# Online learning in MDPs with linear function approximation and bandit feedback

**Gergely Neu**
Universitat Pompeu Fabra
Barcelona, Spain
gergely.neu@gmail.com

**Julia Olkhovskaya**
Universitat Pompeu Fabra
Barcelona, Spain
julia.olkhovskaya@gmail.com

## Abstract

We consider an online learning problem where the learner interacts with a Markov decision process in a sequence of episodes, where the reward function is allowed to change between episodes in an adversarial manner and the learner only gets to observe the rewards associated with its actions. We allow the state space to be arbitrarily large, but we assume that all action-value functions can be represented as linear functions in terms of a known low-dimensional feature map, and that the learner has access to a simulator of the environment that allows generating trajectories from the true MDP dynamics. Our main contribution is developing a computationally efficient algorithm that we call MDP-LINEXP3, and prove that its regret is bounded by $\widetilde{\mathcal{O}}\big(H^2 T^{2/3} (dK)^{1/3}\big)$, where $T$ is the number of episodes, $H$ is the number of steps in each episode, $K$ is the number of actions, and $d$ is the dimension of the feature map. We also show that this bound can be improved to $\widetilde{\mathcal{O}}\big(H^2 \sqrt{dKT}\big)$ under the strong condition that the likelihood ratio between the state distributions generated by any pair of policies is upper bounded by a constant. To our knowledge, MDP-LINEXP3 is the first provably efficient algorithm for this problem setting.

## 1 Introduction

Reinforcement learning (RL) is one of the most important frameworks for sequential decision-making under uncertainty, where a learner interacts with the environment sequentially and aims to improve her strategy over time [39, 41]. Besides well-publicized spectacular empirical successes of RL algorithms, recent years saw a renaissance of theoretical research in the field. Our paper contributes to this line of study by providing algorithms with provable performance guarantees.

In the present work, we study the problem of online learning in episodic Markov Decision Processes (MDP), where the interaction is divided into $T$ episodes of fixed length $H$. At each time step of the episode, the learner observes the current state of the environment, chooses one of the available actions, and earns a reward. Consequently, the state of the environment changes according to the transition function of the underlying MDP, as a function of the previous state and the action taken by the learner. We assume that the reward function can change arbitrarily between episodes, and the learner only has access to bandit feedback: instead of being able to observe the reward function at the end of the episode, the learner only gets to observe the rewards that it actually received. As traditional in this line of work, we aim to design algorithms for the learner with theoretical guarantees on her regret, which is the difference between the total reward accumulated by the learner and the total reward of the best stationary policy fixed in hindsight.

Unlike most previous work on this problem, we allow the state space to be potentially infinite, bringing theory one step closer to practical scenarios where assuming finite state spaces is unrealistic. To address the challenge of learning in large state spaces, we adopt the classic RL technique of using *linear function approximation* and suppose that we have access to a relatively low-dimensional feature map that can be used to represent policies and value functions. We will assume that the feature map is expressive enough so that all action-value functions can be expressed as linear functions of the features. While we do not assume full knowledge of the underlying Markov decision process, we will assume that we have access to a *simulator* of the MDP that will allow the learner to generate sample episodes.

Our main contribution is designing a computationally efficient algorithm called MDP-LINEXP3, and prove that in the setting described above, its regret is at most $\widetilde{\mathcal{O}}\left(H^2T^{2/3}(dK)^{1/3}\right)$, where $K$ is the number of actions and $d$ is the dimensionality of the feature map. We also show that this bound can be improved to $\widetilde{\mathcal{O}}\left(H^2\sqrt{dKT}\right)$ under the strong condition that the likelihood ratio between the state distributions generated by any pair of policies is upper bounded by a constant. These results constitute the first known regret guarantee for any algorithm in this setting.

Our result fits into a long line of work considering online learning in Markov decision processes. The problem of regret minimization in stationary Markov decision processes with a *fixed* reward function has been studied extensively since the work of Burnetas and Katehakis [13], Auer and Ortner [7], Tewari and Bartlett [42], Jaksch et al. [22], with several important advances made in the past decade [17, 18, 10, 21, 23]. While most of these works considered small finite state spaces, the same techniques have been very recently extended to accommodate potentially unbounded state spaces under the assumption of realizable function approximation by Jin et al. [24] and Yang and Wang [44].

Even more relevant is the line of work considering adversarial rewards, initiated by Even-Dar et al. [20], who consider online learning in continuing MDPs with full feedback about the rewards. They proposed a MDP-E algorithm, that achieves $\mathcal{O}(\tau^2\sqrt{T\log K})$ regret, where $\tau$ is an upper bound on the mixing time of the MDP. Later, Neu et al. [33] proposed an algorithm which guarantees $\widetilde{\mathcal{O}}\big(\sqrt{\tau^3KT/\alpha}\big)$ regret with bandit feedback, essentially assuming that all states are reachable with probability $\alpha > 0$ under all policies. In our work, we focus on episodic MDPs with a fixed episode length $H$. The setting was first considered in the bandit setting by Neu et al. [31], who proposed an algorithm with $\mathcal{O}(H^2\sqrt{TK}/\alpha)$. Although the number of states does not show up in the bound, the regret scales at least linearly with the size of the state space $\mathcal{S}$, since $|\mathcal{S}| \leq H/\alpha$. Later work by Zimin and Neu [46], Dick et al. [19] eliminated the dependence on $\alpha$ and proposed an algorithm achieving $\widetilde{\mathcal{O}}(\sqrt{TH|\mathcal{S}|K})$ regret. Regret bounds for the full-information case without prior knowledge of the MDP were achieved by [32] and [37], both of order $\widetilde{\mathcal{O}}(H|\mathcal{S}|K\sqrt{T})$.

As apparent from the above discussion, all work on online learning in MDPs with adversarial rewards considers finite state spaces. The only exception we are aware of is the very recent work [14], whose algorithm OPPO is guaranteed to achieve $\widetilde{\mathcal{O}}\big(\sqrt{d^3H^3T}\big)$ regret. While they remarkably assumed no prior knowledge of the MDP parameters, their guarantees are only achieved in the full-information case. This is to be contrasted with our results that are achieved for the much more restrictive bandit setting, albeit with the more permissive assumption of having access to a simulator of the environment.

Our problem formulation shares many similarities with the framework of the contextual bandit problem, which can be seen as a relatively simple reinforcement learning problem in which the actions do not influence the future evolution of states, but nevertheless can still model a broad range of important real-world problems [43, 5]. Several variants and special cases of the contextual bandit problem exist differing on the assumptions on the set of available actions, context and reward functions [27, 11, 4, 36, 40]. Our own approach in this paper is directly motivated by the recent work of Neu and Olkhovskaya [30], who consider linear contextual bandit problems with adversarial reward functions and i.i.d. context vectors. While we borrow several ideas and techniques from their work, our analysis faces a range of new challenges posed by the fact that the feature vectors our algorithm has to deal with are far from being stationary due to the dependence on the learner's policy. Notably, our work is not the first to make use of ideas originally developed for linear bandit settings in the context of RL with linear function approximation: the recently proposed algorithms of Jin et al. [24] and Yang and Wang [44] addressing this setting are both based on extending ideas from stochastic linear contextual bandits [6, 15, 1].

The rest of the paper is organized as follows. After defining some basic notation, Section 2 presents our problem definition. We present our algorithms and main results in Section 3 and provide the proofs in Section 4. Section 5 concludes the paper by discussing some implications of our results.

**Notation.** We use $\langle \cdot, \cdot \rangle$ to denote inner products in Euclidean space and by $\|\cdot\|$ we denote the Euclidean norm for vectors and the operator norm for matrices. For a symmetric positive definite matrix $A$, we use $\lambda_{\min}(A)$ to denote its smallest eigenvalue. We write $\operatorname{tr}(A)$ for the trace of a matrix $A$ and use $A \succcurlyeq 0$ to denote that an operator $A$ is positive semi-definite, and we use $A \succcurlyeq B$ to denote $A - B \succcurlyeq 0$. For a positive integer $N$, we use $[N]$ to denote the set of positive integers $\{1, 2, \ldots, N\}$.

## 2 Problem definition

An episodic Markovian Decision Process (MDP), denoted by $M(\mathcal{S}, \mathcal{A}, H, P, r)$ is defined by a state space $\mathcal{S}$, action space $\mathcal{A}$, episode length $H \in \mathbb{Z}_+$, transition function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ and a reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$. We assume that $\mathcal{S}$ is a measurable space and $\mathcal{A}$ is a finite set with cardinality $K$. Without significant loss of generality, we will assume that the set of available actions is the same $\mathcal{A} = [K]$ in each state, and furthermore that the MDP has a layered structure, satisfying the following conditions:

- The state set $\mathcal{S}$ can be decomposed into $H$ disjoint sets: $\mathcal{S} = \cup_{h=1}^{H} \mathcal{S}_h$,
- $\mathcal{S}_1 = \{x_1\}$ and $\mathcal{S}_H = \{x_H\}$ are singletons,
- transitions are only possible between consecutive layers, that is, for any $x_h \in \mathcal{S}_h$, the distribution $P(\cdot|x, a)$ is supported on $\mathcal{S}_{h+1}$ for all $a$ and $h \in [H]$.

These assumptions are common in the related literature (e.g., [31, 46, 37]) and are not essential to our analysis; their primary role is simplifying our notation. In the present paper, we consider an *online learning* problem where the learner interacts with its environment in a sequence of episodes $t = 1, 2, \ldots, T$, facing a different *reward function* $r_t$ selected by a (possibly adaptive) adversary at the beginning of each episode $t$. Oblivious to the reward function chosen by the adversary, the learner starts interacting with the MDP in each episode from the initial state $X_{t,1} = x_1$. At each consecutive step $h \in [H]$ within the episode, the learner observes the state $X_{t,h}$, picks an action $A_{t,h}$ and observes the loss $r_t(X_{t,h}, A_{t,h})$. Then, unless $h = H$, the learner moves to the next state $X_{t,h+1}$, which is generated from the distribution $P(\cdot|X_{t,h}, A_{t,h})$. At the end of step $H$, the episode terminates and a new one begins.

Our algorithm and analysis will make use of the concept of (stationary stochastic) *policies* $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$. A policy $\pi$ prescribes a behavior rule to the learner by assigning probability $\pi(a|x)$ to taking action $a$ at state $x$. Each policy $\pi$ generates a probability distribution $\mu_h^\pi$ over each layer $h \in [H]$, and we will refer to the collection of these distributions in each layer as the *occupancy measure* $\mu^\pi$ induced by $\pi$. The instantaneous value function and action-value function with respect to policy $\pi$ in episode $t$ are defined, respectively, as

$$Q_{t,h}^{\pi}(x, a) = \mathbb{E}_\pi \left[ \sum_{k=h}^{H} r_t(\widetilde{X}_k, \widetilde{A}_k) \middle| \widetilde{X}_h = x, \widetilde{A}_h = a \right], \quad V_{t,h}^{\pi}(x) = \sum_a \pi(a|x) Q_{t,h}^{\pi}(x, a),$$

where the notation $\mathbb{E}_\pi[\cdot]$ highlights that the sequence of states $\widetilde{X}_k$ and actions $\widetilde{A}_k$ are generated by following policy $\pi$ in the MDP.

We will be interested in developing learning algorithms that select a policy $\pi_t$ for the learner at the beginning of each episode $t$. With some abuse of notation, we will use $V_{t,l}(x) = V_{t,l}^{\pi_t}(x)$ and $Q_{t,l}(x, a) = Q_{t,l}^{\pi_t}(x, a)$ to denote the value function and the action-value function of policy $\pi_t$ in episode $t$. With this notation, we define our performance metric as the (total expected) *regret*

$$R_T = \sup_\pi \sum_{t=1}^{T} \left( V_{t,1}^{\pi}(x_1) - V_{t,1}(x_1) \right),$$

where the supremum is taken over the set of all stationary policies mapping states to actions. It follows from standard results that there exists a stationary and deterministic policy $\pi^*$ that achieves the corresponding infimum [35, Theorem 4.4.2]. Intuitively, the regret measures the gap between the total loss incurred by the learner and that of the best stationary policy fixed in hindsight, with full knowledge of the sequence of losses chosen by the adversary. This performance measure is standard in the related literature on online learning in MDPs, see, for example [31, 46, 32, 37, 14, 25].

In this paper, we focus on MDPs with potentially infinite state spaces, which makes it difficult to design computationally tractable algorithms with nontrivial guarantees, unless we make some assumptions. We particularly focus on the classic technique of relying on *linear function approximation* by assuming that the action-value functions occurring during the learning process can be written as a linear function of a low-dimensional feature map.

**Assumption 1** (Realizable function approximation)**.** *For any $h \in [H]$, $a \in \mathcal{A}$ there exists a feature map $\varphi : \mathcal{S} \to \mathbb{R}^d$, and there exist vectors $\theta^\pi_{t,a,h} \in \mathbb{R}^d$, such that for any $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [L]$ and a stochastic stationary policy $\pi$, the action-value function can be written as*

$$Q^\pi_{t,h}(x, a) = \left\langle \varphi(x), \theta^\pi_{t,a,h} \right\rangle . \tag{1}$$

*Furthermore, the features and the parameter vectors satisfy $\|\varphi(x)\| \le \sigma$ and $\left\|\theta^\pi_{t,a,h}\right\| \le R$ for all $x \in \mathcal{S}$, $a \in \mathcal{A}$, $h \in [H]$.*

Online learning under this assumption has received substantial attention in the recent literature, and in particular has been shown to be satisfied in the class of so-called *linear MDPs* studied by Jin et al. [24], Cai et al. [14] and *low-rank MDPs* studied by Yang and Wang [44], which are both special cases of *factored linear models* [45, 34].

## 3 Algorithm and main results

Our algorithm design is motivated by the following decomposition of the regret first proposed for online MDP problems by Even-Dar et al. [20] and adapted to finite-horizon MDPs by Neu et al. [31]:

**Lemma 1.** *Let $\mu^*$ denote the occupancy measure induced by $\pi^*$. Then, for any sequence of policies $\pi_t$ selected by the learner, the regret satisfies*

$$R_T = \sum_{h=1}^{H} \mathbb{E}_{X^*_h \sim \mu^*_h} \left[ \sum_{t=1}^{T} \Big( Q_{t,h}(X^*_h, \pi^*(X^*_h)) - V_{t,h}(X^*_h) \Big) \right].$$

As observed in previous work [20, 31], this lemma implies that the global regret minimization problem can be decomposed into a set of local regret minimization problems in each state $x$, where the reward function associated with each action $a$ is defined as $Q_{t,h}(x, a)$. Indeed, letting $\pi_t(\cdot|x)$ denote the policy played by the local algorithm in state $x$ in round $t$, we can define the *local regret* against policy $\pi^*$ as

$$R_{h,T}(x) = \mathbb{E}\left[ \sum_{t=1}^{T} \Big( Q_{t,h}(x, \pi^*(x)) - \sum_a \pi_t(a|x) Q_{t,h}(x, a) \Big) \right],$$

and the regret in layer $h$ as $R_{T,h} = \mathbb{E}_{X^*_h \sim \mu^*_h}\left[R_{h,T}(X^*_h)\right]$. This can be easily seen to be related to the global regret as $R_T = \sum_{h=1}^{H} R_{T,h}$, and thus it is obvious that bounding the local regrets in each state $x$ yields a bound on the global regret.

With this in mind, following the algorithmic template laid out by Even-Dar et al. [20] called MDP-E, we propose an algorithm based on running a variant of the classic EXP3 algorithm of Auer et al. [8] in each state $x$. The key challenge is constructing the inputs to these local algorithms in a way that yields a computationally tractable algorithm with nontrivial performance bounds, and more specifically to achieve runtime and regret guarantees that are *independent of the size of the state space*. Indeed, instead of the possibly unbounded number of states, we prefer to have the *dimensionality of the feature map* appear in our bounds, which is made possible by Assumption 1. Indeed, this assumption allows us to represent each $Q$-function by its parameter vector, which in turn enables an efficient implementation of the local regret minimization algorithms. Specifically, we design an estimator $\widehat{\theta}_{t,a,h}$ of the parameter vector $\theta_{t,a,h}$ corresponding to the action-value function $Q_{t,h}(x, a) = \langle \varphi(x), \theta_{t,a,h} \rangle$ of policy $\pi_t$, and plug the resulting estimates $\langle \varphi(x), \widehat{\theta}_{t,a,h} \rangle$ into a local copy of EXP3. The form of our estimator $\widehat{\theta}_{t,a,h}$ and overall algorithm design is directly influenced by the recently proposed LINEXP3 method of Neu and Olkhovskaya [30], and thus we refer to our algorithm as MDP-LINEXP3. Its pseudocode is presented as Algorithm 1.

We denote the state-action trajectory in episode $t$ as $U_t = ((X_{t,1}, A_{t,1}), \ldots, (X_{t,H}, A_{t,H}))$. For stating many of our technical results, we define the filtration $\mathcal{F}_t = \sigma(U_s, s \le t)$, and the notation

**Algorithm 1** MDP-LinExp3

**Parameters:** Learning rate $\eta > 0$, exploration parameter $\gamma \in (0, 1)$,
**Initialization:** Set $\widehat{\theta}_{1,a,h} = \bar{0}$ for all $a \in \mathcal{A}$, $h \in [H]$.
**For** episode $t = 1, \ldots, T$, **repeat:**

- Draw $Y_t \sim Ber(\gamma)$,

**For** step $h = 1, \ldots, H$, **do:**

1. Observe $X_{t,h}$ and, for all $a \in \mathcal{A}(X_{t,h})$, set
$$w_t(X_{t,h}, a) = \exp\left(\eta \cdot \sum_{s=1}^{t-1} \langle \varphi(X_{t,h}), \widehat{\theta}_{s,a,h} \rangle\right),$$
2. draw $A_{t,h}$ from the policy defined as
$$\pi_t\left(a | X_{t,h}\right) = \frac{w_t(X_{t,h}, a)}{\sum_{a' \in \mathcal{A}(X_{t,h})} w_t(X_{t,h}, a')} \mathbb{I}_{\{Y_t = 0\}} + \frac{1}{K} \mathbb{I}_{\{Y_t = 1\}},$$
3. observe the reward $r_t(X_{t,h}, A_{t,h})$.

**For** step $h = H, \ldots, 1$, **do:**

- Compute $\widehat{\theta}_{t,a,h}$ for all $a \in U_t$.

---

$\mathbb{E}_t\left[\cdot\right] = \mathbb{E}\left[\cdot \,|\mathcal{F}_t\right]$. Our reward estimator will be based on the observed rewards, and particularly the partial sums $G_{t,h} = \sum_{k=h}^{H} r_t(X_{t,k}, A_{t,k})$ for each layer $h$. Another key component will be the following *covariance matrix*:

$$\Sigma_{t,a,h} = \mathbb{E}_t\left[\varphi(X_{t,h})\varphi(X_{t,h})^\mathsf{T} \mathbb{I}_{\{A_{t,h} = a\}}\right].$$

Making sure that $\Sigma_{t,a,h}$ is invertible, we can define the estimator

$$\widetilde{\theta}_{t,a,h} = \Sigma_{t,a,h}^{-1} \varphi(X_{t,h}) G_{t,h} \mathbb{I}_{\{A_{t,h} = a\}}. \tag{2}$$

This estimate shares many similarities with the estimates that are broadly used in the literature on adversarial linear bandits [16, 12]. To appreciate the sensibility of this estimator, first notice that the sum of the rewards over the layers is an unbiased estimator of the action-value function:

$$\mathbb{E}_t\left[G_{t,h} |\, X_{t,h} = x, A_{t,h} = a\right] = Q_{t,h}(x, a) = \langle \varphi(x), \theta_{t,a,h} \rangle.$$

Thus, it is easy to see that $\widetilde{\theta}_{t,a,h}$ is an unbiased estimate of vector $\theta_{t,a,h}$:

$$\mathbb{E}_t\left[\widetilde{\theta}_{t,a,h}\right] = \mathbb{E}_t\left[\Sigma_{t,a,h}^{-1} \varphi(X_{t,h})\varphi(X_{t,h})^\mathsf{T} \theta_{t,a,h} \mathbb{I}_{\{A_{t,h} = a\}}\right] = \Sigma_{t,a,h}^{-1} \Sigma_{t,a,h} \theta_{t,a,h} = \theta_{t,a,h}.$$

The downside of this estimator is that it is virtually impossible to compute, since the matrix $\Sigma_{t,a,h}$ depends on both the policy $\pi_t$ and on the unknown dynamics of MDP in a complicated fashion. To address the difficulty associated with not knowing the MDP, we assume that the learner has access to a *simulator* of the environment that allows drawing sample trajectories $U_t$ from the true dynamics of the MDP without interacting with the environment. Notice that this notion is weaker than the more common concept of a generative model that can generate sample transitions from *any given state-action pair* [9, 38, 26]: the simulator can only produce sample trajectories from a given policy and from the given starting state $x_1$. Armed with this sumulator, we propose a method to directly estimate the inverse of the covariance matrix $\Sigma_{t,a,h}$ by adapting the Matrix Geometric Resampling method of Neu and Olkhovskaya [30] (which itself is originally inspired by the Geometric Resampling method of Neu and Bartók [28, 29]). Our adaptation has two parameters $\beta > 0$ and $M \in \mathbb{Z}_+$, and generates an estimate of the inverse covariance matrix through the following procedure[1]:

[1]The version we present here is a naïve implementation, optimized for readability. We present a more practical variant in Appendix C

**Matrix Geometric Resampling with a Simulator**

**Input:** simulator of $P$, policy $\pi_t$, sequence of actions $(a_1, a_2, \dots, a_H)$.
**For** $i = 1, \dots, M$**, repeat**:

1. Generate a path $U(i) = \{(X_1(i), A_1(i)), \dots, (X_H(i), A_H(i))\}$, following the policy $\pi_t$ in the simulator of $P$,
2. **For** $h = 1, \dots, H$**, repeat**:
   (a) compute $B_{i,a_h,h} = \mathbb{I}_{\{A_h(i)=a_h\}}\varphi(X_h(i))\varphi(X_h(i))^\mathsf{T}$,
   (b) compute $C_{i,a_h,h} = \prod_{j=1}^{i}(I - \beta B_{j,a_h,h})$.

**Return** $\widehat{\Sigma}^+_{t,a_h,h} = \beta I + \beta \sum_{i=1}^{M} C_{i,a_h,h}$ for all $h \in [H]$.

Based on the above procedure, we finally define our reward estimator as

$$\widehat{\theta}_{t,a,h} = \widehat{\Sigma}^+_{t,a,h}\varphi(X_{t,h})G_{t,h}\mathbb{I}_{\{A_{t,h}=a\}}.$$

To get an intuitive understanding of the estimate, assume that $M = \infty$, and take $\beta \le \frac{1}{\sigma^2}$, so that the expectation of the matrix $\widehat{\Sigma}^+_{t,a,h}$ can be seen to be the Neumann-series expansion of the matrix $\Sigma^{-1}_{t,a,h}$:

$$\mathbb{E}_t\left[\widehat{\Sigma}^+_{t,a,h}\right] = \beta I + \beta\sum_{k=1}^{\infty}(I - \beta\Sigma_{t,a,h})^k = \beta(\beta\Sigma_{t,a,h})^{-1} = \Sigma^{-1}_{t,a,h}.$$

This suggests that, for large enough $M$, the matrix $\Sigma^+_{t,a,h}$ should be a good estimator of the inverse covariance matrix, which will be quantified formally in the analysis. With a careful implementation explained in Appendix C, $\widehat{\theta}_{t,a,h}$ can be computed in $O(MHKd)$ time, using $M$ calls to the simulator.

Having access to these estimators, MDP-LINEXP3 constructs an EXP3-style policy defined as $\pi_t(a|x) \propto \exp\bigl(\sum_{k=1}^{t-1}\langle\varphi(x), \widehat{\theta}_{k,a,h}\rangle\bigr)$ for each $x, a$. Notably, the policy only depends on the cumulative parameter vectors and the feature vector $\phi(x)$, and thus does not have to make explicit updates to the individual regret-minimization algorithms acting in the states $x$. For technical reasons, MDP-LINEXP3 follows *the uniform policy* $\pi_U(a|x) = \frac{1}{K}$ with probability $\gamma$ in each episode, and follows the above exponential-weights policy otherwise. We will denote the covariance matrix generated by the uniform policy at layer $h$ as $\Sigma_h$, and make the following assumption:

**Assumption 2.** *The eigenvalues of $\Sigma_h$ for all $h$ are lower bounded by $\lambda_{\min} > 0$.*

Our main result is the following guarantee regarding the performance of MDP-LINEXP3:

**Theorem 1.** *Suppose that the MDP satisfies Assumptions 1 and 2 and $\lambda_{\min} > 0$. Then, for $\gamma \in (0, 1)$, $M \ge 0$, any positive $\eta \le \frac{2}{(M+1)H}$ and any positive $\beta \le \frac{1}{2\sigma^2}$, the expected regret of* MDP-LINEXP3 *over $T$ episodes satisfies*

$$R_T \le 2T\sigma RH \cdot \exp\left(-\frac{\gamma\beta\lambda_{\min}M}{K}\right) + \gamma H^2 T + \eta H^3 d\left(\frac{1}{3} + \frac{K}{\gamma}\frac{\sigma^2}{\lambda_{\min}}\right)T + H \cdot \frac{\log K}{\eta}.$$

*Furthermore, letting $\beta = \frac{1}{2\sigma^2}$, $M = \left\lceil \frac{K\sigma^2\log(T\sigma^2R^2)}{\gamma\lambda_{\min}}\right\rceil$, $\eta = \frac{(\log K)^{2/3}\lambda_{\min}^{1/3}}{T^{2/3}(dK\sigma^2)^{1/3}H}$, $\gamma = \frac{(\sigma^2 dK\log K)^{1/3}}{(T\lambda_{\min})^{1/3}}$ and supposing that $T$ is large enough so that the above constraints on $\gamma, M, \eta$ and $\beta$ are satisfied, we also have*

$$R_T \le 3H^2T^{2/3}\left(\frac{dK\sigma^2\log K}{\lambda_{\min}}\right)^{1/3} + \frac{1}{3}H^2T^{1/3}(d\log K)^{2/3}\left(\frac{\lambda_{\min}}{K\sigma^2}\right)^{1/3} + 4H\sqrt{T}.$$

The downside of the above result is that it scales with the time horizon as $T^{2/3}$, which is likely to be suboptimal in light of the best known bounds of order $\sqrt{T}$ achieved in the tabular setting [46] in the bandit case and the large-scale setting considered by Cai et al. [14] in the full-information case. The next result shows that this dependence can be improved at the price of making stronger assumptions about the MDP. Specifically, assume that $P$ is such that for any policy $\pi$, the occupancy distribution $\mu_h^\pi$ has a density $f_h^\pi(x)$ on the set of states $\mathcal{S}_h$ with respect to some base measure, and denote the density corresponding to $\mu_h^*$ as $f_h^*(x)$. Then, assuming that the likelihood ratio $\frac{f_h^*(x)}{f_h^\pi(x)}$ is uniformly upper bounded, our bounds above can be tightened significantly:

**Theorem 2.** *Suppose that the MDP satisfies Assumptions 1, 2 and that the likelihood ratio between the occupancy measures induced by any policy $\pi$ and $\pi^*$ can be bounded uniformly as $\sup_{\pi,h,x} \frac{f_h^*(x)}{f_h^\pi(x)} \leq \rho$ for some $\rho > 0$. Then, for $\gamma \in (0,1)$, $M \geq 0$, any positive $\eta \leq \frac{2}{(M+1)H}$ and any positive $\beta \leq \frac{1}{2\sigma^2}$, the expected regret of* MDP-LINEXP3 *over $T$ episodes, satisfies*

$$R_T \leq 2T\sigma RH \cdot \exp\left(-\frac{\gamma\beta\lambda_{\min}M}{K}\right) + \gamma H^2 T + \eta H^3 d\left(\frac{1}{3} + K\rho\right)T + H \cdot \frac{\log K}{\eta}.$$

*Furthermore, letting $\beta = \frac{1}{2\sigma^2}$, $M = \left\lceil \frac{K\sigma^2 \log(T\sigma^2 R^2)}{\gamma\lambda_{\min}} \right\rceil$, $\eta = \frac{1}{H}\sqrt{\frac{\log K}{TdK\rho}}$, $\gamma = \sqrt{\frac{d\rho \log K}{T}}$ and supposing that $T$ is large enough so that the above constraints are satisfied, we also have*

$$R_T \leq 3H^2\sqrt{\rho dKT \log K} + \frac{1}{3}H^2\sqrt{\frac{Td\log K}{\rho K}} + 4H\sqrt{T}.$$

# 4 Analysis

As explained in the previous section, our algorithm and analysis is based on decomposing the overall learning problem to a number of local online learning problems corresponding to each state in the MDP. This approach is closely related to the one recently taken by Neu and Olkhovskaya [30], who considered the contextual bandit problem with i.i.d. context vectors and adversarially chosen linear reward functions, and suggested a similar regret decomposition for that problem. Our analysis in the present paper will make use of several tools developed by them, with the added challenge that the feature vectors in our setting are no longer i.i.d.: in any layer, the distribution of states clearly depends on the learner's policy in the previous layers. Concretely, the main challenge in our analysis comes from the fact that the state distribution $\mu^*$ appearing in the regret decomposition of Lemma 1 does not match the actual distribution of states $\mu_t$. In what follows, we highlight the main steps in the analysis. Proofs of the lemmas are given in the Appendix.

We start by rewriting our reward estimator as $\widehat{\theta}_{t,a,h} = \widetilde{\theta}_{t,a,h} + b_{t,a,h}$, where $\widetilde{\theta}_{t,a,h}$ is such that $\mathbb{E}_t\left[\widetilde{\theta}_{t,a,h}\right] = \theta_{t,a,h}$ and $b_{t,a,h}$ is a bias term. The bulk of our analysis is based on the following regret decomposition that further refines the decomposition given in Lemma 1:

**Lemma 2.** *Let $X_h^*$ be sampled from the context distribution generated by $\mu_h^*$. Suppose that $\pi_t \in \mathcal{F}_{t-1}$ and that $\mathbb{E}_t\left[\widetilde{\theta}_{t,a,h}\right] = \theta_{t,a,h}$ for all $t, a, h$. Then, for all $h$,*

$$R_{T,h} = \sum_{t=1}^T \mathbb{E}_{X_h^* \sim \mu_h^*,t}\left[\sum_{a=1}^K \left(\pi^*(a|X_h^*) - \pi_t(a|X_h^*)\right)\left\langle \varphi(X_h^*), \widetilde{\theta}_{t,a,h}\right\rangle\right].$$

The proof is presented in Appendix A.1. This suggests that we can define an auxiliary regret minimization game for every layer $h$ and every state $x$ with reward $\left\langle\varphi(x), \widehat{\theta}_{t,a,h}\right\rangle$ assigned to action $a$ in each round $t$. The regret in this auxiliary game can be written as

$$\widehat{R}_{T,h}(x) = \sum_{t=1}^T\sum_{a=1}^K\big(\pi_t^*(a|x) - \pi(a|x)\big)\big\langle\varphi(x), \widehat{\theta}_{t,a,h}\big\rangle,$$

and the above lemma suggests that the regret in layer $h$ can be simply bounded as

$$R_{T,h} \leq \mathbb{E}\left[\widehat{R}_{T,h}(X_h^*)\right] + 2\sum_{t=1}^T \max_a \left|\mathbb{E}\left[\langle\varphi(X_h^*), b_{t,a,h}\rangle\right]\right|.$$

Thus, we are left with the problem of controlling the auxiliary regret in each state, and the bias of our estimators. The following lemma, which is a straightforward application of standard ideas from the classical EXP3 analysis [8], gives bounds for the regret in the auxiliary game:

**Lemma 3.** *Fix any $h \in [H]$, $x \in \mathcal{S}_h$ and suppose that $\widehat{\theta}_{t,a,h}$ is such that $\left|\eta\big\langle\varphi(x), \widehat{\theta}_{t,a,h}\big\rangle\right| < 1$. Then, the regret in the auxiliary game at $x$ satisfies*

$$\widehat{R}_{T,h}(x) \leq \frac{\log K}{\eta} + \gamma U_T(x) + \eta\sum_{t=1}^T\sum_{a=1}^K \pi_t(a|x)\big\langle\varphi(x), \widehat{\theta}_{t,a,h}\big\rangle^2,$$

*where $U_T(x) = \sum_{t=1}^T\left(\big\langle\varphi(x), \widehat{\theta}_{t,\pi^*(x),h}\big\rangle - \frac{1}{K}\sum_{a=1}^K\big\langle\varphi(x), \widehat{\theta}_{t,a,h}\big\rangle\right)$.*

While the proof is standard, we provide it for completeness in Appendix A.2. The main term on the right-hand side of this bound is handled in the next lemma:

**Lemma 4.** *Suppose that $\varphi(X_{t,h})$ is satisfying $\|\varphi(X_{t,h})\|_2 \leq \sigma$, $0 < \beta \leq \frac{1}{2\sigma^2}$ and $M > 0$. Then for each $t$ and $h$,*

$$\mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X_h^*)\big\langle\varphi(X_h^*),\widehat{\theta}_{t,a,h}\big\rangle^2\right] \leq (H-h)^2 d\left(\frac{1}{3}+\frac{K}{\gamma}\frac{\sigma^2}{\lambda_{\min}}\right).$$

The proof of this claim is rather complicated and is presented in Appendix A.3. The main difficulty in the analysis comes from the mismatch of distribution of $X_h^*$ and $X_{t,h}$. To illustrate this difficulty, consider replacing $\widehat{\theta}_{t,a,h}$ by the ideal estimator $\widetilde{\theta}_{t,a,h}$ defined in Equation (2) in the quadratic term bounded in the above lemma. Introducing the notation $\Sigma_{t,a,h}^* = \mathbb{E}\left[\pi_t(a|X_h^*)\varphi(X_h^*)\varphi(X_h^*)^\intercal\right]$, each term in the sum can be bounded as

$$\begin{aligned}
\mathbb{E}_t\left[\pi_t(a|X_h^*)\big\langle\varphi(X_h^*),\widetilde{\theta}_{t,a}\big\rangle^2\right] &= \mathbb{E}_t\left[\pi_t(a|X_h^*)\left(\varphi(X_h^*)^\intercal\Sigma_{t,a,h}^{-1}\varphi(X_t)G_{t,h}\mathbb{I}_{\{A_{t,h}=a\}}\right)^2\right]\\
&\leq (H-h)^2\cdot\mathbb{E}_t\left[\mathrm{tr}\left(\pi_t(a|X_h^*)\varphi(X_h^*)\varphi(X_h^*)^\intercal\Sigma_{t,a,h}^{-1}\varphi(X_{t,h})\varphi(X_{t,h})^\intercal\Sigma_{t,a,h}^{-1}\mathbb{I}_{\{A_{t,h}=a\}}\right)\right]\\
&= (H-h)^2\cdot\mathrm{tr}\left(\Sigma_{t,a,h}^*\Sigma_{t,a,h}^{-1}\right).
\end{aligned}$$

Unfortunately, this latter term cannot be bounded without further assumptions on $\Sigma_{t,a,h}$ due to the mismatch between the distributions of $X_h^*$ and $X_{t,a,h}$. We address this issue by mixing the exponential-weights distribution with the uniform policy and appealing to Assumption 2, which together ensure that the smallest eigenvalue of matrix $\Sigma_{t,a,h}$ is at least $\lambda_{\min}\frac{\gamma}{K}$. This yields a bound on the operator norm of the matrix inverse $\Sigma_{t,a,h}^{-1}$, and eventually the bound of order $H^2Kd/(\gamma\lambda_{\min})$ above. The tighter bounds of Theorem 2 are derived by using a stronger assumption to bound $\mathrm{tr}\big(\Sigma_{t,a,h}^*\Sigma_{t,a,h}^{-1}\big)$—the details of these tighter bounds are presented in Appendix B.

The final element in the proof is the following lemma that bounds the bias of the estimator:

**Lemma 5.** *For $M \geq 0$, $\beta = \frac{1}{2\sigma^2}$, we have*

$$\left|\mathbb{E}_t\left[\big\langle\varphi(X_h^*),\theta_{t,a,h}-\widehat{\theta}_{t,a,h}\big\rangle\right]\right| \leq \sigma R\exp\left(-\frac{\gamma\beta}{K}\lambda_{\min}M\right).$$

The proof can be found in Appendix A.4. Putting these lemmas together and verifying that the reward estimators indeed satisfy the condition of Lemma 3 (done in Lemma 6 in Appendix A.5), we obtain the following bound on the regret in layer $h$:

$$R_{T,h} \leq 2T\sigma R\cdot\exp\left(-\frac{\gamma\beta\lambda_{\min}M}{K}\right)+2\gamma(H-h)T+\eta(H-h)^2\left(\frac{d}{3}+d\frac{K}{\gamma}\frac{\sigma^2}{\lambda_{\min}}\right)T+\frac{\log K}{\eta}.$$

Summing up the bound for all $h \in [H]$ proves Theorem 1.

## 5 Discussion

This paper is studied the problem of online learning in MDPs, merging two important lines of work on this problem concerned with linear function approximation [24, 14] and bandit feedback with adversarial rewards [31, 33, 46]. Our results are the first in this setting and not directly comparable with any previous work. Consequently, it is presently unclear if our guarantees can be substantially improved: while Theorem 2 shows that regret bounds of order $\sqrt{T}$ are indeed possible in this challenging setting, this result comes at the cost of an extremely strong assumption on the MDP dynamics. We note that this assumption on bounded likelihood ratios is essentially identical to the assumption made by Neu et al. [31, 33] in the tabular setting that the probability of visiting any state under any policy is lower bounded by $\alpha > 0$. We also point out that our approach bears many similarities to that of Abbasi-Yadkori et al. [2, 3], whose regret bounds also depend on a quantity analogous to $\lambda_{\min}$ appearing in our Theorem 1. We remain optimistic that assumptions on such problem-dependent quantities can be eventually relaxed, especially since such improvements have already been demonstrated in the tabular setting by Zimin and Neu [46], Dick et al. [19]. We believe that the techniques developed in this paper will be essential for making progress in this direction, and are particularly confident that the Matrix Geometric Resampling technique will be a part of the eventual solution.

## References


[1] Y. Abbasi-Yadkori, D. Pál, and Cs. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320. 2011.

[2] Y. Abbasi-Yadkori, P. Bartlett, K. Bhatia, N. Lazic, Cs. Szepesvári, and G. Weisz. POLITEX: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702, 2019.

[3] Y. Abbasi-Yadkori, N. Lazic, Cs. Szepesvári, and G. Weisz. Exploration-enhanced POLITEX. *arXiv preprint arXiv:1908.10479*, 2019.

[4] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *ICML 2014*, page 1638–1646, 2014.

[5] A. Agarwal, A. Krishnamurthy, J. Langford, H. Luo, and S. R. E. Open problem: First-order regret bounds for contextual bandits. In *Proceedings of the 30th Conference on Learning Theory*, pages 4–7, 2017.

[6] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

[7] P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *NIPS-18*, pages 49–56, 2006. ISBN 0-262-23253-7.

[8] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

[9] M. G. Azar, R. Munos, and H. J. Kappen. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1707–1714, 2012.

[10] M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272, 2017.

[11] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *AISTATS 2011*, pages 19–26, 2011.

[12] S. Bubeck, N. Cesa-Bianchi, and S. M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. *arXiv e-prints*, art. arXiv:1202.3079, Feb. 2012.

[13] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for Markov Decision Processes. 22(1):222–255, 1997.

[14] Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably Efficient Exploration in Policy Optimization. *arXiv e-prints*, art. arXiv:1912.05830, Dec. 2019.

[15] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

[16] V. Dani, S. M. Kakade, and T. P. Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 20*, pages 345–352. 2008.

[17] C. Dann and E. Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

[18] C. Dann, T. Lattimore, and E. Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems 30*, pages 5713–5723. 2017.

[19] T. Dick, A. Gyorgy, and C. Szepesvari. Online learning in markov decision processes with changing cost sequences. In *International Conference on Machine Learning*, pages 512–520, 2014.

[20] E. Even-Dar, S. M. Kakade, and Y. Mansour. Online Markov decision processes. *Math. Oper. Res.*, 34(3):726–736, 2009.

[21] R. Fruit, M. Pirotta, A. Lazaric, and R. Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pages 1573–1581, 2018.

[22] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, August 2010. ISSN 1532-4435.

[23] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

[24] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably Efficient Reinforcement Learning with Linear Function Approximation. *arXiv e-prints*, art. arXiv:1907.05388, July 2019.

[25] T. Jin and H. Luo. Learning adversarial mdps with bandit feedback and unknown transition. *ArXiv*, abs/1912.01192, 2019.

[26] T. Lattimore and Cs. Szepesvári. Learning with good feature representations in bandits and in RL with a generative model. *arXiv preprint arXiv:1911.07676*, 2019.

[27] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[28] G. Neu and G. Bartók. An efficient algorithm for learning with semi-bandit feedback. In *ALT 2013*, pages 234–248, 2013.

[29] G. Neu and G. Bartók. Importance weighting without importance weights: An efficient algorithm for combinatorial semi-bandits. *Journal of Machine Learning Research*, 17:1–21, 2016.

[30] G. Neu and J. Olkhovskaya. Efficient and Robust Algorithms for Adversarial Linear Contextual Bandits. *arXiv e-prints*, art. arXiv:2002.00287, Feb. 2020.

[31] G. Neu, A. György, and Cs. Szepesvári. The online loop-free stochastic shortest-path problem. pages 231–243, 01 2010.

[32] G. Neu, A. György, and Cs. Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 805–813, 2012.

[33] G. Neu, A. György, Cs. Szepesvári, and A. Antos. Online Markov decision processes under bandit feedback. volume 59, pages 1804–1812, 01 2013. doi: 10.1109/TAC.2013.2292137.

[34] B. Á. Pires and Cs. Szepesvári. Policy error bounds for model-based reinforcement learning with factored linear models. In *Conference on Learning Theory*, pages 121–151, 2016.

[35] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

[36] A. Rakhlin and K. Sridharan. BISTRO: An efficient relaxation-based method for contextual bandits. In *International Conference on Machine Learning*, pages 1977–1985, 2016.

[37] A. Rosenberg and Y. Mansour. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5478–5486, 2019.

[38] A. Sidford, M. Wang, X. Wu, L. Yang, and Y. Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems 31*, pages 5186–5196. 2018.

[39] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction (second edition)*. online draft, 2018.

[40] V. Syrgkanis, A. Krishnamurthy, and R. E. Schapire. Efficient algorithms for adversarial contextual learning, 2016.

[41] Cs. Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

[42] A. Tewari and P. L. Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In *NIPS-20*, pages 1505–1512, 2008.

[43] A. Tewari and S. A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health - Sensors, Analytic Methods, and Applications*, pages 495–517. 2017.

[44] L. F. Yang and M. Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019.

[45] H. Yao, Cs. Szepesvári, B. Pires, and X. Zhang. Pseudo-mdps and factored linear action models. 10 2014. doi: 10.1109/ADPRL.2014.7010633.

[46] A. Zimin and G. Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1583–1591. Curran Associates, Inc., 2013.

# A Omitted proofs

## A.1 The proof of Lemma 2

By Lemma 1, and since $\widetilde{\theta}_{t,a,h}$ is unbiased, we have

$$
\begin{aligned}
R_{T,h} &= \sum_{t=1}^{T} \mathbb{E}_{X_h^* \sim \mu_h^*,t} \left[ Q_{t,h}(X_h^*, \pi^*(X_h^*)) - V_{t,h}(X_h^*) \right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{X_h^* \sim \mu_h^*,t} \left[ \sum_{a=1}^{K} \left( \pi^*(a|X_h^*) - \pi_t(a|X_h^*) \right) \left\langle \varphi(X_h^*), \theta_{t,a,h} \right\rangle \right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{X_h^* \sim \mu_h^*,t} \left[ \sum_{a=1}^{K} \left( \pi^*(a|X_h^*) - \pi_t(a|X_h^*) \right) \left\langle \varphi(X_h^*), \widetilde{\theta}_{t,a,h} \right\rangle \right].
\end{aligned}
$$

□

## A.2 The proof of Lemma 3

We omit index $h$ for ease of readability. The proof follows the standard analysis of EXP3 originally due to Auer et al. [8]. We begin by recalling the notation $w_t(x,a) = \exp\left(\eta \sum_{s=1}^{t-1} \left\langle \varphi(x), \widehat{\theta}_{s,a} \right\rangle\right)$ and introducing $W_t(x) = \sum_{a=1}^{K} w_t(x,a)$. The proof is based on analyzing $\log W_{T+1}(x)$, which can be thought of as a potential function in terms of the cumulative losses. We first observe that $\log W_{T+1}(x)$ can be lower-bounded in terms of the cumulative loss:

$$
\log\left(\frac{W_{T+1}(x)}{W_1(x)}\right) \geq \log\left(\frac{w_{T+1}(x,\pi^*(x))}{W_1(x)}\right) = \eta \sum_{t=1}^{T} \varphi(x)^{\mathsf{T}} \widehat{\theta}_{t,\pi^*(x)} - \log K.
$$

On the other hand, for any $t$, we can prove the upper bound

$$
\begin{aligned}
\log \frac{W_{t+1}(x)}{W_t(x)} &= \log\left(\sum_{a=1}^{K} \frac{w_{t+1}(x,a)}{W_t(x)}\right) = \log\left(\sum_{a=1}^{K} \frac{w_t(x,a) e^{\eta\langle \varphi(x), \widehat{\theta}_{t,a}\rangle}}{W_t(x)}\right) \\
&= \log\left(\sum_{i=1}^{K} \frac{\pi_t(a|x) - \gamma/K}{1-\gamma} \cdot e^{\eta\langle \varphi(x), \widehat{\theta}_{t,a}\rangle}\right) \\
&\overset{(a)}{\leq} \log\left(\sum_{i=1}^{K} \frac{\pi_t(a|x) - \gamma/K}{1-\gamma} \left(1 + \eta\left\langle \varphi(x), \widehat{\theta}_{t,a}\right\rangle + \left(\eta\left\langle \varphi(x), \widehat{\theta}_{t,a}\right\rangle\right)^2\right)\right) \\
&\overset{(b)}{\leq} \sum_{a=1}^{K} \frac{\pi_t(a|x)}{1-\gamma} \left(\eta\left\langle \varphi(x), \widehat{\theta}_{t,a}\right\rangle + \left(\eta\left\langle \varphi(x), \widehat{\theta}_{t,a}\right\rangle\right)^2\right) - \frac{\eta\gamma}{K(1-\gamma)} \sum_{a} \left\langle \varphi(x), \widehat{\theta}_{t,a}\right\rangle,
\end{aligned}
$$

where in step $(a)$ we used the inequality $e^z \leq 1 + z + z^2$, which holds for $z < 1.79$, and in step $(b)$ we used the inequality $\log(1+z) \leq z$ that holds for any $z$. Noticing that $\sum_{t=1}^{T} \log \frac{W_{t+1}}{W_t} = \log \frac{W_{T+1}}{W_1}$, we can sum both sides of the above inequality for all $t = 1, \ldots, T$ and compare with the lower bound to get

$$
\eta \sum_{t=1}^{T} \varphi(x)^{\mathsf{T}} \widehat{\theta}_{t,\pi^*(x)} - \ln K \leq \sum_{t=1}^{T} \sum_{a=1}^{K} \frac{\pi_t(a|x)}{1-\gamma} \left(\eta\left\langle \varphi(x), \widehat{\theta}_{t,a}\right\rangle + \left(\eta\left\langle \varphi(x), \widehat{\theta}_{t,a}\right\rangle\right)^2\right) - \frac{\eta\gamma \sum_a \left\langle \varphi(x), \widehat{\theta}_{t,a}\right\rangle}{K(1-\gamma)}.
$$

Reordering and multiplying both sides by $\frac{1-\gamma}{\eta}$ gives

$$
\begin{aligned}
&\sum_{t=1}^{T} \left( \left\langle \varphi(x), \widehat{\theta}_{t,\pi^*(x)} - \sum_{a=1}^{K} \pi_t(a|x) \left\langle \varphi(x), \widehat{\theta}_{t,a}\right\rangle \right\rangle \right) \\
&\leq \frac{(1-\gamma)\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} \left(\left\langle \varphi(x), \widehat{\theta}_{t,a}\right\rangle\right)^2 + \gamma \sum_{t=1}^{T} \left( \left\langle \varphi(x), \widehat{\theta}_{t,\pi^*(x)}\right\rangle - \frac{1}{K} \sum_{a} \left\langle \varphi(x), \widehat{\theta}_{t,a}\right\rangle \right).
\end{aligned}
$$

This concludes the proof. □

### A.3 The proof of Lemma 4

The proof relies on a repeated use of the following identity that holds for any symmetric positive definite matrix $S$:

$$\sum_{k=0}^{M} (I-S)^k = S^{-1} - (I-S)^M S^{-1}.$$

For ease of readability, we will omit the indices $h$ in this section. We denote the covariance of states, generated by policy $\pi^*$ as $\Sigma_a^* = \mathbb{E}\left[\pi_t(a|X^*)\varphi(X^*)\varphi(X^*)^\intercal\right]$. We start by plugging in the definition of $\widehat{\theta}_{t,a}$ and writing

$$\begin{aligned}
&\mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X^*)\left\langle\varphi(X^*),\widehat{\theta}_{t,a}\right\rangle^2\right]\\
&\quad= \mathbb{E}_t\left[\sum_{a=1}^{K}\pi_t(a|X^*)\left(\varphi(X^*)^\intercal\widehat{\Sigma}_{t,a}^{+}\varphi(X_t)G_{t,h}\mathbb{I}_{\{A_t=a\}}\right)^2\right] \qquad (3)\\
&\quad\le (H-h)^2\cdot\mathbb{E}_t\left[\sum_{a=1}^{K}\text{tr}\left(\pi_t(a|X^*)\varphi(X^*)\varphi(X^*)^\intercal\widehat{\Sigma}_{t,a}^{+}\varphi(X_t)\varphi(X_t)^\intercal\widehat{\Sigma}_{t,a}^{+}\mathbb{I}_{\{A_t=a\}}\right)\right],
\end{aligned}$$

where we used $\left\langle X_t,\theta_{t,a}\right\rangle \le H-h$ in the inequality. Using the definition of $\Sigma_{t,a}^{+}$ and elementary manipulations, we can get

$$\begin{aligned}
&\mathbb{E}_t\left[\text{tr}\left(\pi_t(a|X^*)\varphi(X^*)\varphi(X^*)^\intercal\widehat{\Sigma}_{t,a}^{+}\varphi(X_t)\varphi(X_t)^\intercal\widehat{\Sigma}_{t,a}^{+}\mathbb{I}_{\{A_t=a\}}\right)\right]\\
&\quad= \mathbb{E}_t\left[\text{tr}\left(\Sigma_a^*\Sigma_{t,a}^{+}\Sigma_{t,a}\Sigma_{t,a}^{+}\right)\right] = \beta^2\cdot\mathbb{E}_t\left[\text{tr}\left(\Sigma_a^*\left(\sum_{k=0}^{M}C_{k,a}\right)\Sigma_{t,a}\left(\sum_{j=0}^{M}C_{j,a}\right)\right)\right]\\
&\quad= \beta^2\mathbb{E}_t\left[\sum_{k=0}^{M}\sum_{j=0}^{M}\text{tr}\left(\Sigma_a^*C_{k,a}\Sigma_{t,a}C_{j,a}\right)\right]\\
&\quad= \beta^2\mathbb{E}_t\left[\sum_{k=0}^{M}\text{tr}\left(\Sigma_a^*C_{k,a}\Sigma_{t,a}C_{k,a}\right)\right] + 2\beta^2\mathbb{E}_t\left[\sum_{k=0}^{M}\sum_{j=k+1}^{M}\text{tr}\left(\Sigma_a^*C_{k,a}\Sigma_{t,a}C_{j,a}\right)\right].
\end{aligned}$$

Let us first address the first term on the right hand side. To this end, consider any symmetric positive definite matrix $S$ that commutes with $\Sigma_{t,a}$ and observe that

$$\begin{aligned}
&\mathbb{E}_t\left[(I-\beta B_{k,a})S(I-\beta B_{k,a})\right]\\
&\quad= \mathbb{E}\left[(I-\beta\varphi(X(k))\varphi(X(k))^\intercal\mathbb{I}_{\{A(k)=a\}})S(I-\beta\varphi(X(k))\varphi(X(k))^\intercal\mathbb{I}_{\{A(k)=a\}})\right]\\
&\quad= S-\beta\mathbb{E}\left[\varphi(X(k))\varphi(X(k))^\intercal\mathbb{I}_{\{A(k)=a\}}S\right]-\beta\mathbb{E}_t\left[S\varphi(X(k))\varphi(X(k))^\intercal\mathbb{I}_{\{A(k)=a\}}\right]\\
&\quad+\beta^2\mathbb{E}_t\left[\varphi(X(k))\varphi(X(k))^\intercal S\varphi(X(k))\varphi(X(k))^\intercal\mathbb{I}_{\{A(k)=a\}}\right]\\
&\quad\preccurlyeq S-2\beta S\Sigma_{t,a}+\beta^2\sigma^2S\Sigma_{t,a} = S\left(I-\beta(2-\beta\sigma^2)\Sigma_{t,a}\right),
\end{aligned}$$

where we used our assumption that $\|\varphi(X(k))\| \le \sigma$, which implies $\mathbb{E}_t\left[\|\varphi(X(k))\|_2^2\,\varphi(X(k))\varphi(X(k))^\intercal\mathbb{I}_{\{A(k)=a\}}\right] \preccurlyeq \sigma^2\Sigma_{t,a}$. Now, recalling the definition $C_{k,a} = \prod_{j=1}^{k}(I-\beta B_{j,a})$ and using the above relation repeatedly, we can obtain

$$\begin{aligned}
\text{tr}\left(\mathbb{E}_t\left[\Sigma_a^*C_{k,a}\Sigma_{t,a}C_{k,a}\right]\right) &= \text{tr}\left(\mathbb{E}_t\left[\Sigma_a^*C_{k-1,a}\mathbb{E}_t\left[(I-\beta B_{k,a})\Sigma_{t,a}(I-\beta B_{k,a})\right]C_{k-1,a}\right]\right)\\
&\le \text{tr}\left(\mathbb{E}_t\left[\Sigma_a^*C_{k-1,a}\Sigma_{t,a}\left(I-\beta(2-\beta\sigma^2)\Sigma_{t,a}\right)C_{k-1,a}\right]\right) \qquad (4)\\
&\le \ldots \le \text{tr}\left(\Sigma_a^*\Sigma_{t,a}(I-\beta(2-\beta\sigma^2)\Sigma_{t,a})^k\right).
\end{aligned}$$

Thus, we can see that

$$\beta^2\sum_{k=0}^{M}\text{tr}\left(\mathbb{E}_t\left[\Sigma_a^*C_{k,a}\Sigma_{t,a}C_{k,a}\right]\right) = \beta^2\sum_{k=0}^{M}\text{tr}\left(\Sigma_a^*\Sigma_{t,a}(I-\beta(2-\beta\sigma^2)\Sigma_{t,a})^k\right)$$

$$= \frac{\beta^2}{\beta(2-\beta\sigma^2)} \text{tr}\left(\Sigma_a^* \Sigma_{t,a} \Sigma_{t,a}^{-1}\left(I - (I - \beta(2-\beta\sigma^2)\Sigma_{t,a})^M\right)\right) \le \frac{\beta \text{tr}\left(\Sigma_a^*\right)}{2-\beta\sigma^2} \le \frac{2\beta\text{tr}\left(\Sigma_a^*\right)}{3},$$

where we used the condition $\beta \le \frac{1}{2\sigma^2}$ and the fact that $(I - \beta(2-\beta\sigma^2)\Sigma_{t,a})^M \succcurlyeq 0$ by the same condition. We can finally observe that our assumption on the contexts implies $\text{tr}\left(\Sigma_a^*\right) \le \text{tr}\left(\sigma^2 I\right) = \sigma^2 d$, so again by our condition on $\beta$ we have $\beta\text{tr}\left(\Sigma_a^*\right) \le \frac{d}{2}$, and the first term is bounded by $\frac{d}{3}$.

Moving on to the second term, we first note that for any $j > k$, the conditional expectation of $B_{j,a}$ given $B_{\le k,a} = (B_{1,a}, B_{2,a}, \dots B_{k,a})$ satisfies $\mathbb{E}\left[C_{k,a} \middle| B_{\le k,a}\right] = C_{k,a}(I - \beta\Sigma)^{j-k}$ due to conditional independence of all $B_{j,a}$ given $B_{k,a}$, for $i > k$. We make use of this equality by writing

$$\begin{aligned}
\beta^2 \sum_{k=0}^{M} \sum_{j=k+1}^{M} \mathbb{E}\left[\text{tr}\left(\Sigma_a^* C_{k,a} \Sigma_{t,a} C_{j,a}\right)\right] &= \beta^2 \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^{M} \text{tr}\left(\Sigma_a^* C_{k,a} \Sigma_{t,a} C_{j,a}\right) \middle| B_{\le k,a}\right]\right] \\
&= \beta^2 \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\sum_{j=k+1}^{M} \text{tr}\left(\Sigma_a^* C_{k,a} \Sigma_{t,a} C_{j,az} (I - \beta\Sigma_{t,a})^{j-k}\right) \middle| B_{\le k,a}\right]\right] \\
&= \beta \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\text{tr}\left(\Sigma_a^* C_{k,a} \Sigma_{t,a} C_{k,a} \Sigma_{t,a}^{-1}\left(I - (I - \beta\Sigma_{t,a})^{M-k}\right)\right) \middle| B_{\le k,a}\right]\right] \\
&\le \beta \sum_{k=0}^{M} \mathbb{E}\left[\mathbb{E}\left[\text{tr}\left(\Sigma_a^* C_{k,a} \Sigma_{t,a} C_{k,a} \Sigma_{t,a}^{-1}\right) \middle| B_{\le k,a}\right]\right] \\
&\qquad \text{(due to } (I - \beta\Sigma_{t,a})^{M-k} \succcurlyeq 0\text{)} \\
&\le \beta \sum_{k=0}^{M} \text{tr}\left(\Sigma_a^* \Sigma_{t,a} (I - \beta(2-\beta\sigma^2)\Sigma_{t,a})^k \Sigma_{t,a}^{-1}\right) \\
&\qquad \text{(by the same argument as in Equation (4))} \\
&\le \frac{1}{(2-\beta\sigma^2)} \text{tr}\left(\Sigma_a^* \Sigma_{t,a} \Sigma_{t,a}^{-1}\left(I - (I - \beta(2-\beta\sigma^2)\Sigma_{t,a})^M \Sigma_{t,a}^{-1}\right)\right) \\
&\le \text{tr}\left(\Sigma_a^* \Sigma_{t,a}^{-1}\right) = \text{tr}\left(\Sigma_a^*\left(\Sigma_{t,a}' + \frac{\gamma}{K}\Sigma\right)^{-1}\right) \le \frac{K}{\gamma} \text{tr}\left(\Sigma_a^* \Sigma^{-1}\right),
\end{aligned}$$

where in the last line we used that $\Sigma_{t,a}$ can be written as $\Sigma_{t,a} = (1-\gamma)\Sigma_{t,a}' + \frac{\gamma}{K}\Sigma$ for $\Sigma_{t,a}' = \mathbb{E}_t\left[\varphi(X_t)\varphi(X_t)^\mathsf{T}\mathbb{I}_{\{A_t=a, Y_t=0\}}\right]$. Now, turning back to the sum over actions in (3) and recalling the definition of $\Sigma^*$, we observe that $\Sigma^* = \sum_a \Sigma_a^*$ so that we can write

$$\begin{aligned}
&\mathbb{E}_t\left[\sum_{a=1}^{K} \text{tr}\left(\pi_t(a|X^*)\varphi(X^*)\varphi(X^*)^\mathsf{T}\widehat{\Sigma}_{t,a}^{+}\varphi(X_t)\varphi(X_t)^\mathsf{T}\widehat{\Sigma}_{t,a}^{+}\mathbb{I}_{\{A_t=a\}}\right)\right] \\
&\qquad \le \frac{d}{3} + \frac{K}{\gamma}\sum_{a=1}^{K} \text{tr}\left(\Sigma_a^*\Sigma^{-1}\right) = \frac{d}{3} + \frac{K}{\gamma}\text{tr}\left(\Sigma^*\Sigma^{-1}\right) \le \frac{d}{3} + \frac{K}{\gamma}\sqrt{\text{tr}\left(\left(\Sigma^*\right)^2\right)\text{tr}\left(\left(\Sigma^{-1}\right)^2\right)} \qquad (5) \\
&\qquad \le \frac{d}{3} + \frac{Kd}{\gamma}\frac{\sigma^2}{\lambda_{\min}},
\end{aligned}$$

where we used the Cauchy–Schwarz inequality in the last step. This proves the statement. □

### A.4 The proof of Lemma 5

We first observe that the bias of $\widehat{\theta}_{t,a,h}$ can be easily expressed as

$$\begin{aligned}
\mathbb{E}_t\left[\widehat{\theta}_{t,a,h}\right] &= \mathbb{E}_t\left[\widehat{\Sigma}_{t,a,h}^{+}\varphi(X_{t,h})\varphi(X_{t,h})^\mathsf{T}\theta_{t,a,h}\mathbb{I}_{\{A_{t,h}=a\}}\right] \\
&= \mathbb{E}_t\left[\widehat{\Sigma}_{t,a,h}^{+}\right]\mathbb{E}_t\left[\varphi(X_{t,h})\varphi(X_{t,h})^\mathsf{T}\mathbb{I}_{\{A_{t,h}=a\}}\right]\theta_{t,a,h} \\
&= \mathbb{E}_t\left[\widehat{\Sigma}_{t,a,h}^{+}\right]\Sigma_{t,a,h}\theta_{t,a,h} = \theta_{t,a,h} - (I - \beta\Sigma_{t,a,h})^M\theta_{t,a,h}.
\end{aligned}$$

Thus, the bias is bounded as

$$\left|\mathbb{E}_t\left[\varphi(X_h^*)^\intercal (I-\beta\Sigma_{t,a,h})^M\theta_{t,a,h}\right]\right| \le \|\varphi(X_h^*)\|_2 \cdot \|\theta_{t,a,h}\|_2 \left\|(I-\beta\Sigma_{t,a,h})^M\right\|_{\text{op}}.$$

In order to bound the last factor above, observe that $\Sigma_{t,a,h} \succcurlyeq \frac{\gamma}{K}\Sigma_h$ due to the uniform exploration used in the first layer by MDP-LinExp3, which implies that

$$\left\|(I-\beta\Sigma_{t,a,h})^M\right\|_{\text{op}} \le \left(1-\frac{\gamma\beta\lambda_{\min}}{K}\right)^M \le \exp\left(-\frac{\gamma\beta}{K}\lambda_{\min}M\right),$$

where the second inequality uses $1-z \le e^{-z}$ that holds for all $z$. This concludes the proof. □

### A.5 The boundedness of the estimates

**Lemma 6.** *The loss estimates satisfy* $\eta\left|\left\langle\varphi(X_h^*), \widehat{\theta}_{t,a,h}\right\rangle\right| < 1$ *for* $\eta \le \frac{2}{H(M+1)}$ .

*Proof.* The claim is proven by the following straightforward calculation:

$$\begin{aligned}
\eta \cdot \left|\left\langle\varphi(X_h^*), \widehat{\theta}_{t,a,h}\right\rangle\right| &= \eta \cdot \left|\varphi(X_h^*)^\intercal \widehat{\Sigma}^+_{t,a,h}\varphi(X_{t,h})\left\langle\varphi(X_{t,h}), \theta_{t,a,h}\right\rangle \mathbb{I}_{\{A_t=a\}}\right| \\
&\le \eta(H-h)\cdot\left|\varphi(X_h^*)^\intercal\widehat{\Sigma}^+_{t,a,h}\varphi(X_{t,h})\right| \le \eta(H-h)\sigma^2\left\|\widehat{\Sigma}^+_{t,a,h}\right\|_{\text{op}} \\
&\le \eta(H-h)\sigma^2\beta\left(1+\sum_{k=1}^M \|C_{k,a,h}\|_{\text{op}}\right) \le \eta(H-h)(M+1)/2,
\end{aligned}$$

where we used the fact that our choice of $\beta$ ensures $\|C_{k,a,h}\|_{\text{op}} = \left\|\prod_{j=0}^k (I-\beta B_{j,a,h})\right\|_{\text{op}} \le 1.$ □

## B Proof of Theorem 2

The improvement in the regret bound comes from applying an importance-weighting trick in the proof of Lemma 4 to bound the problematic term $\text{tr}\left(\Sigma^*_{a,h}\Sigma^{-1}_{t,a,h}\right)$. Specifically, we write

$$\begin{aligned}
\text{tr}\left(\Sigma^*_{a,h}\Sigma^{-1}_{t,a,h}\right) &= \text{tr}\left(\mathbb{E}_t\left[\pi_t(a|X_h^*)\varphi(X_h^*)\varphi(X_h^*)^\intercal\right]\Sigma^{-1}_{t,a,h}\right) \\
&= \text{tr}\left(\mathbb{E}_t\left[\frac{f_h^*(X_{t,h})}{f_h^{\pi_t}(X_{t,h})}\pi_t(a|X_{t,h})\varphi(X_{t,h})\varphi(X_{t,h})^\intercal\right]\Sigma^{-1}_{t,a,h}\right) \\
&\le \rho\cdot\text{tr}\left(\mathbb{E}_t\left[\pi_t(a|X_{t,h})\varphi(X_{t,h})\varphi(X_{t,h})^\intercal\right]\Sigma^{-1}_{t,a,h}\right) = \rho d,
\end{aligned}$$

where we used our assumption on the likelihood ratio in the inequality. Using this bound instead of the one in Equation (5) at the end of the proof of Lemma 4 yields the improved bound

$$\mathbb{E}_t\left[\sum_{a=1}^K \pi_t(a|X_h^*)\left\langle\varphi(X_h^*), \widehat{\theta}_{t,a,h}\right\rangle^2\right] \le (H-h)^2 d\left(\frac{1}{3}+\rho K\right).$$

The proof of Theorem 2 is then concluded similarly as the proof of Theorem 1. □

## C Fast Matrix Geometric Resampling

The naïve implementation of the MGR procedure presented in the main text requires $O(MKHd + KHd^2)$ time due to the matrix-matrix multiplications involved. In this section we explain how to compute $\widehat{\theta}_{t,a}$ in $O(MKHd)$ time, exploiting the fact that the matrices $\widehat{\Sigma}_{t,a,h}$ never actually need to be computed, since the algorithm only works with products of the form $\widehat{\Sigma}_{t,a,h}\varphi(X_{t,h})$ for vectors $X_{t,h}$, $h \in [H]$. This motivates the following procedure:

**Fast Matrix Geometric Resampling**

---

**Input:** simulator of transition function $P$, policy $\pi_t$, trajectory $(x_1, a_1, x_2, a_2, \ldots, x_H, a_H)$

**Initialization:** Compute $Y_{0,h} = \varphi(x_h)$ for all $h \in [H]$.

**For** $k = 1, \ldots, M$**, repeat**:

1. Generate a path $U(i) = \{(X_1(i), A_1(i)), \ldots, (X_H(i), A_H(i))\}$, following the policy $\pi_t$ in the simulator of $P$,
2. **For** $h = 1, \ldots, H$**, repeat**:
   (a) if $A_h(k) = a_h$, set $Y_{k,h} = Y_{k-1,h} - \beta \langle Y_{k-1,h}, \varphi(X_h(k)) \rangle \varphi(X_h(k))$,
   (b) otherwise, set $Y_{k,h} = Y_{k-1,h}$.

**Return** $q_{t,a,h} = \beta Y_{0,h} + \beta \sum_{k=1}^{M} Y_{k,h}$ for all $h \in [H]$.

It is easy to see from the above procedure that each iteration $k$ can be computed using $(K + 1)Hd$ vector-vector multiplications: sampling each action $A_h(k)$ takes $Kd$ time due to having to compute the products $\left\langle \varphi(X_h(k)), \sum_{s=1}^{t-1} \widehat{\theta}_{s,a,h} \right\rangle$ for each action $a$, and updating $Y_{k,h}$ can be done by computing the product $\langle Y_{k-1,h}, \varphi(X_h(k)) \rangle$. Overall, this results in a total runtime of order $MKHd$ as promised above.