilde{\mathbb{P}}_h^k \widetilde{W}_{h+1}^k(s,a)\right]_0^{2\beta(H-h+1)/\sqrt{\gamma}} \\ &= \operatorname{clip}\left[b_h^k(s,a) + \widetilde{\mathbb{P}}_h^k \widetilde{W}_{h+1}^k(s,a)\right]_0^{\infty} \\ &\geq b_h^k(s,a) + \widetilde{\mathbb{P}}_h^k \widetilde{W}_{h+1}^k(s,a) \\ &\geq b_h^k(s,a) + \phi(s,a)^\top \widehat{\mathbf{w}}_h^k + b_h^{\mathbb{P},k}(s,a) \\ &\geq b_h^k(s,a) + \mathbb{P}_h \widetilde{W}_{h+1}^k(s,a), && \text{(Equation (45))}\end{aligned}$$

which completes the proof. □

Lemma 14 (Approximate backup operator error bound)**.** *Let $\mathcal{D}_h^k$ be the dataset used for episode k of size $\widetilde{O}((dHK)^4)$, and $\Lambda_h^k = \lambda I + \sum_{i \in \mathcal{D}_h^k} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top$, with $\lambda \geq 1$. Further, let $\mathcal{V}$ be a function class with $\log \mathcal{N}_\epsilon(\mathcal{V}) \leq c_{\text{cov}} d^2 \log\left(\frac{d\beta\beta^{\mathbb{P}}K}{\epsilon}\right)$ for any $\epsilon > 0$, and $\|f\|_\infty \leq B_{\max}$ for all $f \in \mathcal{V}$. Then there exists a constant $C > 0$ depending only on c_{cov}, such that letting*

$$\beta^{\mathbb{P}} \geq CB_{\max} d \log\left(\frac{d\beta KH}{\delta}\right),$$

ensures that with probability $\geq 1-\delta$ it holds that for all $f \in \mathcal{V}$ and all s, a, h, k;

$$\left|\phi(s,a)^\top \widehat{w}_f - \mathbb{P}_h f(s,a)\right| \leq \beta^{\mathbb{P}} \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}},$$

where $\widehat{w}_f = \left(\Lambda_h^k\right)^{-1} \sum_{i \in \mathcal{D}_h^k} \phi(s_h^i, a_h^i) f(s_{h+1}^i)$.

Proof. Fix k, h, and define $w_f^\star$ by

$$\mathbb{P}_h f(s,a) = \phi(s,a)^\top \int \psi_h(s')f(s')\mathrm{d}s' := \phi(s,a)^\top w_f^\star.$$

Note that by normalization assumptions in Assumption 1, we have that $\left\|w_f^\star\right\| \le \sqrt{d}B_{\max}$, thus, by Lemma 17;

$$\left\|\widehat{w}_f - w_f^\star\right\|_{\Lambda_h^k} \le \left\|\sum_{i\in\mathcal{D}_h^k} \phi(s_h^i, a_h^i)\left(f(s_{h+1}^i) - \phi(s_h^i,a_h^i)^\top w_f^\star\right)\right\|_{(\Lambda_h^k)^{-1}} + \sqrt{\lambda d}B_{\max}. \tag{48}$$

In addition, by Lemma 20, we have that w.p. $\ge 1-p$;

$$\begin{aligned}
&\left\|\sum_{i\in\mathcal{D}_h^k} \phi(s_h^i, a_h^i)\left(f(s_{h+1}^i) - \phi(s_h^i,a_h^i)^\top w_f^\star\right)\right\|^2_{(\Lambda_h^k)^{-1}} \\
&\le 4B_{\max}^2\left(\frac{d}{2}\log\left(\frac{|\mathcal{D}_h^k|+\lambda}{\lambda}\right) + \log\frac{\mathcal{N}_{\epsilon_{\mathrm{cov}}}(\mathcal{V})}{p}\right) + \frac{8|\mathcal{D}_h^k|^2\epsilon^2}{\lambda}, \\
&\le 2B_{\max}^2 d\log\left(\frac{|\mathcal{D}_h^k|+\lambda}{\lambda}\right) + 4c_{\mathrm{cov}}B_{\max}^2 d^2\log\left(\frac{d\beta\beta^{\mathbb{P}}K}{\epsilon_{\mathrm{cov}}p}\right) + \frac{8|\mathcal{D}_h^k|^2\epsilon^2}{\lambda} \\
&\le c(\epsilon_{\mathrm{cov}}|\mathcal{D}_h^k|B_{\max}d)^2\log\left(\frac{d\beta\beta^{\mathbb{P}}K}{\epsilon_{\mathrm{cov}}p}\right),
\end{aligned}$$

for some constant $c \ge 1$ that depends only on c_{cov}. Now, using that $|\mathcal{D}_h^k| = \widetilde{O}((dHK)^4)$, with an appropriate choice of ϵ_{cov} and we can further bound the last display by

$$c'(B_{\max}d)^2\log\left(\frac{d\beta\beta^{\mathbb{P}}K}{p}\right),$$

where c' is another constant ≥ 1. Combining this with Equation (48), we get that w.p. $1-p$;

$$\left\|\widehat{w}_f - w_f^\star\right\|_{\Lambda_h^k} \le 2c'B_{\max}d\sqrt{\log\left(\frac{d\beta\beta^{\mathbb{P}}K}{p}\right)}.$$

By the union bound over k, h, choosing $\delta = p/(KH)$, we have that w.p. $1-\delta$, it holds that for all k, h;

$$\left\|\widehat{w}_f - w_f^\star\right\|_{\Lambda_h^k} \le 4c'B_{\max}d\sqrt{\log\left(\frac{d\beta\beta^{\mathbb{P}}LKH}{\delta}\right)}.$$

Now, by Lemma 15, setting

$$\beta^{\mathbb{P}} = 8c'B_{\max}d\log\left(\frac{d\beta KH}{\delta}\right) \ge 4c'B_{\max}d\log\left(\frac{d\beta KH}{\delta}\times 4c'B_{\max}d\right)$$

ensures that $\left\|\widehat{w}_f - w_f^\star\right\|_{\Lambda_h^k} \le \beta^{\mathbb{P}}$. Finally, observe that for all s, a;

$$\left|\phi(s,a)^\top\widehat{w}_f - \mathbb{P}_h f(s,a)\right| = \left|\phi(s,a)^\top\left(\widehat{w}_f - w_f^\star\right)\right| \le \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\left\|\widehat{w}_f - w_f^\star\right\|_{\Lambda_h^k} \le \beta^{\mathbb{P}}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}},$$

which complete the proof. □

Lemma 15. *Let $R, z \ge 1$, and $x \ge 2z\log(Rz)$. Then $z\log(Rx) \le x$.*

Proof. If $x = 2z\log(Rz)$;

$$\begin{aligned} z\log(Rx) &= z\log R + z\log(2z\log(Rz)) \\ &= z\log R + z\log(2z) + z\log\log(Rz) \\ &\le z\log R + z\log z + z\log(Rz) \\ &= 2z\log R + 2z\log z \\ &= x. \end{aligned}$$

For larger values, the result follows by noting $x - z\sqrt{\log(Rx)}$ is monotonically increasing in x for all $x \ge z$. □

The next lemma bounds the norm of the weights $\widehat{\mathbf{w}}_h^k$ computed in the OLSPE algorithm. We note a tighter bound can be shown, as in (Jin et al., 2020b) Lemma B.2, but the simpler argument below is sufficient for our purposes.

Lemma 16. *For all $k \in [K], h \in [H]$, assuming running* OLSPE *(Algorithm 2) with dataset $\mathcal{D}_h^k$, we have $\|\widehat{\mathbf{w}}_h^k\| \le 2\beta H|\mathcal{D}_h^k|/\sqrt{\gamma\lambda}$.*

Proof. We have;

$$\begin{aligned} \|\widehat{\mathbf{w}}_h^k\| &= \left\| \left(\Lambda_h^k\right)^{-1} \sum_{i\in\mathcal{D}_h^k} \phi(s_h^i, a_h^i)\widetilde{W}_{h+1}^k(s_{h+1}^i) \right\| \\ &\le (2\beta H/\sqrt{\gamma}) \left\|\left(\Lambda_h^k\right)^{-1}\right\| \left\| \sum_{s_h,a_h\in\mathcal{D}_h^k} \phi(s_h,a_h)\right\| \le \frac{2\beta H|\mathcal{D}_h^k|}{\sqrt{\gamma\lambda}}, \end{aligned}$$

where the first inequality follows from $\|\widetilde{W}_{h+1}^k\|_\infty \le \|\widetilde{B}_{h+1}^k\|_\infty \le 2\beta H/\sqrt{\gamma}$, as per definition Equation (14). □

Lemma 17. *Let $\{\phi_i\}_{i=1}^n \in \mathbb{R}^d$, $\{y_i\}_{i=1}^n \in \mathbb{R}$, $\lambda \in \mathbb{R}$, and set $\Lambda := \sum_{i=1}^N \phi_i\phi_i^\top + \lambda I$, and $\widehat{w} = \Lambda^{-1}\sum_{i=1}^N \phi_i y_i$. Then*

$$\|\widehat{w} - w^\star\|_\Lambda \le \left\| \sum_{i=1}^N \phi_i\left(y_i - \phi^\top w^\star\right)\right\|_{\Lambda^{-1}} + \sqrt{\lambda}\,\|w^\star\|$$

Proof. We have

$$\widehat{w} - w^\star = \Lambda^{-1}\sum_{i=1}^N \phi_i y_i - \Lambda^{-1}\left(\sum_{i=1}^N \phi_i\phi_i^\top + \lambda I\right)w^\star = \Lambda^{-1}\sum_{i=1}^N \phi_i\left(y_i - \phi^\top w^\star\right) + \lambda\Lambda^{-1}w^\star,$$

which implies

$$\|\widehat{w} - w^\star\|_\Lambda \le \left\| \sum_{i=1}^N \phi_i\left(y_i - \phi^\top w^\star\right)\right\|_{\Lambda^{-1}} + \lambda\|w^\star\|_{\Lambda^{-1}} \le \left\| \sum_{i=1}^N \phi_i\left(y_i - \phi^\top w^\star\right)\right\|_{\Lambda^{-1}} + \sqrt{\lambda}\,\|w^\star\|,$$

as required. □

D.1 Uniform concentration for bonus value functions

In this section we provide lemmas that support uniform concentration over bonus value functions explored by the algorithm. The bound on the covering number of the euclidean ball stated below is standard.

Lemma 18 (Covering number of Euclidean Ball)**.** *For any $\epsilon > 0$, the ϵ-covering of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is upper bounded by $(1 + 2R/\epsilon)^d$.*

The next lemma follows from (relatively standard) arguments that are essentially the same as those of Lemma D.6 in (Jin et al., 2020b).

Lemma 19. *Let $\mathcal{N}_\epsilon(\mathcal{F})$ denote the $\|\cdot\|_\infty$ covering number of a function class $\mathcal{F}$. For some universal constant $c > 0$, we have*

$$\log \mathcal{N}_\epsilon(\mathcal{W}(\beta, \lambda_{\Sigma^+}, \beta^{\mathbb{P}}, \lambda_\Lambda, L, B_{\max}, \pi)) \le cd^2 \log\left(\frac{d\beta^{\mathbb{P}}\beta\lambda_{\Sigma^+}L}{\lambda_\Lambda\epsilon}\right),$$

for the function class $\mathcal{W}$ as defined in Equation (44).

Proof. First, we remove clipping (that can only decrease the covering number), and reparameterize the $\mathcal{B}$ function class Equation (43) with $A = \left(\beta^{\mathbb{P}}\right)^2 \Lambda^{-1}$ and $E = \beta^2 \Sigma^+$, to consider functions of the form

$$B(s,a;E,A,w) = \|\phi(s,a)\|_E + \sum_a \pi(a'|s) \|\phi(s,a')\|_E + \phi(s,a)^\top w + \|\phi(s,a)\|_A ,$$

with parameters $\|w\| \le L$, $\|A\| \le \left(\beta^{\mathbb{P}}\right)^2 \lambda_\Lambda^{-1}$, and $\|E\| \le \beta^2 \lambda_{\Sigma^+}$. Recall that $\|\phi(s,a)\| \le 1$, and observe,

$$\begin{aligned}
&|B(s,a;E_1,A_1,w_1) - B(s,a;E_2,A_2,w_2)| \\
&\le \left|\sqrt{\phi(s,a)^\top E_1 \phi(s,a)} - \sqrt{\phi(s,a)^\top E_2 \phi(s,a)}\right| + \sum_{a'} \pi(a'|s) \left|\sqrt{\phi(s,a')^\top E_1 \phi(s,a')} - \sqrt{\phi(s,a')^\top E_2 \phi(s,a')}\right| \\
&\quad + \|\phi(s,a)\| \|w_1 - w_2\| + \left|\sqrt{\phi(s,a)^\top A_1 \phi(s,a)} - \sqrt{\phi(s,a)^\top A_2 \phi(s,a)}\right| \\
&\le \sqrt{|\phi(s,a)^\top (E_1 - E_2) \phi(s,a)|} + \sum_{a'} \pi(a'|s) \sqrt{|\phi(s,a')^\top (E_1 - E_2) \phi(s,a')|} \\
&\quad + \|w_1 - w_2\| + \sqrt{|\phi(s,a)^\top (A_1 - A_2) \phi(s,a)|} \\
&\le 2\sqrt{\|E_1 - E_2\|} + \|w_1 - w_2\| + \sqrt{\|A_1 - A_2\|} \\
&\le 2\sqrt{\|E_1 - E_2\|_F} + \|w_1 - w_2\| + \sqrt{\|A_1 - A_2\|_F}
\end{aligned}$$

Now, we consider an $\epsilon^2/16$ net over $\left\{E \subset \mathbb{R}^{d\times d} \mid \|E\|_F \le \sqrt{d}\beta^2 \lambda_{\Sigma^+}\right\}$, an $\epsilon/2$ net over $\left\{w \in \mathbb{R}^d \mid \|r\| \le L\right\}$, and an $\epsilon^2/4$ net over $\left\{A \subset \mathbb{R}^{d\times d} \mid \|A\|_F \le \sqrt{d} \left(\beta^{\mathbb{P}}\right)^2 \lambda_\Lambda^{-1}\right\}$. Noting that for any matrix M, $\|M\|_F \le \sqrt{d}\,\|M\|$, we have that the product of these three nets provides an ϵ-net over the original parameter space. By Lemma 18, this implies

$$\log \mathcal{N}_\epsilon(\mathcal{B}) \le d \log(1 + 4L/\epsilon) + d^2 \log\left(1 + 8\sqrt{d}\left(\beta^{\mathbb{P}}\right)^2 \lambda_\Lambda^{-1} \epsilon^{-2}\right) + d^2 \log\left(1 + 8\sqrt{d}\beta^2 \lambda_{\Sigma^+} \epsilon^{-2}\right).$$

Finally, noting that π is a parameter that is held fixed, and that $W(s)$ just averages over values of $B(s,\cdot)$, we have $\log \mathcal{N}_\epsilon(\mathcal{W}) \le \log \mathcal{N}_\epsilon(\mathcal{B})$, and the result follows. ◻

The next lemma is brought as is from (Jin et al., 2020b), except from slight adaptation of notation. We remark that due to the blocking structure / simulator in our algorithms, we could in fact use a similar weaker version of this lemma suitable for random design least squares regression, rather than the one below which is suitable for a martingale setting.

Lemma 20 (**Uniform concentration of self normalized processes**, (Jin et al., 2020b) Lemma D.4)**.** *Let $\{x_\tau\}$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_\tau\}_{\tau=1}^\infty$. Let $\{\phi_\tau\}$ be an $\mathbb{R}^d$-valued stochastic process where $\phi_\tau \in \mathcal{F}_\tau$, and $\|\phi_\tau\| \le 1$. Further, let $\Lambda_n = \lambda I + \sum_{\tau=1}^n \phi_\tau \phi_\tau^\top$. Then for any $\delta > 0$, with probability at least $1-\delta$, for all $n \ge 1$ and any $V \in \mathcal{V}$ so that $\|V\|_\infty \le D$, we have;*

$$\left\|\sum_{\tau=1}^n \phi_\tau \Big(V(x_\tau) - \mathbb{E}\left[V(x_\tau)|\mathcal{F}_{\tau-1}\right]\Big)\right\|_{\Lambda_n^{-1}}^2 \le 4D^2 \left(\frac{d}{2}\log\left(\frac{n+\lambda}{\lambda}\right) + \log\frac{\mathcal{N}_\epsilon(\mathcal{V})}{\delta}\right) + \frac{8n^2\epsilon^2}{\lambda},$$

where $\mathcal{N}_\epsilon(\mathcal{V})$ is $\|\cdot\|_\infty$ covering number of $\mathcal{V}$.

E Matrix Geometric Resampling Lemma Proof

As mentioned, our Algorithm 3 is similar to that of Luo et al. (2021), which itself is the original proposed by Neu & Olkhovskaya (2020a) (see also Neu & Olkhovskaya, 2021; 2020b), but with averaging over multiple estimators. We present here a different analysis to obtain tighter bounds in the 2'nd moment term analysis given in Lemma 9.

Proof of Lemma 3. First, note that since $\gamma < 1/2$ and $c = 1/2$;

$$\begin{aligned}
\left\|\widehat{\Sigma}_{m,\gamma}^{(n)}\right\| \le (1-c\gamma)^n \implies \left\|\widehat{\Sigma}_{m,\gamma}^{+}\right\| \le c\sum_{n=0}^N (1-c\gamma)^n \le \frac{1}{\gamma} \\
\implies \left\|\widehat{\Sigma}_{\gamma}^{+}\right\| \le \frac{1}{\gamma}.
\end{aligned}$$

For the bias claim, using independence of samples;

$$\begin{aligned}
\mathbb{E}\widehat{\Sigma}^{(n)}_{m,\gamma} &= \prod_{i=1}^{n}\left(I - c\mathbb{E}\left[\gamma I + \phi_{m,i}\phi_{m,i}^{\top}\right]\right) = \prod_{i=1}^{n}(I - c\Sigma_\gamma) = (I - c\Sigma_\gamma)^n \\
\implies \mathbb{E}\widehat{\Sigma}^{+}_{m,\gamma} &= cI + c\sum_{n=1}^{N}(I - c\Sigma_\gamma)^n = c\sum_{n=0}^{N}(I - c\Sigma_\gamma)^n,
\end{aligned}$$

hence,

$$\mathbb{E}\widehat{\Sigma}^{+}_{\gamma} = c\sum_{n=0}^{N}(I - c\Sigma_\gamma)^n = \Sigma_\gamma^{-1} - \sum_{n=N+1}^{\infty}(I - c\Sigma_\gamma)^n,$$

where we use that $\gamma < 1/2$ and $c = 1/2$ imply all eigenvalues of $I - c\Sigma_\gamma$ are in $(0, 1)$, and $A^{-1} = \sum_{n=0}^{\infty}(I - A)^n$ for any invertible matrix A with all eigenvalues $\in (0, 1)$. Now,

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

where in the last step we substitute $c = 1/2$ and $N = \frac{2}{\gamma}\log\frac{1}{\gamma\epsilon}$.

Now for the last claim, note that for any m, $\widehat{\Sigma}^{+}_{m,\gamma}$ is a sum of positive definite matrices, with the first term being cI, thus $\lambda_{\min}\left(\widehat{\Sigma}^{+}_{m}\right) \ge 1/2$. In addition, by Lemma 15,

$$M = \frac{48d}{\gamma\sigma}\log\frac{72d}{\gamma^2\sigma} \ge \frac{12d}{\gamma(\sigma/2)}\log\frac{3M}{\gamma} \implies \sigma/2 \ge \frac{12d}{\gamma M}\log\frac{3M}{\gamma},$$

therefore our assumption that $\sigma \le 1/4$ verifies the conditions for Lemma 21 are met. Thus, we obtain;

$$\mathbb{E}\left[\widehat{\Sigma}^{+}_{\gamma}\Sigma_\gamma\widehat{\Sigma}^{+}_{\gamma}\right] \preceq 2\mathbb{E}\left[\widehat{\Sigma}^{+}_{\gamma}\right] + \left(3\epsilon + \frac{12d}{\gamma M}\log\frac{3M}{\gamma}\right)I,$$

and by the previous display,

$$\mathbb{E}\left[\widehat{\Sigma}^{+}_{\gamma}\Sigma_\gamma\widehat{\Sigma}^{+}_{\gamma}\right] \preceq 2\mathbb{E}\left[\widehat{\Sigma}^{+}_{\gamma}\right] + (3\epsilon + \sigma/2)\,I.$$

The proof is complete by our assumption that $\epsilon \le \sigma/6$. □

Lemma 21. *Let $0 < \epsilon < 1/16, 0 < \gamma < 1/2$, and assume $\widehat{\Sigma}^{+}_{1}, \ldots, \widehat{\Sigma}^{+}_{M} \in \mathbb{R}^{d\times d}$ are M i.i.d. random matrices and $\Sigma \in \mathbb{R}^{d\times d}$ is a fixed matrix such that $\gamma \preceq \Sigma \preceq I$, and $\left\|\mathbb{E}\left[\widehat{\Sigma}^{+}\right] - \Sigma^{-1}\right\| \le \epsilon$ where $\widehat{\Sigma}^{+} := \frac{1}{M}\sum_{m=1}^{M}\widehat{\Sigma}^{+}_{m}$. Further, assume that $(1/2)I \preceq \widehat{\Sigma}^{+}_{m} \preceq (1/\gamma)I$ almost surely for all m, and $\frac{8d}{\gamma M}\log\frac{3M}{\gamma} < 1/8$. Then,*

$$\mathbb{E}\left[\widehat{\Sigma}^{+}\Sigma\widehat{\Sigma}^{+}\right] \preceq 2\mathbb{E}\left[\widehat{\Sigma}^{+}\right] + \left(3\epsilon + \frac{12d}{\gamma M}\log\frac{3M}{\gamma}\right)I$$

Proof. Denote $\Sigma^{+} := \mathbb{E}\left[\widehat{\Sigma}^{+}\right]$. By assumption,

$$\Sigma^{+} = \Sigma^{-1} + \left(\Sigma^{+} - \Sigma^{-1}\right) \preceq \Sigma^{-1} + \epsilon I,$$

thus by Lemma 22,

$$\begin{aligned}
&\widehat{\Sigma}^{+} \preceq 2\Sigma^{+} + \alpha I \preceq 2\Sigma^{-1} + (2\epsilon + \alpha)I \\
\iff\ &\widehat{\Sigma}^{+} - (2\epsilon + \alpha)I \preceq 2\Sigma^{-1} \qquad (49)
\end{aligned}$$

holds with probability $\geq 1-\delta$ and $\alpha := \frac{4d}{\gamma M}\log\frac{3M}{\delta}$. Now, as long as $\alpha' := 2\epsilon + \alpha < 1/4$, we have that

$$\lambda_{\min}\left(\widehat{\Sigma}^{+} - \alpha' I\right) \geq 1/2 - \alpha' \geq 1/4,$$

therefore the matrices on both sides of Equation (49) are positive definite, hence

$$\Sigma \preceq 2\left(\widehat{\Sigma}^{+} - \alpha' I\right)^{-1}.$$

This implies that,

$$\widehat{\Sigma}^{+}\Sigma\widehat{\Sigma}^{+} = \left(\widehat{\Sigma}^{+} - \alpha' I\right)\Sigma\widehat{\Sigma}^{+} + \alpha'\Sigma\widehat{\Sigma}^{+} \preceq 2\widehat{\Sigma}^{+} + \alpha'\Sigma\widehat{\Sigma}^{+},$$

holds w.p. $\geq 1-\delta$. This, and considering that $\left\|\widehat{\Sigma}^{+}\Sigma\widehat{\Sigma}^{+} - 2\Sigma^{+} - \alpha'\Sigma\widehat{\Sigma}^{+}\right\| \leq \frac{1}{\gamma^2} + \frac{2}{\gamma} + \frac{1}{\gamma} \leq \frac{4}{\gamma^2}$, implies that for any $\delta > 0$;

$$\begin{aligned}
\mathbb{E}\left[\widehat{\Sigma}^{+}\Sigma\widehat{\Sigma}^{+}\right] &\preceq 2\mathbb{E}\left[\widehat{\Sigma}^{+}\right] + \alpha'\Sigma\mathbb{E}\left[\widehat{\Sigma}^{+}\right] + \frac{4\delta}{\gamma^2}I \\
&= 2\Sigma^{+} + \alpha'\Sigma\Sigma^{+} + \frac{4\delta}{\gamma^2}I \\
&= 2\Sigma^{+} + \alpha' I + \alpha'\Sigma\left(\Sigma^{+} - \Sigma^{-1}\right) + \frac{4\delta}{\gamma^2}I \\
&\preceq 2\Sigma^{+} + \alpha' I + \epsilon I + \frac{4\delta}{\gamma^2}I \\
&\preceq 2\Sigma^{+} + \left(3\epsilon + \frac{4d}{\gamma M}\log\frac{3M}{\delta} + \frac{4\delta}{\gamma^2}\right)I,
\end{aligned}$$

with the last equality following simply by plugging in the definition of α'. Choosing $\delta = \gamma/M$, we may now see that by our assumptions,

$$\alpha' := 2\epsilon + \alpha = 2\epsilon + \frac{4d}{\gamma M}\log\frac{3M^2}{\gamma} \leq \frac{1}{8} + \frac{8d}{\gamma M}\log\frac{3M}{\gamma} < 1/4,$$

which verifies our earlier requirement on α'. The proof is complete by plugging our choice of δ into the previous display. □

Lemma 22. *Assume $\widehat{\Sigma}^{+}_1, \ldots, \widehat{\Sigma}^{+}_M \in \mathbb{R}^{d\times d}$ are M i.i.d. random matrices such that $\left\|\widehat{\Sigma}^{+}_m\right\| \leq 1/\gamma$ almost surely and $\mathbb{E}\widehat{\Sigma}^{+}_m = \Sigma^{+}$. Then, for $\widehat{\Sigma}^{+} = \frac{1}{M}\sum_{i=1}^{M}\widehat{\Sigma}^{+}_m$ and $\alpha = \frac{4d}{\gamma M}\log\frac{3M}{\delta}$, we have*

$$\widehat{\Sigma}^{+} \preceq 2\Sigma^{+} + \alpha I.$$

Proof. For any fixed $\phi \in \mathbb{R}^d$ with $\|\phi\| = 1$, we have by Lemma 23 that w.p. $\geq 1-\delta$:

$$\begin{aligned}
&\sum_{m=1}^{M}\phi^{\top}\widehat{\Sigma}^{+}_m\phi \leq 2\sum_{m=1}^{M}\phi^{\top}\Sigma^{+}\phi + \frac{1}{\gamma}\log\frac{1}{\delta} \\
\implies\ & \phi^{\top}\widehat{\Sigma}^{+}\phi \leq 2\phi^{\top}\Sigma^{+}\phi + \frac{1}{\gamma M}\log\frac{1}{\delta}.
\end{aligned}$$

Consider now an ϵ-net over the unit sphere in $\mathbb{R}^d$ of size $(1+2/\epsilon)^d$, which exists by Lemma 18. By the union bound we have that w.p. $1-\delta$, for all $\tilde{\phi}$ in the net it holds that;

$$\tilde{\phi}^{\top}\widehat{\Sigma}^{+}\tilde{\phi} \leq 2\tilde{\phi}^{\top}\Sigma^{+}\tilde{\phi} + \frac{d}{\gamma M}\log\frac{3}{\delta\epsilon},$$

Thus, w.p. $1-\delta$, for any $\phi \in \mathbb{R}^d$, $\|\phi\| = 1$;

$$\phi^{\top}\widehat{\Sigma}^{+}\phi \leq 2\phi^{\top}\Sigma^{+}\phi + \frac{3\epsilon^2}{\gamma} + \frac{d}{\gamma M}\log\frac{3}{\delta\epsilon} \leq \frac{4d}{\gamma M}\log\frac{3M}{\delta} = \alpha,$$

with the last inequality following from choosing $\epsilon = 1/M$. This implies that

$$\forall \phi, \|\phi\| = 1; \quad \phi^\top \widehat{\Sigma}^+ \phi \le \phi^\top \left(2\Sigma^+ + \alpha I\right) \phi$$
$$\implies \forall \phi \in \mathbb{R}^d; \quad \phi^\top \widehat{\Sigma}^+ \phi \le \phi^\top \left(2\Sigma^+ + \alpha I\right) \phi,$$

which completes the proof. □

Lemma 23. *Let $\{X_i\}_{i=1}^N$ be a sequence of i.i.d. random variables supported on $[0, B]$. Then with probability $\ge 1 - \delta$, we have that;*

$$\sum_{i=1}^N X_i \le 2 \sum_{i=1}^N \mathbb{E}\left[X_i\right] + B \log \frac{1}{\delta}.$$

Proof. Let $Z_i \coloneqq X_i/B, \mu_i \coloneqq \mathbb{E}[Z_i]$, and observe;

$$\mathbb{E}\left[e^{Z_i}\right] \le \mathbb{E}\left[1 + Z_i + Z_i^2\right] \le 1 + 2\mu_i \le e^{2\mu_i},$$

where the first inequality follows from $e^z \le 1 + z + z^2$ for $z \in [0, 1]$, and the last from $1 + z \le e^z$. By independence of the Z_i, this implies that

$$\mathbb{E}\left[e^{\sum_{i=1}^N Z_i - 2\mu_i}\right] = \prod_{i=1}^N \mathbb{E}\left[e^{Z_i - 2\mu_i}\right] \le 1,$$

and therefore by Markov's inequality,

$$\Pr\left(\sum_{i=1}^N Z_i - 2\mu_i \ge w\right) = \Pr\left(e^{\sum_{i=1}^N Z_i - 2\mu_i} \ge e^w\right) \le \mathbb{E}\left[e^{\sum_{i=1}^N Z_i - 2\mu_i}\right] e^{-w} \le e^{-w}.$$

Setting $\delta \coloneqq e^{-w}$, we get that w.p.$\ge 1 - \delta$, $\sum_{i=1}^N Z_i \le 2 \sum_{i=1}^N \mu_i + \log \frac{1}{\delta}$. The result follows by substituting Z_i for X_i/B and rearranging. □

F Additional Lemmas

Lemma 24 (See Lemma D.4 in (Rosenberg et al., 2020))**.** *Let $(\mathcal{F}_i)_{i=1}^\infty$ be a filtration, and let $(X_i)_{i=1}^\infty$ be a sequence of random variables that are $\mathcal{F}_i$-measurable, and supported on $[0, B]$. Then with probability $\ge 1 - \delta$, we have that for any $N \ge 1$;*

$$\sum_{i=1}^N \mathbb{E}\left[X_i \mid \mathcal{F}_{i-1}\right] \le 2 \sum_{i=1}^N X_i + 4B \log \frac{2K}{\delta}.$$

Lemma 25 (Elliptical potential lemma, see (Lattimore & Szepesvári, 2020), Lemma 19.4)**.** *Let $(\phi_i)_{i=1}^N \subset \mathbb{R}^d$ with $\|\phi_i\| \le 1$, and set $\Lambda_i \coloneqq \lambda I + \sum_{t=1}^{i-1} \phi_t \phi_t^\top$ where $\lambda \ge 1$. Then,*

$$\sum_{i=1}^N \|\phi_i\|_{\Lambda_i^{-1}}^2 \le 2d \log\left(1 + \frac{N}{d\lambda}\right)$$

Proof. Note that $\lambda \ge 1$ implies $\|\phi_i\|_{\Lambda_i^{-1}}^2 \le \lambda_{\max}(\Lambda_i^{-1}) \|\phi_i\|^2 \le \lambda^{-1} \le 1$. Thus

$$\sum_{i=1}^N \|\phi_i\|_{\Lambda_i^{-1}}^2 = \sum_{i=1}^N \min\left\{1, \|\phi_i\|_{\Lambda_i^{-1}}^2\right\}.$$

The rest of the proof is identical to (Lattimore & Szepesvári, 2020), with $L = 1$ and $V_0 = \lambda I$. □

Lemma 26 (Extended value difference, (Shani et al., 2020) Lemma 1, see also (Cai et al., 2020))**.** *Let $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \ell)$ be any MDP and $\pi, \pi' \in \mathcal{S} \to \Delta(\mathcal{A})$ be any two policies. Then, for any sequence of functions $\widehat{Q}_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}, \widehat{V}_h^\pi : \mathcal{S} \to \mathbb{R}$, where $\widehat{V}_h^\pi(s) := \left\langle \pi_h(\cdot|s), \widehat{Q}_h^\pi(s,\cdot) \right\rangle$, $h = 1, \ldots, H$, we have*

$$\widehat{V}_1^\pi - V_1^{\pi'} = \sum_{h=1}^{H} \mathbb{E}_{s_h \sim d_h^{\pi'}} \left[\left\langle \widehat{Q}_h^\pi(s_h, \cdot), \pi_h(\cdot|s_h) - \pi'_h(\cdot|s_h) \right\rangle \right] + \sum_{h=1}^{H} \mathbb{E}_{s_h, a_h \sim d_h^{\pi'}} \left[\widehat{Q}_h^\pi(s_h, a_h) - \ell_h(s_h, a_h) - \mathbb{P}_h \widehat{V}_{h+1}^\pi(s_h, a_h) \right].$$

Proof. For any $s \in \mathcal{S}, h \in [H]$, we have

$$\begin{aligned}\widehat{V}_h^\pi(s) - V_h^{\pi'}(s) &= \left\langle \pi_h(\cdot|s), \widehat{Q}_h^\pi(s,\cdot) \right\rangle - \left\langle \pi'_h(\cdot|s), Q_h^{\pi'}(s,\cdot) \right\rangle \\ &= \left\langle \pi_h(\cdot|s) - \pi'_h(\cdot|s), \widehat{Q}_h^\pi(s,\cdot) \right\rangle + \left\langle \pi'_h(\cdot|s), \widehat{Q}_h^\pi(s,\cdot) - Q_h^{\pi'}(s,\cdot) \right\rangle\end{aligned}$$

Further, by the Bellman consistency equations, for all a; $Q_h^{\pi'}(s,a) = \ell_h(s,a) + \mathbb{P}_h V_{h+1}^{\pi'}(s,a)$, thus

$$\begin{aligned}\left\langle \pi'_h(\cdot|s), \widehat{Q}_h^\pi(s,\cdot) - Q_h^{\pi'}(s,\cdot) \right\rangle &= \mathbb{E}_{a \sim \pi'(\cdot|s)} \left[\widehat{Q}_h^\pi(s,a) - \ell_h(s,a) - \mathbb{P}_h V_{h+1}^{\pi'}(s,a) \right] \\ &= \mathbb{E}_{a \sim \pi'(\cdot|s)} \left[\widehat{Q}_h^\pi(s,a) - \ell_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^{\pi}(s,a) \right] \\ &\quad + \mathbb{E}_{a \sim \pi'(\cdot|s)} \left[\mathbb{P}_h \widehat{V}_{h+1}^{\pi}(s,a) - \mathbb{P}_h V_{h+1}^{\pi'}(s,a) \right] \\ &= \mathbb{E}_{a \sim \pi'(\cdot|s)} \left[\widehat{Q}_h^\pi(s,a) - \ell_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^{\pi}(s,a) \right] \\ &\quad + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a), a \sim \pi'(\cdot|s)} \left[\widehat{V}_{h+1}^{\pi}(s') - V_{h+1}^{\pi'}(s') \right].\end{aligned}$$

Combining the last two displays we obtain

$$\begin{aligned}\widehat{V}_h^\pi(s) - V_h^{\pi'}(s) &= \left\langle \pi_h(\cdot|s) - \pi'_h(\cdot|s), \widehat{Q}_h^\pi(s,\cdot) \right\rangle + \mathbb{E}_{a \sim \pi'(\cdot|s)} \left[\widehat{Q}_h^\pi(s,a) - \ell_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^{\pi}(s,a) \right] \\ &\quad + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a), a \sim \pi'(\cdot|s)} \left[\widehat{V}_{h+1}^{\pi}(s') - V_{h+1}^{\pi'}(s') \right].\end{aligned}$$

Unrolling the above relation, the result follows. □

The next lemma is standard, for proof see e.g., Hazan et al. (2016); Lattimore & Szepesvári (2020).

Lemma 27 (Entropy regularized OMD). *Let $\eta > 0$, and $g_k \in \mathbb{R}^n$, $x_k \in \Delta(n)$ be a sequence of vectors such that for all a, $x_1(a) = 1/n$, for all $k \in [K], a \in [n]$, $\eta g_k(a) \geq -1$ and*

$$x_{k+1}(a) = \frac{x_k(a) e^{-\eta g_t(a)}}{\sum_{a' \in [n]} x_k(a') e^{-\eta g_k(a')}}.$$

Then,

$$\max_{x \in \Delta_n} \left\{ \sum_{k=1}^{K} \langle g_k, x_k - x \rangle \right\} \leq \frac{\log n}{\eta} + \eta \sum_{k=1}^{K} \sum_{i=1}^{n} x_k(i) g_k(i)^2.$$

The next lemma establishes a regret bound for OMD with blocking, and follows from standard arguments. We provide a proof for completeness.

Lemma 28 (Entropy regularized OMD with blocking). *Let $K \in \mathbb{Z}_+, \tau \leq K, J = \lceil K/\tau \rceil$, and set $T_j := \{\tau(j-1)+1, \ldots, \tau j\}$ for all $j \in [J]$. Assume $\eta > 0$, let $g_k \in \mathbb{R}^n$ be a sequence of vectors such that $\forall a, k; \eta g_k(a) \geq -1$,* *and set*

$$g_{(j)} = \frac{1}{\tau}\sum_{k\in T_j} g_k \ \forall j \in [J]$$

$$x_{(j+1)}(a) = \frac{x_{(j)}(a)e^{-\eta g_{(j)}(a)}}{\sum_{a'\in[n]} x_{(j)}(a')e^{-\eta g_{(j)}(a')}}.$$

Then if $x_k \in \Delta(n)$ *are such that* $x_k = x_{(j)}$ *for all* $k \in T_j, j \in [J]$ *we have*

$$\max_{x\in\Delta_n}\left\{\sum_{k=1}^{K}\langle g_k, x_k - x\rangle\right\} \le \frac{\tau\log n}{\eta} + \tau\max_k \|g_k\|_\infty + \eta\sum_{k=1}^{K}\sum_{i=1}^{n} x_k(i)g_k(i)^2.$$

Proof. By applying Lemma 27 on $g_{(j)}, x_{(j)}$, we get

$$\sum_{j=1}^{J}\left\langle g_{(j)}, x_{(j)} - x^\star\right\rangle \le \frac{\log n}{\eta} + \eta\sum_{j=1}^{J}\sum_{i=1}^{n} x_{(j)}(i)g_{(j)}(i)^2.$$

In addition,

$$\sum_{j=1}^{J}\left\langle g_{(j)}, x_{(j)} - x^\star\right\rangle = \sum_{j=1}^{J}\left\langle \frac{1}{|T_j|}\sum_{k\in T_j} g_k, x_{(j)} - x^\star\right\rangle = \sum_{j=1}^{J}\frac{1}{|T_j|}\sum_{k\in T_j}\left\langle g_k, x_k - x^\star\right\rangle \ge \frac{1}{\tau}\sum_{k=1}^{K}\left\langle g_k, x_k - x^\star\right\rangle$$

Further, by Jensen's inequality,

$$g_{(j)}(i)^2 = \left(\frac{1}{|T_j|}\sum_{k\in T_j} g_k(i)\right)^2 = \frac{1}{|T_j|^2}\left(\sum_{k\in T_j} g_k(i)\right)^2 \le \frac{1}{|T_j|}\sum_{k\in T_j} g_k(i)^2,$$

thus

$$\frac{1}{\tau}\sum_{k=1}^{K}\left\langle g_k, x_k - x^\star\right\rangle \le \frac{\log n}{\eta} + \frac{\eta}{\tau}\sum_{k=1}^{K'}\sum_{i=1}^{n} x_k(i)g_k(i)^2 + \frac{\eta}{|T_J|}\sum_{k\in T_J}\sum_{i=1}^{n} x_k(i)g_k(i)^2,$$

where $K' = \max\{k \in T_{J-1}\}$. Finally,

$$\begin{aligned}\sum_{k=1}^{K}\left\langle g_k, x_k - x^\star\right\rangle &\le \frac{\tau\log n}{\eta} + \eta\sum_{k=1}^{K'}\sum_{i=1}^{n} x_k(i)g_k(i)^2 + \frac{\tau\eta}{|T_J|}\sum_{k\in T_J}\sum_{i=1}^{n} x_k(i)g_k(i)^2, \\ &\le \frac{\tau\log n}{\eta} + \eta\sum_{k=1}^{K'}\sum_{i=1}^{n} x_k(i)g_k(i)^2 + \frac{\tau}{|T_J|}\sum_{k\in T_J}\sum_{i=1}^{n} x_k(i)g_k(i) \\ &\le \frac{\tau\log n}{\eta} + \eta\sum_{k=1}^{K}\sum_{i=1}^{n} x_k(i)g_k(i)^2 + \tau\max_k \|g_k\|_\infty,\end{aligned}$$

which concludes the proof. □