

---

# Optimistic Planning by Regularized Dynamic Programming

---

Antoine Moulin<sup>1</sup> Gergely Neu<sup>1</sup>

## Abstract

We propose a new method for optimistic planning in infinite-horizon discounted Markov decision processes based on the idea of adding regularization to the updates of an otherwise standard approximate value iteration procedure. This technique allows us to avoid contraction and monotonicity arguments that are typically required by existing analyses of approximate dynamic programming methods, and in particular to use approximate transition functions estimated via least-squares procedures in MDPs with linear function approximation. We use our method to provide a computationally efficient algorithm for learning near-optimal policies in discounted linear kernel MDPs from a single stream of experience, and show that it achieves near-optimal statistical guarantees.

## 1. Introduction

The idea of constructing a confidence set of statistically plausible models and picking a policy that maximizes the expected return in the best of these models can be traced back to the pioneering work of [Lai & Robbins \(1985\)](#) in the context of multi-armed bandit problems, and has been successfully extended to address the exploration-exploitation dilemma in reinforcement learning (RL, [Sutton & Barto, 2018](#)). This popular design principle came to be known as *optimism in the face of uncertainty*, and the associated optimization task as the problem of *optimistic planning*. The optimistic principle has driven the development of statistically efficient reinforcement learning algorithms for a variety of problem settings. Following the early work of [Brafman & Tennenholtz \(2002\)](#); [Strehl et al. \(2009\)](#) on optimistic exploration methods for reinforcement learning in Markov decision processes (MDPs), a big breakthrough was achieved by [Jaksch, Ortner, and Auer \(2010\)](#), whose UCRL2 algorithm was shown to achieve near-optimal regret guaran-

tees in a broad class of tabular MDPs. In subsequent years, their work has inspired an impressive amount of follow-up work, leading to a variety of extensions, improvements, and other mutations.

The computational efficiency of such optimistic methods crucially hinges on the implementation of the optimistic planning subroutine. In the work of [Jaksch et al. \(2010\)](#), this was addressed by a procedure called *extended value iteration* (EVI), which performs dynamic programming in an auxiliary MDP where the confidence set of models is projected to the space of actions, allowing the realization of arbitrary transitions that are statistically plausible given all past experience. After mild adjustments, the EVI procedure can be shown to give near-optimal solutions to the optimistic planning problem in a computationally efficient manner (cf. [Fruit et al., 2018](#) and Section 38.5.2 in [Lattimore & Szepesvári, 2020](#)). Other, even more effective optimistic dynamic programming procedures have been proposed and analyzed over the years ([Fruit et al., 2018](#); [Qian et al., 2018](#)). However, all of these computational developments have been largely restricted to the relatively simple tabular setting.

In recent years, the RL theory literature has seen a massive revival largely due to the breakthrough achieved by [Jin, Yang, Wang, and Jordan \(2020\)](#), who successfully extended the idea of optimistic exploration to a class of large-scale MDPs using linear function approximation. While extremely influential, their approach (and virtually all of its numerous follow-ups) are limited to the relatively simple setting of *finite-horizon MDPs*. The reason for this limitation is inherent in their algorithm design that crucially uses the fact that optimistic planning in finite-horizon MDPs can be solved via a simple backward recursion over the time indices within each episode ([Neu & Pike-Burke, 2020](#)). This idea completely fails for infinite-horizon problems where dynamic programming methods should aim to approximate the solution of a *fixed-point equation*. Solving such fixed-point equations is possible in the tabular case but no known efficient method exists for linear function approximation, the short reason being that the least-squares transition estimator used in the construction of [Jin et al. \(2020\)](#) cannot be straightforwardly used to build a approximate Bellman operator that satisfies the necessary contraction properties.

---

<sup>1</sup>Universitat Pompeu Fabra, Barcelona, Spain. Correspondence to: Antoine Moulin <antoine.moulin@upf.edu>, Gergely Neu <gergely.neu@gmail.com>.

The best attempt at attacking the infinite-horizon setting under function approximation we are aware of is by Wei, Jahromi, Luo, and Jain (2021), who propose a set of algorithms that are either statistically or computationally efficient, but eventually fall short of providing an algorithm with both of these desired properties. Another good contribution was made by Vial, Parulekar, Shakkottai, and Srikant (2022), who provided approximate DP methods for stochastic shortest path problems with linear transition functions, and analyzed them via studying the concentration properties of the empirical transition operator. This technique did allow them to prove regret bounds, but the guarantees did not reach optimality in terms of scaling with the time horizon unless strong assumptions are made. Notably, Vial et al. (2022) only managed to perform a tight analysis in the special case where the features are orthogonal, which allowed them to reason about contraction properties of the empirical Bellman operator. Lacking a general contraction argument, or another idea that would enable computationally efficient optimistic planning, efficient exploration-exploitation in infinite-horizon MDPs under function approximation has remained unsolved so far.

This is the problem we address in this paper in the context of *discounted* infinite-horizon Markov decision processes. Instead of relying on a contraction argument (or an approximate version thereof), we propose to solve the infinite-horizon optimistic planning problem using *regularized dynamic programming*. In particular, we consider a variant of the Mirror-Descent Modified Policy Iteration (MD-MPI) algorithm of Geist, Scherrer, and Pietquin (2019) that uses a least-squares estimator of the transition kernel and an exploration bonus to define an optimistic regularized Bellman operator. Using arguments from the classic analysis of mirror descent methods, we show that each application of this optimistic operator improves the quality of the policy up to an additive error term that telescopes over the iterations. In other words, we show that each iteration improves over the last one in an *average* sense. This is in stark contrast to arguments used for analyzing previous optimistic planning methods that relied on contraction arguments which guarantee *strict* improvements to the policy in each iteration. The advantage of our method is that it remains applicable even when the approximate dynamic programming operator is not contractive or monotone (even approximately).

Our concrete contribution is applying the above scheme to discounted linear kernel MDPs and showing that it achieves a near-optimal regret bound of order  $\sqrt{(B^2 dH + d^2 H^3 + \log |\mathcal{A}| H^4) T}$ , where  $d$  is the feature dimension,  $B$  is a bound on the norm of the features, and  $H = \frac{1}{1-\gamma}$  is the effective horizon. This result implies that our algorithm produces an  $\varepsilon$ -optimal policy after about  $(B^2 dH + d^2 H^3 + \log |\mathcal{A}| H^4) / \varepsilon^2$  iterations. Each policy update takes  $\text{poly}(d, H, T)$  iterations of regularized dynamic

programming, each consisting of  $\text{poly}(d, H, T)$  elementary operations. This is to be contrasted with previous contributions on a similar<sup>1</sup> setting by Zhou, He, and Gu (2021), whose policy updates rely on a version of EVI adapted to linear function approximation. Their EVI variants require globally constraining the model parameters in a way that the model is a valid transition kernel. While this last constraint allowed them to reason about contractive properties of the EVI iterations, it is practically impossible to enforce without making very strong assumptions on the feature maps and the MDP itself. The difficulty remains even when the property is only required to hold locally in each state. In this sense, our method is the first to obtain near-optimal statistical rates while also being entirely computationally feasible.

The rest of this paper is organized as follows. After providing some general notations at the end of this section and all the necessary background in Section 2, we introduce our algorithmic framework in Section 3, instantiate it in the context of linear kernel MDPs, and provide performance guarantees in this setting. We explain the key steps of the analysis in Section 4 and conclude in Section 5 with a detailed discussion of our contribution along with its limitations.

**Notations.** For a natural number  $N > 0$ , we denote  $[N] = \{1, 2, \dots, N\}$ . For a real number  $M$ , we define the truncation operator  $\Pi_M$  that acts on functions  $f$  defined on a domain  $A$  via  $\Pi_M f : x \mapsto \max(\min[f(x), M], 0)$ . For a measurable space  $(A, \mathcal{F})$ , we define the set of all probability distributions  $\Delta(A)$ , and for any two distributions  $P, Q \in \Delta(A)$  such that  $P \gg Q$ , we define the relative entropy as  $\mathcal{D}_{\text{KL}}(P||Q) = \mathbb{E}_{a \sim P} \left[ \ln \left( \frac{dP}{dQ}(a) \right) \right]$ . For a distribution  $P \in \Delta(A)$  and a bounded function  $f \in \mathbb{R}^A$ , we write  $\langle P, f \rangle = \mathbb{E}_{a \sim P} [f(a)]$  to denote the expectation of  $f$  under  $P$ , and we will use the same notation for finite-dimensional vector spaces to denote inner products. For a finite-dimensional vector  $v \in \mathbb{R}^d$  and a square matrix  $Z \in \mathbb{R}^{d \times d}$ , we will use the notation  $\|v\|_Z = \sqrt{\langle v, Zv \rangle}$ .

## 2. Preliminaries

We consider a discounted MDP  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, r, P, \gamma, \nu_0)$ , where  $\mathcal{X}$  is the finite state space<sup>2</sup>,  $\mathcal{A}$  is the finite action space,  $r : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$  is the deterministic reward function assumed to be known<sup>3</sup>,  $P : \mathcal{X} \times \mathcal{A} \rightarrow \Delta(\mathcal{X})$  is the

<sup>1</sup>We provide a detailed discussion about the differences between our settings in Section 5.

<sup>2</sup>Our results extend to the case where  $\mathcal{X}$  is a measurable space. The precise definitions require measure-theoretic concepts (Bertsekas & Shreve, 1996). For the sake of readability and because they are well understood, we only consider finite state spaces here.

<sup>3</sup>It is a standard assumption, and removing it only costs a constant factor in the regret (Jaksch et al., 2010).

transition probability distribution,  $\gamma \in (0, 1)$  is the discount factor, and  $\nu_0 \in \Delta(\mathcal{X})$  is the initial state distribution. The model describes a sequential interaction scheme between a decision-making *agent* and its environment, where the following steps are repeated for a sequence of rounds  $t = 1, 2, \dots$  after the initial state is drawn as  $X_0 \sim \nu_0$ : the agent observes the state  $X_t \in \mathcal{X}$ , selects an action  $A_t \in \mathcal{A}$ , obtains reward  $r(X_t, A_t)$ , and the environment generates the next state  $X_{t+1} \sim P(\cdot|X_t, A_t)$ . The goal of the agent is to pick its sequence of actions in a way that the total discounted return  $\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t)$  is as large as possible.

Below we describe the most fundamental objects relevant to our work, and refer the reader to the classic book of [Puterman \(2014\)](#) for more context and details. A (stationary) policy is a mapping  $\pi : \mathcal{X} \rightarrow \Delta(\mathcal{A})$  from a state to a probability measure over actions. The value function and action-value function of a policy  $\pi$  are respectively defined as the functions  $V^\pi \in \mathbb{R}^{\mathcal{X}}$  and  $Q^\pi \in \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$  mapping each state  $x$  and state-action pair  $x, a$  to

$$V^\pi(x) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r(X_t, A_t) \middle| X_0 = x \right],$$

$$Q^\pi(x, a) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r(X_t, A_t) \middle| X_0 = x, A_0 = a \right].$$

Here,  $\mathbb{E}_\pi$  denotes the expectation with respect to the probability measure  $\mathbb{P}_\pi$ , generated by the interaction between the environment and the policy  $\pi$ . With some abuse of notation, we define the transition operator  $P : \mathbb{R}^{\mathcal{X}} \rightarrow \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$  as  $(Pf)(x, a) = \sum_{x' \in \mathcal{X}} P(x'|x, a) f(x')$ , for  $f \in \mathbb{R}^{\mathcal{X}}$ . Its adjoint  $P^\top$  acts on distributions  $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$  as  $(P^\top \mu)(x') = \sum_{x, a \in \mathcal{X} \times \mathcal{A}} P(x'|x, a) \mu(x, a)$ . With these notations, we can simply state the *Bellman equations* tying together the value function  $V^\pi(x) = \mathbb{E}_{a \sim \pi(a|x)} [Q^\pi(x, a)]$  and the action-value function as

$$Q^\pi = r + \gamma P V^\pi.$$

We also introduce the operator  $E : \mathbb{R}^{\mathcal{X}} \rightarrow \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$  acting on functions  $f \in \mathbb{R}^{\mathcal{X}}$  via the assignment  $(Ef)(x, a) = f(x)$ , and its adjoint via its action  $E^\top \mu(x) = \sum_a \mu(x, a)$  on distributions  $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$ .

In a discounted MDP, a policy  $\pi$  induces a unique *normalized discounted occupancy measure* over the state space, defined for any state  $x \in \mathcal{X}$  as

$$\nu^\pi(x) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi [X_t = x].$$

The normalization term  $(1 - \gamma)$  guarantees that  $\nu^\pi$  is a probability measure over the state space. We call the inverse of this normalization constant the *effective horizon* and denote it by  $H = \frac{1}{1 - \gamma}$ . We also define the associated state-action occupancy measure  $\mu^\pi$ , defined as

$\mu^\pi(x, a) = \nu^\pi(x) \pi(a|x)$ . State-action occupancy measures are known to satisfy the following recurrence relation that is sometimes called the system of *Bellman flow constraints*:

$$E^\top \mu^\pi = \gamma P^\top \mu^\pi + (1 - \gamma) \nu_0. \quad (1)$$

Using the state-action occupancy measure, the discounted return of a policy can be written as  $R_\gamma^\pi = \frac{1}{1 - \gamma} \langle \mu^\pi, r \rangle$ . We will use  $\mu^*$  to denote an occupancy measure with maximal return and  $\nu^* = E^\top \mu^*$  to denote the associated state-occupancy measure. Finally, given two policies  $\pi, \pi'$ , we denote  $\mathcal{D}_{\text{KL}}(\pi \| \pi') = (\mathcal{D}_{\text{KL}}(\pi(\cdot|x) \| \pi'(\cdot|x)))_{x \in \mathcal{X}}$ , and we define  $\mathcal{H}(\pi \| \pi') = \langle \nu^\pi, \mathcal{D}_{\text{KL}}(\pi \| \pi') \rangle$ , the relative conditional entropy<sup>4</sup>.

In this paper, we will consider the setting of online learning in discounted MDPs, where the agent aims to produce an  $\varepsilon$ -optimal policy  $\pi_{\text{out}}$  satisfying  $\langle \mu^* - \mu^{\pi_{\text{out}}}, r \rangle \leq \varepsilon$  based on a single stream of experience in the MDP. We will assume that the learner has access to a reset action that drops the agent back to a state randomly drawn from the initial-state distribution  $\nu_0$ , and that the learner follows a stationary policy  $\pi_t$  in each round  $t$ . We will measure the performance in terms of the number of samples necessary to guarantee that the output policy is  $\varepsilon$ -optimal. As an auxiliary performance measure, we will also consider the *regret*<sup>5</sup> of the learner defined as

$$\text{regret}_T = \sum_{t=1}^T (\langle \mu^* - \mu^{\pi_t}, r \rangle).$$

It is easy to see that a regret bound can be converted into sample complexity guarantees. In particular, selecting a time index  $I$  uniformly at random from  $1, \dots, T$  and returning  $\pi_{\text{out}} = \pi_I$  guarantees that

$$\mathbb{E}[\langle \mu^* - \mu^{\pi_{\text{out}}}, r \rangle] = \frac{\text{regret}_T}{T},$$

which can be made arbitrarily small if  $\text{regret}_T$  grows sub-linearly and  $T$  is set large enough. We note here that, while superficially similar to the discounted regret criterion considered in earlier works like [Liu & Su \(2020\)](#); [He et al. \(2021\)](#) or [Zhou et al. \(2021\)](#), there are some major differences between our objectives. We only point out here that we consider the complexity of producing a good policy to execute from the initial state distribution, whereas theirs

<sup>4</sup>Technically, this is the conditional relative entropy between the *occupancy measures*  $\mu^\pi$  and  $\mu^{\pi'}$ , but we will keep referring to it in terms of the policies to keep our notations light. We refer to [Neu et al. \(2017\)](#) for further discussion.

<sup>5</sup>This is probably closer to the notion of *pseudo-regret* in the terminology of [Bubeck & Cesa-Bianchi \(2012\)](#); [Lattimore & Szepesvári \(2020\)](#), but the distinction cannot be made so clearly without ambiguity in the MDP setting. We will keep things simple by writing “regret” in the rest of the paper.

measures the suboptimality of the policies along the trajectory traversed by the learner. We defer a further discussion of the two settings to Section 5.

For our technical results, we will focus on a class of MDPs commonly referred to as *linear kernel MDPs* formally defined as follows.

**Assumption 2.1** (Linear kernel MDP). There exist a known feature map  $\psi : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \rightarrow \mathbb{R}^d$ , and an unknown  $\theta \in \mathbb{R}^d$  with  $\|\theta\|_2 \leq B$  such that  $P(x'|x, a) = \sum_{i=1}^d \theta_i \psi_i(x, a, x')$ . Furthermore, for any  $(x, a) \in \mathcal{X} \times \mathcal{A}$ ,  $V \in [0, H]^{\mathcal{X}}$ ,

$$\left\| \sum_{x' \in \mathcal{X}} \psi(x, a, x') V(x') \right\|_2 \leq BH.$$

While remotely related to the recently successful notion of linear MDPs introduced by Jin et al. (2020) and Yang & Wang (2019), linear kernel MDPs are a distinct class of models that cannot be captured in that framework. Linear kernel MDPs are also known as linear mixture MDPs (cf. Modi et al., 2020 and Ayoub et al., 2020), and have been studied almost as widely in the past few years as linear MDPs—we refer to Zhou et al. (2021) for further discussion.

### 3. Algorithm and main result

Our approach implements the principle of optimism in the face of uncertainty in discounted Markov decision processes. Instead of aiming to solve an optimistic version of the Bellman optimality equations via extended value iteration as done by Jaksch et al. (2010), our method draws on techniques from convex optimization aiming at *average policy improvement*. In particular, our planning procedure is based on a *regularized* version of approximate value iteration and incorporates an *optimistic* estimate of the associated Bellman operator. Consequently, we refer to our algorithm as **RAVI-UCB**, standing for Regularized Approximate Value Iteration with Upper Confidence Bounds.

**RAVI-UCB** performs a sequence of regularized Q-functions as follows. Starting with an initial estimate  $V_0 = 0$  and an initial policy  $\pi_0$ , it calculates a sequence of updates for  $k = 1, \dots, K$  as

$$Q_{k+1}(x, a) = \Pi_H \left[ r(x, a) + \text{CB}_k(x, a) + \gamma (\widehat{P}V_k)(x, a) \right],$$

$$V_{k+1}(x) = \frac{1}{\eta} \log \left( \sum_a \pi_k(a|x) e^{\eta Q_{k+1}(x, a)} \right),$$

$$\pi_{k+1}(a|x) = \frac{\pi_k(a|x) e^{\eta Q_{k+1}(x, a)}}{\sum_{a'} \pi_k(a'|x) e^{\eta Q_{k+1}(x, a')}}.$$

Here,  $\widehat{P}$  is a nominal transition model and  $\text{CB}_k$  is an exploration bonus defined to be large enough to ensure that

$\gamma \widehat{P}V_k + \text{CB}_k \geq \gamma PV_k$  and so that  $Q_{k+1}$  is an upper bound on the regularized Bellman update  $r + \gamma PV_k$ . The Q-functions are truncated to the range  $[0, H]$  to make sure that the optimistic property above can be ensured by setting a reasonably sized exploration bonus  $\text{CB}_k$ . It is important to note that  $Q_{k+1}$  does not directly attempt to approximate the optimal action-value function  $Q^*$  in the true MDP, which marks a clear departure from previously known optimism-based regret analyses. Instead, our analysis will show that  $(1 - \gamma) \langle \nu_0, V_k \rangle$  acts as an optimistic estimate of the optimal return  $(1 - \gamma) \langle \nu_0, V^* \rangle$  in an “average” sense, and that the total reward of our algorithm can also be bounded in terms of the same quantity.

The overall procedure is presented for the concrete setting of linear kernel MDPs as Algorithm 1. The algorithm proceeds in a sequence of *epochs*  $k = 1, 2, \dots$ , where a new epoch is started by taking the reset action with probability  $1 - \gamma$  in each round, which results in epochs of average length  $H = \frac{1}{1-\gamma}$ . At the beginning of each epoch, we update the model estimate via a least-squares regression and perform one step of online mirror descent to produce the new value estimate as well as the associated policy. This value estimate is used to compute a feature map and the bonuses. Finally, we apply the approximate Bellman operator to produce a state-action value estimate.

We highlight that the assignments in the planning subroutine are only made symbolically for all  $x, a$ , and a practical implementation will not need to loop over the entire state-action space. Rather, all quantities of interest can be computed on demand while executing the policy in runtime. As often assumed in the related literature, we further suppose that the sum  $\sum_{x'} \psi(x, a, x') V_k(x')$  can be computed (or approximated) efficiently, for instance by requiring  $\psi(x, a, \cdot)$  to have sparse support. We provide a more detailed discussion of all such computational matters in Section 5.

Our main technical result regarding the performance **RAVI-UCB** is the following regret bound.

**Theorem 3.1.** *Suppose that **RAVI-UCB** is run with the uniform policy as  $\pi_0$ ,  $V_0 = 0$ ,  $\lambda = 1$ , learning rate  $\eta = \sqrt{2(1-\gamma)^2 \log |\mathcal{A}|/T}$ , and exploration parameter  $\beta = H \sqrt{2 \left( \frac{d}{2} \log \left[ 1 + \frac{TB^2H^2}{d} \right] + \log T \right)} + B$ . Then, the expected regret of **RAVI-UCB** satisfies*

$$\mathbb{E}[\text{regret}_T] = \tilde{O} \left( \sqrt{(d^2H^3 + B^2dH + H^4 \log |\mathcal{A}|)T} \right)$$

In the above statement,  $\tilde{O}(\cdot)$  hides logarithmic factors of  $T$ ,  $B$ ,  $d$ , and  $H$ . A perhaps more useful result is the following that can be easily derived from the above regret bound via an online-to-batch conversion. In particular, suppose that **RAVI-UCB** returns a policy  $\pi_{\text{out}} = \pi_U$  with  $U$  being an

**Algorithm 1 RAVI-UCB** for linear kernel MDPs.

**Inputs:** Horizon  $T$ , learning rate  $\eta > 0$ , confidence parameter  $\beta > 0$ , regularization parameter  $\lambda$ , initial value  $V_0$ , initial policy  $\pi_0$ .

**Initialize:**  $t = 1$ ,  $\Lambda_1 = \lambda I$ ,  $b_1 = 0$ ,  $Q_1 = EV_0$ .

**for**  $k = 1, \dots$  **do**

$T_k = t$ .

$\hat{\theta}_k = \Lambda_{T_k}^{-1} b_{T_k}$ .

$\hat{P}_k = \sum_i \hat{\theta}_{k,i} \psi_i$ .

$V_k(x) = \frac{1}{\eta} \log \left( \sum_a \pi_{k-1}(a|x) e^{\eta Q_k(x,a)} \right)$ .

$\pi_k(a|x) = \pi_{k-1}(a|x) e^{\eta(Q_k(x,a) - V_k(x))}$ .

$\varphi_k(x, a) = \sum_{x'} \psi(x, a, x') V_k(x')$ .

$\text{CB}_k(x, a) = \beta \|\varphi_k(x, a)\|_{\Lambda_{T_k}^{-1}}$ .

$Q_{k+1} = \Pi_H \left[ r + \text{CB}_k + \gamma \hat{P}_k V_k \right]$ .

**repeat**

Play  $a_t \sim \pi_k(\cdot|x_t)$ .

Observe  $r(x_t, a_t)$  and  $x_{t+1}$ .

Update  $\Lambda_{t+1} = \Lambda_t + \varphi_k(x_t, a_t) \varphi_k(x_t, a_t)^\top$ .

Update  $b_{t+1} = b_t + \varphi_k(x_t, a_t) V_k(x_{t+1})$ .

$t = t + 1$ .

With probability  $1 - \gamma$ , reset to initial distribution:

$x_t \sim \nu_0$  and **break**.

**until**  $t = T$

**end for**

epoch index chosen uniformly at random from the range of epochs. The following corollary provides a guarantee on the quality of this policy.

**Corollary 3.2.** *Let  $\varepsilon > 0$ . Then, **RAVI-UCB** run with the same parameters as above outputs a policy  $\pi_{out}$  satisfying  $\mathbb{E}[\langle \mu^* - \mu^{\pi_{out}}, r \rangle] \leq \varepsilon$  after  $T_\varepsilon$  steps of interaction in the MDP with*

$$T_\varepsilon = \tilde{O} \left( \frac{B^2 d H + d^2 H^3 + H^4 \log |\mathcal{A}|}{\varepsilon^2} \right).$$

The expectation appearing in the first statement is with respect to the random transitions in the MDP and the epoch scheduling, whereas the expectation in the second one is also with respect to the random choice of the policy. It is possible to remove the former expectation at the price of a slightly more involved analysis, but the expectation over the randomization of the output policy is inherent to the online-to-batch conversion process used by our analysis. We will return to this point in Section 5.

## 4. Analysis

We now present the proof of our main result, Theorem 3.1. In particular, we state a sequence of lemmas whose combination will yield the complete proof. We will provide the proofs that we believe to be most insightful in the main text,

and relegate the more technical ones to Appendix A.

The analysis will be split into two main parts: one pertaining to the general properties of our optimistic planning procedure and to the eventual regret bound that can be derived from it, and one concerning the specifics of the linear kernel MDPs. In particular, we first analyze **RAVI-UCB** using a generic exploration bonus that we will suppose to be “valid” in an appropriate sense, and then show how to derive such valid exploration bonuses in the concrete setting we consider.

### 4.1. Optimistic Planning

We first study the properties of our optimistic planning procedure, without making explicit references to the linear kernel MDP model. For this general analysis, we will fix an epoch index  $k$ , assume that  $\hat{P}_k$  is some estimator of the transition kernel  $P$  and that the exploration bonus  $\text{CB}_k$  satisfies the following condition for all  $x, a$

$$\left| \langle \gamma P(\cdot|x, a) - \gamma \hat{P}_k(\cdot|x, a), V_k \rangle \right| \leq \text{CB}_k(x, a), \quad (2)$$

where  $V_k$  is the value-function estimate produced at the beginning of epoch  $k$ . We will refer to exploration bonuses satisfying the above condition as *valid*.

Our key result regarding the quality of the policies produced by **RAVI-UCB** is the following.

**Lemma 4.1.** *Let  $K$  be a fixed number of epochs, and let  $\pi_k$  and  $\text{CB}_k$  be the policy and exploration bonus produced by **RAVI-UCB** in epoch  $k$ , where the input is  $V_0 = 0$  and any policy  $\pi_0$ . Suppose that  $\{\text{CB}_k\}_{k \in [K]}$  is a sequence of valid exploration bonuses in the sense of Equation (2). Then, the sequence  $\{\pi_k\}_{k \in [K]}$  satisfies the following bound:*

$$\begin{aligned} \sum_{k=1}^K (\langle \mu^*, r \rangle - \langle \mu^{\pi_k}, r \rangle) &\leq 2 \sum_{k=1}^K \langle \mu^{\pi_k}, \text{CB}_k \rangle + 2H \\ &\quad + \frac{1}{\eta} \mathcal{H}(\pi^* \| \pi_0) + \frac{\eta H^3 K}{2}. \end{aligned}$$

*Proof.* The main idea of the proof is to show that, under the validity condition of the exploration bonuses,  $(1 - \gamma) \langle \nu_0, V_k \rangle$  acts as an approximate upper bound on the optimal return  $\langle \mu^*, r \rangle$ , up to some additional terms resulting from the use of incremental updates. Thanks to the use of regularization, we can show that these additional terms are small on average, and that the gap between the optimistic value and the return of  $\pi_k$  can be bounded in terms of  $\langle \mu^{\pi_k}, \text{CB}_k \rangle$ . With this in mind, we begin by rewriting the performance gap of the output policy as follows:

$$\sum_{k=1}^K (\langle \mu^*, r \rangle - \langle \mu^{\pi_k}, r \rangle) = \sum_{k=1}^K (\Delta_k^* + \Delta_k),$$

where we defined  $\Delta_k^* = \langle \mu^*, r \rangle - (1 - \gamma) \langle \nu_0, V_k \rangle$  and  $\Delta_k = (1 - \gamma) \langle \nu_0, V_k \rangle - \langle \mu^{\pi_k}, r \rangle$  for all  $k$ .

Let us now fix some  $k$  and consider the first term,  $\Delta_k^*$ . We start by observing that  $(1 - \gamma) \nu_0 = E^\top \mu^* - \gamma P^\top \mu^*$ , which allows us to write

$$\begin{aligned} \Delta_k^* &= \langle \mu^*, r \rangle - (1 - \gamma) \langle \nu_0, V_k \rangle \\ &= \langle \mu^*, r + \gamma PV_k \rangle - \langle \mu^*, EV_k \rangle. \end{aligned} \quad (3)$$

In order to treat the first term, we will now show that the inequality  $Q_{k+1} \geq r + \gamma PV_k$  holds. For each state-action pair  $(x, a)$ , we need to handle two separate cases corresponding to whether or not  $Q_{k+1}(x, a)$  is truncated from above in the  $k^{\text{th}}$  epoch. In the first case, we have  $Q_{k+1}(x, a) = H$ , which implies

$$Q_{k+1}(x, a) = H = 1 + \gamma H \geq r(x, a) + \gamma(PV_k)(x, a). \quad (4)$$

Here, we have crucially used the condition  $V_k \leq H$  in the inequality, which was made possible by truncating the  $Q$ -functions to the range  $[0, H]$ . In the other case where  $Q_{k+1}(x, a) \leq H$ , we use the validity of  $\text{CB}_k$  to show the following inequality:

$$\begin{aligned} Q_{k+1}(x, a) &\geq r(x, a) + \text{CB}_k(x, a) + \gamma(\widehat{P}_k V_k)(x, a) \\ &\geq r(x, a) + \gamma(PV_k)(x, a), \end{aligned}$$

where the first inequality is valid even when a truncation from below happens. Having covered both cases, we combine our freshly proven inequality  $Q_{k+1} \geq r + \gamma PV_k$  with Equation (3) to obtain

$$\begin{aligned} \Delta_k^* &\leq \langle \mu^*, Q_{k+1} - EV_k \rangle \\ &= \langle \mu^*, Q_{k+1} - EV_{k+1} \rangle + \langle \mu^*, EV_{k+1} - EV_k \rangle. \end{aligned}$$

Summing up for all  $k = 1, \dots, K$ , we get

$$\begin{aligned} \sum_{k=1}^K \Delta_k^* &\leq \langle \mu^*, \overline{Q}_{K+1} - E\overline{V}_{K+1} \rangle \\ &\quad + \langle \mu^*, E(V_{K+1} - V_1) \rangle, \end{aligned}$$

where we defined  $\overline{Q}_k = \sum_{i=1}^k Q_i$  and  $\overline{V}_k = \sum_{i=1}^k V_i$  for any  $k$ . By a classic telescoping argument (presented in Lemma B.1), one can show that, for all  $k$ ,

$$\begin{aligned} \overline{V}_k(x) &= \max_{p \in \Delta(\mathcal{A})} \left\{ \langle p, \overline{Q}_k(x, \cdot) \rangle - \frac{1}{\eta} \mathcal{D}_{\text{KL}}(p \| \pi_0(\cdot | x)) \right\} \\ &\geq \langle \pi^*(\cdot | x), \overline{Q}_k(x, \cdot) \rangle - \frac{1}{\eta} \mathcal{D}_{\text{KL}}(\pi^*(\cdot | x) \| \pi_0(\cdot | x)). \end{aligned}$$

Combining this with the previous inequality, we get

$$\sum_{k=1}^K \Delta_k^* \leq \frac{1}{\eta} \mathcal{H}(\pi^* \| \pi_0) + \langle \mu^*, EV_{K+1} \rangle, \quad (5)$$

where we used the definition of the conditional entropy and that  $V_1 = 0$ .

We now move on to bounding  $\Delta_k$ . Proceeding similarly as before, we will begin by proving the inequality  $Q_{k+1} \leq r + 2\text{CB}_k + \gamma PV_k$  for the two cases corresponding to whether or not  $Q_{k+1}(x, a)$  is truncated from below in each state-action pair. First considering the case where  $Q_{k+1}(x, a) = 0$ , we observe that

$$Q_{k+1}(x, a) = 0 \leq r(x, a) + \gamma(PV_k)(x, a),$$

from which the claim follows due to non-negativity of  $\text{CB}_k$ . As for the other case, we have

$$\begin{aligned} Q_{k+1}(x, a) &\leq r(x, a) + \text{CB}_k(x, a) + \gamma(\widehat{P}_k V_k)(x, a) \\ &\leq r(x, a) + 2\text{CB}_k(x, a) + \gamma(PV_k)(x, a), \end{aligned}$$

where the last step follows from the validity condition on  $\text{CB}_k$ . Then, using the above inequality to lower-bound  $r$ , we bound  $\Delta_k$  as follows:

$$\begin{aligned} \Delta_k &= (1 - \gamma) \langle \nu_0, V_k \rangle - \langle \mu^{\pi_k}, r \rangle \\ &\leq (1 - \gamma) \langle \nu_0, V_k \rangle - \langle \mu^{\pi_k}, Q_{k+1} - 2\text{CB}_k - \gamma PV_k \rangle \\ &= \langle E^\top \mu^{\pi_k} - \gamma P^\top \mu^{\pi_k}, V_k \rangle \\ &\quad - \langle \mu^{\pi_k}, Q_{k+1} - \gamma PV_k \rangle + 2 \langle \mu^{\pi_k}, \text{CB}_k \rangle \\ &= \langle \mu^{\pi_k}, EV_k - Q_{k+1} \rangle + 2 \langle \mu^{\pi_k}, \text{CB}_k \rangle \\ &= \langle \mu^{\pi_k}, EV_k \rangle - \langle \mu^{\pi_{k+1}}, Q_{k+1} \rangle \\ &\quad + \langle \mu^{\pi_{k+1}} - \mu^{\pi_k}, Q_{k+1} \rangle + 2 \langle \mu^{\pi_k}, \text{CB}_k \rangle, \end{aligned}$$

where we have used  $(1 - \gamma) \nu_0 = E^\top \mu^{\pi_k} - \gamma P^\top \mu^{\pi_k}$  in the third line. To proceed, we use Lemma B.1 to note that

$$\langle \mu^{\pi_{k+1}}, Q_{k+1} \rangle = \langle E^\top \mu^{\pi_{k+1}}, V_{k+1} + \frac{1}{\eta} \mathcal{D}_{\text{KL}}(\pi_{k+1} \| \pi_k) \rangle,$$

which allows us to continue as

$$\begin{aligned} \Delta_k &\leq \langle \mu^{\pi_k}, EV_k \rangle - \langle \mu^{\pi_{k+1}}, EV_{k+1} \rangle \\ &\quad + \langle \mu^{\pi_{k+1}} - \mu^{\pi_k}, Q_{k+1} \rangle - \frac{1}{\eta} \mathcal{H}(\pi_{k+1} \| \pi_k) \\ &\quad + 2 \langle \mu^{\pi_k}, \text{CB}_k \rangle. \end{aligned} \quad (6)$$

The last remaining difficulty is to control the second term in the last inequality. This can be done thanks to the regularization, that makes the occupancy measures change ‘‘slowly enough’’. To proceed, we use Pinsker’s inequality and the boundedness of  $Q_{k+1}$  to show

$$\langle \mu^{\pi_{k+1}} - \mu^{\pi_k}, Q_{k+1} \rangle \leq H \sqrt{2\mathcal{D}_{\text{KL}}(\mu^{\pi_{k+1}} \| \mu^{\pi_k})}.$$

Appealing to Lemma A.1, we can bound the last term as

$$\mathcal{D}_{\text{KL}}(\mu^{\pi_{k+1}} \| \mu^{\pi_k}) \leq H \cdot \mathcal{H}(\pi_{k+1} \| \pi_k).$$

Using these results, we can bound the first two terms in the bound of Equation (6), we obtain

$$\begin{aligned} & \langle \mu^{\pi_{k+1}} - \mu^{\pi_k}, Q_{k+1} \rangle - \frac{1}{\eta} \mathcal{H}(\pi_{k+1} \| \pi_k) \\ & \leq \sqrt{2H^3 \mathcal{H}(\pi_{k+1} \| \pi_k)} - \frac{1}{\eta} \mathcal{H}(\pi_{k+1} \| \pi_k) \\ & \leq \sup_z \left\{ \sqrt{2H^3} \cdot z - \frac{1}{\eta} z^2 \right\} = \frac{\eta H^3}{2}, \end{aligned}$$

where the last step follows from the Fenchel–Young inequality applied to the convex function  $f(z) = z^2/2$ . Then, summing up both sides of Equation (6) for all  $k = 1, \dots, K$ , we get

$$\begin{aligned} \sum_{k=1}^K \Delta_k & \leq -\langle \mu^{\pi_{K+1}}, EV_{K+1} \rangle + \frac{\eta H^3}{2} K \\ & \quad + 2 \sum_{k=1}^K \langle \mu^{\pi_k}, CB_k \rangle, \end{aligned} \quad (7)$$

where we used  $V_1 = 0$ . Combining Equations (5) and (7), we finally obtain

$$\begin{aligned} \sum_{k=1}^K (\langle \mu^*, r \rangle - \langle \mu^{\pi_k}, r \rangle) & \leq 2 \sum_{k=1}^K \langle \mu^{\pi_k}, CB_k \rangle + 2H \\ & \quad + \frac{1}{\eta} \mathcal{H}(\pi^* \| \pi_0) + \frac{\eta H^3}{2} K, \end{aligned}$$

where we used  $\langle \mu^* - \mu^{\pi_{K+1}}, EV_{K+1} \rangle \leq 2H$  that follows from the boundedness of  $V_{K+1}$  ensured by the truncation of the iterates. This concludes the proof.  $\square$

## 4.2. Valid Exploration Bonuses for Linear Kernel MDPs

We now turn our attention to the specific setting of linear kernel MDPs and show that they satisfy the validity condition as required by Lemma 4.1.

**Lemma 4.2.** *Suppose that the MDP  $\mathcal{M}$  satisfies Assumption 2.1, and let  $\delta \in (0, 1)$ . Then, setting the coefficient  $\beta = H \sqrt{2 \left( \frac{d}{2} \log \left[ 1 + \frac{TB^2 H^2}{\lambda d} \right] + \log \frac{1}{\delta} \right)} + \sqrt{\lambda} B$  guarantees that, with probability  $1 - \delta$ , the validity condition of Equation (2) is satisfied by  $CB_k$  as defined in Algorithm 1 for all  $k$ .*

The proof is deferred to Appendix A. It relies on standard techniques regarding linear kernel MDPs that we learned from Zhou et al. (2021) and Cai et al. (2020). One important property that is required for this analysis to go through is the boundedness of each  $V_k$  that is guaranteed by the truncation operation performed by **RAVI-UCB**.

## 4.3. The Epoch Schedule

The final technical part is to account for the effects of the randomized epoch schedule. To make these arguments

more convenient to state, we first introduce some notation. We let  $T_k$  denote the first time index of epoch  $k$ ,  $\mathcal{T}_k = \{T_k, T_k + 1, \dots, T_{k+1} - 1\}$  denote the set of time indices belonging to epoch  $k$ , and  $K(T)$  denote the total number of epochs. For the sake of analysis, it will be useful to pad out the trajectory of states and actions with the artificial observations  $(X_{T+1}, A_{T+1}, \dots, X_{T^+}, A_{T^+})$ , where  $T^+$  is the first time that a reset would have occurred had the algorithm been executed beyond time step  $T$ . These additional observations are well-defined random variables, and are introduced to make sure that the last epoch does not require special treatment in the analysis.

Given all this notation, we are almost ready to conclude our analysis. Provided that the exploration bonuses are valid, we need to control the sum  $\sum_{t=1}^T \langle \mu^{\pi_t}, CB_t \rangle$ . We relate it to a more easily tractable sum in the next lemma.

**Lemma 4.3.** *The sequence of policies selected by **RAVI-UCB** satisfies*

$$\mathbb{E} \left[ \sum_{t=1}^T \langle \mu^{\pi_t}, CB_t \rangle \right] \leq \mathbb{E} \left[ \sum_{t=1}^{T^+} CB_t(X_t, A_t) \right]$$

The proof is provided in Appendix A.3. This bound is guaranteed by the epoch schedule used by **RAVI-UCB** that ensures that within each epoch  $k$  of geometric length, the sequence of realized state-action trajectory is distributed according to the occupancy measure of  $\pi_k$ . Then, we can bound the sum of the exploration bonuses with the following lemma.

**Lemma 4.4.** *The sum of exploration bonuses generated by **RAVI-UCB** satisfy*

$$\mathbb{E} \left[ \sum_{t=1}^{T^+} CB_t(X_t, A_t) \right] = \mathcal{O} \left( \beta \sqrt{dHT} \log(T) \right).$$

The proof follows from a series of small (but somewhat tedious) adjustments of a classic result often referred to as the ‘‘elliptical potential lemma’’, the main challenge being dealing with the randomized epoch schedule. The full argument is presented in Appendix A.4.

## 4.4. Putting everything together

The proof of Theorem 3.1 is concluded by putting the above claims together. In particular, we let  $\delta = 1/T$  and choose  $\beta$  as specified in Lemma 4.2. This implies that the resulting exploration bonuses are valid with probability at least  $1 - \delta$ , so on this event we can appeal to Lemma 4.1 to bound the expected regret of **RAVI-UCB**. Setting  $\pi_0$  as the uniform

policy, we get

$$\begin{aligned} \mathbb{E} [\text{regret}_T] &\leq 2\mathbb{E} \left[ \sum_{t=1}^T \langle \mu^{\pi_t}, \text{CB}_t \rangle \right] \\ &\quad + H\mathbb{E} \left[ \frac{1}{\eta} \log |\mathcal{A}| + \frac{\eta H^3}{2} K(T) + 2H \right], \end{aligned}$$

where we used that the expected epoch length is  $H$ . Noticing that  $\mathbb{E}[K] = (1 - \gamma)T$  and setting the learning rate  $\eta = \sqrt{2 \log |\mathcal{A}| / (H^2 T)}$ , the expected optimization error becomes

$$\begin{aligned} \mathbb{E} \left[ \frac{1}{\eta} \log |\mathcal{A}| + \frac{\eta H^3 K}{2} \right] &= \frac{\log |\mathcal{A}|}{\eta} + \frac{\eta H^2 T}{2} \\ &= \sqrt{2 H^2 T \log |\mathcal{A}|}. \end{aligned}$$

The remaining terms in the regret bound corresponding to the sum of exploration bonuses can be bounded by appealing to Lemmas 4.3 and 4.4. This concludes the proof.

We are almost ready to conclude our analysis. The proof of Theorem 3.1 is concluded by putting the above two claims together, and accounting for the effects of the epoch schedule.

Given all this notation, we are almost ready to conclude our analysis. We first appeal to Lemma 4.2 that guarantees the validity of the exploration bonuses with probability  $1 - \delta$ , and then use Lemma 4.1 to show that, under the same event, an appropriately tuned run of **RAVI-UCB** guarantees

where we extended the sum on the right hand side to run up to  $T^+$  via adding a number of nonnegative terms and used the fact that the average epoch length is  $H$ . On average, the number of epochs is then  $T/H$ , thus we can set the learning rate  $\eta = \sqrt{\frac{2 \log |\mathcal{A}|}{H^2 T}}$  to minimize the second term, which becomes  $\sqrt{2 \log |\mathcal{A}| H^4 T} + 2H^2$ .

$$\sum_{t=1}^T \langle \mu^* - \mu^{\pi_t}, r \rangle \leq 2 \sum_{t=1}^{T^+} \langle \mu^{\pi_t}, \text{CB}_t \rangle + \varepsilon T,$$

It remains to bound the first term in the above bound. We relate it to a more easily tractable sum in the next lemma.

This bound is guaranteed by the epoch schedule used by **RAVI-UCB** that guarantees that within each epoch  $k$  of geometric length, the sequence of realized state-action trajectory is distributed according to the occupancy measure of  $\pi_k$ . The final element of the proof is the following result that bounds the sum of the exploration bonuses.

## 5. Discussion

We now discuss the merits and limitations of our results, and point out directions for future research.

There are many remarkable differences between our approach and all previously proposed optimistic exploration methods that we are aware of. Perhaps the most interesting novelty in our method is that it radically relaxes the optimistic properties that previous methods strive for: instead of calculating estimates of the value function or the MDP model that are strictly optimistic, we only guarantee that our value estimates are optimistic in an average sense. Thus, during its runtime, our algorithm may execute several policies that do not individually satisfy any optimistic properties, even approximately. We find this property to be curious and believe that the ideas we develop to tackle such notions of “average optimism” may find other applications. We note though that our planning procedure can be used to produce optimistic policies in a stricter sense by executing several regularized value iteration steps per policy update, until the resulting optimization error vanishes. Doing so results in an improved dependency on  $H$  by a factor  $\sqrt{H}$  but comes at the cost of a major computational overhead.

While our algorithm is closely related to the MD-MPI method of Geist, Scherrer, and Pietquin (2019) and our proofs feature several similar steps, we remark that the purpose of our analysis is quite different from theirs, even when disregarding the optimistic adjustment we make to the Bellman operators. Taking a close look at their proofs for the special case of zero approximation errors, one can deduce bounds on our quantity of interest that are of the order  $(H + \mathcal{H}(\pi^* \|\pi_K)) / K$  after  $K$  iterations. This is faster than what our analysis provides for approximate dynamic programming, which is due to the monotonicity of the exact Bellman operator which allows fast last-iterate convergence. The same rate appears in the analysis of regularized policy iteration methods by Agarwal et al. (2021) (cf. their Theorem 16). In one way or another, all of these analyses use classic tools from the analysis of mirror descent first developed by Martinet (1970), Rockafellar (1976), and Nemirovski & Yudin (1983) (see also Beck & Teboulle, 2003). We note in passing that, due to the fact that these regularization-based methods come with guarantees that hold on arbitrary data sequences, our regret guarantees trivially generalize to the case where the rewards change over time in a potentially adversarial fashion (as in, e.g., Even-Dar et al., 2009; Cai et al., 2020).

Another line of work that our contribution seemingly fits into is the one initiated by Liu & Su (2020) on the topic of regret minimization for discounted MDPs (see also He et al., 2021; Zhou et al., 2021). A closer look reveals that their objective is quite different from ours, in that they aim to upper bound  $\sum_{t=1}^T (V^*(X_t) - V^{\pi_t}(X_t))$  along the trajectory traversed by the learning agent. This notion of regret has been motivated by a formerly popular notion of “sample complexity of exploration” in discounted MDPs—we highlight Kakade (2003); Strehl et al. (2009) out of the abundant “PAC-MDP”



literature on this subject. This performance measure is in fact not comparable to ours in almost any possible sense. In fact, it is easy to see that this notion may fail to capture the sample complexity of learning a good policy in a meaningful way: a policy that immediately enters a “trap” state that yields zero reward until the end of time will only incur a *constant* regret of order  $\frac{1}{1-\gamma}$ , even if there is a policy that yields a steady stream of  $+1$  rewards in each round. Thus, without making stringent assumptions about the MDP that rule out such undesirable scenarios, the value of minimizing this notion of discounted regret may be questionable.

On a related note, our own method suffers from the limitation of requiring access to a reset action taking the agent back to the initial-state distribution  $\nu_0$  at any desired time. It is easy to see that in general, this device is necessary to achieve our objectives. Indeed, in MDPs where all states around the initial distribution are transient, it is impossible to learn a good policy from a single stream of experience without resets since the agent only gets to visit the relevant part of the state space once. We thus believe that all these issues are inherent to learning in discounted MDPs.

Another limitation of our result is that our guarantees only hold on expectation as opposed to high probability. It is easy to see that several of our results can be strengthened to hold in this stronger sense, albeit at the cost of a more involved technical analysis. In particular, the only parts of our analysis that need to be changed are Lemmas 4.3 and 4.4, to deal with the fact that the epoch schedule is randomized. The first of these can be handled via a simple martingale argument and the second by bounding the number and length of the epochs with high probability. Both of these changes are conceptually simple, but practically tedious so we omit them for the sake of keeping the paper easier to access. On the other hand, the result of Corollary 3.2 relies on a randomized online-to-batch conversion, and the result is stated on expectation with respect to the randomization step. Once again, this result can be strengthened to hold with high probability by running a “best-policy-selection” subroutine on the entire sequence of policies produced by the algorithm. This post-processing step is standard in the related literature and we omit details here to preserve clarity.

Based on our current results, generalizing our techniques to the infinite-horizon average-reward setting seems to be challenging but not impossible. The key step in our proof that requires discounting is setting the truncation level at  $H = \frac{1}{1-\gamma}$ , which serves the purpose of guaranteeing that our approximate Bellman operator is optimistic. In particular, the truncation level needs to be set large enough so that the inequality of Equation 4 goes through. We see no natural way to extend this condition to the undiscounted setup. We remain hopeful that this challenge can be overcome with more effort (but may potentially need some significant new ideas).

Finally, let us remark on the linear kernel MDP assumption that we have used in our analysis. While arguably very well-studied in the past years, this model for linear function approximation has several limitations that make it rather difficult to adapt to practical scenarios. The biggest of these limitations is that learning algorithms in this model need access to an oracle to evaluate sums of the form  $\sum_{x'} \psi(x, a, x') V(x')$ , which can only be performed efficiently in special cases. Options include assuming that the  $\psi(x, a, \cdot)$  is sparse or the integral can be approximated effectively via Monte Carlo sampling. A major inconvenience that this causes in the implementation of our method is that Q-functions (and thus policies) cannot be represented effectively with a single low-dimensional object, so these values have to be recalculated on the fly while executing the policy, thus necessitating excessive Monte Carlo integration in runtime. We thus wish to extend our analysis to more tractable MDP models like the model of Jin et al. (2020). While it is straightforward to implement our algorithm for linear MDPs, unfortunately the covering number of the value function class used by our algorithm appears to be too large to allow proving strong performance bounds. On a more positive note, we wish to point out that linear kernel MDPs are still a rich family of models that in general is incomparable to linear MDPs, and can subsume many interesting models such as linear mixture MDPs (Modi et al., 2020) and linear-factor MDPs (Yang & Wang, 2020)—we refer to (Ayoub et al., 2020) for further discussion. We are optimistic that the limitations of our current analysis can be eventually removed and our method can be adapted to a much broader class of infinite-horizon MDPs.

## References

- Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.
- 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.
- 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, 2020.
- Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.
- Bertsekas, D. and Shreve, S. E. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

- Brafman, R. I. and Tenenbholz, M. R-max: A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct): 213–231, 2002.
- Bubeck, S. and Cesa-Bianchi, N. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Now Publishers Inc, 2012.
- Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.
- Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.
- de la Peña, Victor, H., Lai, T. L., and Shao, Q.-M. *Self-normalized processes: Limit theory and Statistical Applications*, volume 204. Springer, 2009. Self-normalized tail bound appears in Thm. 14.7.
- Even-Dar, E., Kakade, S. M., and Mansour, Y. Online Markov decision processes. *Math. Oper. Res.*, 34(3): 726–736, 2009.
- Fruit, R., Pirota, 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.
- Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.
- He, J., Zhou, D., and Gu, Q. Nearly minimax optimal reinforcement learning for discounted MDPs. *Advances in Neural Information Processing Systems*, 34:22288–22300, 2021.
- Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.
- 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, 2020.
- Kakade, S. *On the sample complexity of reinforcement learning*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
- Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22, 1985.
- Lai, T. L. and Wei, C. Z. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.
- Lai, T. L., Robbins, H., and Wei, C. Z. Strong consistency of least squares estimates in multiple regression ii. *Journal of multivariate analysis*, 9(3):343–361, 1979.
- Lattimore, T. and Szepesvári, Cs. *Bandit algorithms*. Cambridge University Press, 2020.
- Liu, S. and Su, H. Regret bounds for discounted MDPs. *arXiv preprint arXiv:2002.05138*, 2020.
- Martinet, B. Régularisation d’inéquations variationnelles par approximations successives. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 4(R3):154–158, 1970.
- 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, 2020.
- Nemirovski, A. and Yudin, D. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.
- Neu, G. and Pike-Burke, C. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.
- Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.
- Ouhama, R., Basu, D., and Maillard, O.-A. Bilinear exponential family of mdps: Frequentist regret bound with tractable exploration and planning. *arXiv preprint arXiv:2210.02087*, 2022.
- Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- Qian, J., Fruit, R., Pirota, M., and Lazaric, A. Exploration bonus for regret minimization in undiscounted discrete and continuous markov decision processes. *arXiv preprint arXiv:1812.04363*, 2018.
- Rockafellar, R. T. Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

- Strehl, A. L., Li, L., and Littman, M. L. Reinforcement learning in finite MDPs: PAC analysis. *The Journal of Machine Learning Research*, 10:2413–2444, 2009.
- Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction. 2nd edition.* 2018.
- Vial, D., Parulekar, A., Shakkottai, S., and Srikant, R. Regret bounds for stochastic shortest path problems with linear function approximation. In *International Conference on Machine Learning*, pp. 22203–22233, 2022.
- 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, 2020.
- Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted MDPs with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021.

## A. Omitted proofs

### A.1. Technical tools for the proof of Lemma 4.1

**Lemma A.1.** *Let  $\pi$  and  $\pi'$  be two policies, with their corresponding state-action occupancy measures being  $\mu^\pi$  and  $\mu^{\pi'}$ , and their state occupancy measures being  $\nu^\pi$  and  $\nu^{\pi'}$ . Then,*

$$\mathcal{D}_{\text{KL}}\left(\mu^\pi \parallel \mu^{\pi'}\right) \leq \frac{1}{1-\gamma} \mathcal{H}(\pi \parallel \pi').$$

*Proof.* Using the chain rule of the relative entropy, we write

$$\mathcal{D}_{\text{KL}}\left(\mu^\pi \parallel \mu^{\pi'}\right) = \mathcal{D}_{\text{KL}}\left(\nu^\pi \parallel \nu^{\pi'}\right) + \mathcal{H}(\pi \parallel \pi').$$

By the Bellman flow constraints in Equation (1) and the joint convexity of the relative entropy, we bound the second term as

$$\begin{aligned} \mathcal{D}_{\text{KL}}\left(\nu^\pi \parallel \nu^{\pi'}\right) &= \mathcal{D}_{\text{KL}}\left(\gamma P^\top \mu^\pi + (1-\gamma)\nu_0 \parallel \gamma P^\top \mu^{\pi'} + (1-\gamma)\nu_0\right) \\ &\leq (1-\gamma) \mathcal{D}_{\text{KL}}(\nu_0 \parallel \nu_0) + \gamma \mathcal{D}_{\text{KL}}\left(P^\top \mu^\pi \parallel P^\top \mu^{\pi'}\right) \\ &= \gamma \mathcal{D}_{\text{KL}}\left(P^\top \mu^\pi \parallel P^\top \mu^{\pi'}\right) \leq \gamma \mathcal{D}_{\text{KL}}\left(\mu^\pi \parallel \mu^{\pi'}\right), \end{aligned}$$

where we also used the data-processing inequality in the last step. The proof is concluded by reordering the terms.  $\square$

### A.2. Proof of Lemma 4.2

Let us fix  $k \in [K]$ ,  $t \in \{T_k, T_k + 1, \dots, T_{k+1} - 1\}$ ,  $\delta \in (0, 1)$ . We start by recalling the definition of the nominal transition model  $\widehat{P}_k$  acting on functions  $V$  as  $(\widehat{P}_k V)(x, a) = \langle \varphi_V(x, a), \widehat{\theta}_k \rangle$ , where we denoted the state-action feature map  $\varphi_V(x, a) = \sum_{x' \in \mathcal{X}} \psi(x, a, x') V(x')$ , and the parameter  $\widehat{\theta}_k$  can be written out as

$$\widehat{\theta}_k = \Lambda_{T_k}^{-1} b_{T_k} = \left( \sum_{i=0}^{k-1} \sum_{j=T_i}^{T_{i+1}-1} \varphi_i(x_j, a_j) \varphi_i(x_j, a_j)^\top + \lambda I \right)^{-1} \sum_{i=0}^{k-1} \sum_{j=T_i}^{T_{i+1}-1} \varphi_i(x_j, a_j) V_i(x_{j+1}).$$

To proceed, we notice that the true transition operator acting on  $V$  can be written in a similar form as

$$\begin{aligned} (PV)(x, a) &= \sum_{x' \in \mathcal{X}} P(x'|x, a) V(x') && \text{(by definition of } P) \\ &= \sum_{x' \in \mathcal{X}} \langle \theta, \psi(x, a, x') \rangle V(x') && \text{(by Assumption 2.1)} \\ &= \left\langle \theta, \sum_{x' \in \mathcal{X}} \psi(x, a, x') V(x') \right\rangle \\ &= \langle \theta, \varphi_V(x, a) \rangle, \end{aligned}$$

where we used the definition of  $\varphi_V$  in the last line. Proceeding further with the same expression, we write

$$\begin{aligned} (PV)(x, a) &= \langle \varphi_V(x, a), \Lambda_{T_k}^{-1} \Lambda_{T_k} \theta \rangle \\ &= \left\langle \varphi_V(x, a), \Lambda_{T_k}^{-1} \sum_{i=0}^{k-1} \sum_{j=T_i}^{T_{i+1}-1} \varphi_i(x_j, a_j) \varphi_i(x_j, a_j)^\top \theta + \lambda \Lambda_{T_k}^{-1} \theta \right\rangle \\ &= \left\langle \varphi_V(x, a), \Lambda_{T_k}^{-1} \sum_{i=0}^{k-1} \sum_{j=T_i}^{T_{i+1}-1} \varphi_i(x_j, a_j) (PV_i)(x_j, a_j) + \lambda \Lambda_{T_k}^{-1} \theta \right\rangle, \end{aligned}$$

where we used the definition of  $\Lambda_{T_k}$  and the linear kernel MDP assumption in the last line. Comparing the expressions for  $PV$  and  $\widehat{P}_k V$ , we obtain

$$\left| \widehat{P}_k V(x, a) - PV(x, a) \right| = \left| \left\langle \varphi_V(x, a), \Lambda_{T_k}^{-1} \sum_{i=0}^{k-1} \sum_{j=T_i}^{T_{i+1}-1} \varphi_i(x_j, a_j) [V_i(x_{j+1}) - (PV_i)(x_j, a_j)] - \lambda \Lambda_{T_k}^{-1} \theta \right\rangle \right|.$$

Using the Cauchy–Schwartz inequality and taking  $V = V_k$ , we get

$$\left| \widehat{P}_k V_k(x, a) - PV_k(x, a) \right| \leq \|\varphi_k(x, a)\|_{\Lambda_{T_k}^{-1}} (|\xi_k| + |b_k|),$$

where  $\xi_k = \left\| \sum_{i=0}^{k-1} \sum_{j=T_i}^{T_{i+1}-1} \varphi_i(x_j, a_j) [V_i(x_{j+1}) - (PV_i)(x_j, a_j)] \right\|_{\Lambda_{T_k}^{-1}}$ , and  $b_k = \lambda \|\theta\|_{\Lambda_{T_k}^{-1}}$ . The second term can be easily bounded as  $|b_k| \leq \sqrt{\lambda} \|\theta\|_2 \leq \sqrt{\lambda} B$ , using that  $\lambda_{\min}(\Lambda_{T_k}) \geq \lambda$  and the boundedness of the features.

For the first term, observe that  $V_i(x_{j+1}) - (PV_i)(x_j, a_j)$  forms a martingale difference sequence, with increments bounded in  $[-H, H]$  by the truncation made in the algorithm. Additionally, the feature vectors are bounded as  $\|\varphi_i(x_j, a_j)\|_2 \leq BH$  and the true parameter as  $\|\theta\|_2 \leq B$  by Assumption 2.1. Therefore, we can apply the self-normalized concentration result in Theorem B.2 (stated in Appendix B.2), which guarantees that with probability at least  $1 - \delta$ , the following bound holds for all  $k \in [K]$ :

$$\xi_k \leq H \sqrt{2 \log \left[ \frac{\det(\Lambda_{T_k})^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right]}.$$

The determinants appearing in the bound can be further upper bounded by using that  $\det(\lambda I) = \lambda^d$  and

$$\begin{aligned} \det(\Lambda_{T_k}) &\leq \left( \frac{\text{tr}(\Lambda_{T_k})}{d} \right)^d && \text{(by the trace-determinant inequality)} \\ &= \frac{1}{d^d} \left( \lambda d + \sum_{i=0}^{k-1} \sum_{j=T_i}^{T_{i+1}-1} \|\varphi_i(x_j, a_j)\|_2^2 \right)^d && \text{(by the definition of } \Lambda_{T_k} \text{)} \\ &\leq \left( \lambda + \frac{T_k B^2 H^2}{d} \right)^d && \text{(by the boundedness of the features)} \\ &\leq \left( \lambda + \frac{T B^2 H^2}{d} \right)^d, \end{aligned}$$

where in the last step we used  $T_k \leq T$ . We plug this back in the upper-bound on  $\xi_k$  to obtain the bound

$$\xi_k \leq H \sqrt{2 \left( \frac{d}{2} \log \left[ 1 + \frac{T B^2 H^2}{\lambda d} \right] + \log \frac{1}{\delta} \right)}.$$

Putting everything together, we have verified that, for all  $k \in [K]$ ,

$$\begin{aligned} \left| \left\langle P(\cdot|x, a) - \widehat{P}_k(\cdot|x, a), V_k \right\rangle \right| &\leq \|\varphi_k(x, a)\|_{\Lambda_{T_k}^{-1}} \left( H \sqrt{2 \left( \frac{d}{2} \log \left[ 1 + \frac{T B^2 H^2}{\lambda d} \right] + \log \frac{1}{\delta} \right)} + \sqrt{\lambda} B \right) \\ &= \beta \|\varphi_k(x, a)\|_{\Lambda_{T_k}^{-1}}. \end{aligned}$$

holds with probability at least  $1 - \delta$ , where we have defined  $\beta$  as

$$\beta = H \sqrt{2 \left( \frac{d}{2} \log \left[ 1 + \frac{T B^2 H^2}{\lambda d} \right] + \log \frac{1}{\delta} \right)} + \sqrt{\lambda} B. \quad (8)$$

This concludes the proof.  $\square$

### A.3. Proof of Lemma 4.3

For the sake of this proof, we slightly update our notation for  $\mathcal{T}_k$  by setting  $\mathcal{T}_{K(T)} = \{T_{K(T)}, T_{K(T)} + 1, \dots, T^+\}$ . We will use  $\mathcal{F}_{k-1}$  to denote the filtration generated by the observations up to the end of epoch  $k - 1$ , and  $L_k$  to denote the length of epoch  $k$ . We start by rewriting the sum of exploration bonuses up to step  $T^+$  as

$$\sum_{t=1}^{T^+} \text{CB}_t(X_t, A_t) = \sum_{k=1}^{K(T)} \sum_{t \in \mathcal{T}_k} \text{CB}_t(X_t, A_t). \quad (9)$$

By virtue of the definition of  $T^+$ , all epochs are of geometric length with mean  $\frac{1}{1-\gamma}$ . Now, let us consider a fixed epoch  $k$  and define the auxiliary infinite sequence of state-action pairs  $X_{k,0}, A_{k,0}, X_{k,1}, A_{k,1}, \dots$  that is generated independently from the realized sample trajectory  $(X_t, A_t)_{t \in \mathcal{T}_k}$  given  $\mathcal{F}_{k-1}$  as follows. The initial state  $X_{k,0}$  is drawn from  $\nu_0$ , and then subsequently for each  $i = 0, 1, \dots$ , the actions are drawn as  $A_{k,i} \sim \pi_k(\cdot | X_{k,i})$  and follow-up states are drawn as  $X_{k,i+1} \sim P(\cdot | X_{k,i}, A_{k,i})$ . Recalling the notational convention that  $\text{CB}_t = \text{CB}_k$  for all  $t \in \mathcal{T}_k$ , we observe that for any  $k$ , we have

$$\begin{aligned} \mathbb{E} \left[ \sum_{t \in \mathcal{T}_k} \text{CB}_t(X_t, A_t) \middle| \mathcal{F}_{k-1} \right] &= \mathbb{E} \left[ \sum_{i=0}^{L_k-1} \text{CB}_k(X_{k,i}, A_{k,i}) \middle| \mathcal{F}_{k-1} \right] \\ &= \mathbb{E} \left[ \sum_{i=0}^{\infty} \mathbb{1}_{\{i < L_k\}} \text{CB}_k(X_{k,i}, A_{k,i}) \middle| \mathcal{F}_{k-1} \right] \\ &= \mathbb{E} \left[ \sum_{i=0}^{\infty} \gamma^i \text{CB}_k(X_{k,i}, A_{k,i}) \middle| \mathcal{F}_{k-1} \right] \\ &= \sum_{i=0}^{\infty} \gamma^i \langle u_{k,i}, \text{CB}_k \rangle = \frac{\langle \mu^{\pi_k}, \text{CB}_k \rangle}{1-\gamma} \\ &= \mathbb{E} [L_k \langle \mu^{\pi_k}, \text{CB}_k \rangle | \mathcal{F}_{k-1}] = \mathbb{E} \left[ \sum_{t \in \mathcal{T}_k} \langle \mu^{\pi_k}, \text{CB}_k \rangle \middle| \mathcal{F}_{k-1} \right], \end{aligned}$$

where in the third line we have observed that  $L_k$  follows a geometric law with parameter  $1 - \gamma$ , and is independent of  $(X_{k,i}, A_{k,i})_i$ . In the fourth line we introduced the notation  $u_{k,i}$  to denote the joint distribution of  $X_{k,i}, A_{k,i}$  given  $\mathcal{F}_{k-1}$  and noticed that the discounted sum of these distributions exactly matches the definition of the occupancy measure  $\mu^{\pi_k}$  up to the normalization constant  $(1 - \gamma)$ , and finally concluded by observing that  $\mathbb{E}[L_k | \mathcal{F}_{k-1}] = \frac{1}{1-\gamma}$ .

The proof is completed by summing up over all epochs, taking marginal expectations, and noticing that

$$\mathbb{E} \left[ \sum_{t=1}^T \langle \mu^{\pi_t}, \text{CB}_t \rangle \right] \leq \mathbb{E} \left[ \sum_{t=1}^{T^+} \langle \mu^{\pi_t}, \text{CB}_t \rangle \right] = \mathbb{E} \left[ \sum_{k=1}^{K(T)} \sum_{t \in \mathcal{T}_k} \langle \mu^{\pi_k}, \text{CB}_k \rangle \right].$$

□

### A.4. Proof of Lemma 4.4

The proof is based on a classic ‘‘pigeonhole’’ argument often called the ‘‘elliptical potential lemma’’ (e.g., Lemma 19.4 in Lattimore & Szepesvari, 2020, or Section 11.7 in Cesa-Bianchi & Lugosi, 2006, but see also Lai et al., 1979; Lai & Wei, 1982). The main challenge of adapting this result to our setting is accounting for the randomized epoch schedule. Another subtle difficulty comes from the fact that Lemma 4.1 only bounds the total regret as opposed to the instantaneous regrets in each round, which necessitates arguments that are slightly more involved than what is commonly seen in closely related work.

As for the actual proof, we start by introducing some useful notation that we will use throughout the proof. For  $t \in [T]$ , we use  $k_t$  to denote the index of the epoch that  $t$  belongs to. For simplicity, for all  $k$  and  $t$ , we will write  $\varphi_{k,t} = \varphi_k(X_t, A_t)$ ,  $\Lambda_k = \Lambda_{T_k}$ . We also define  $\mathcal{N}(T) = \left\{ t \in [T] : \|\varphi_{k_t,t}\|_{\Lambda_{k_t}^{-1}} \geq 1 \right\}$  as the set of ‘‘bad’’ time indices where state-action

pairs with large feature norms are observed, and  $\mathcal{E}(T) = \left\{k \in [K(T)] : \exists t \in \mathcal{T}_k, \|\varphi_{k,t}\|_{\Lambda_k^{-1}} \geq 1\right\}$  be the set of epochs containing at least one bad time index. Using these definitions, we rewrite the sum of exploration bonuses as follows:

$$\begin{aligned}
 \mathbb{E} \left[ \sum_{t=1}^{T^+} \text{CB}_t(X_t, A_t) \right] &= \beta \mathbb{E} \left[ \sum_{k \in \mathcal{E}(T)} \sum_{t \in \mathcal{T}_k} \|\varphi_{k,t}\|_{\Lambda_k^{-1}} + \sum_{k \notin \mathcal{E}(T)} \sum_{t \in \mathcal{T}_k} \|\varphi_{k,t}\|_{\Lambda_k^{-1}} \right] \\
 &\leq \beta \mathbb{E} \left[ \frac{BH}{\sqrt{\lambda}} \sum_{k \in \mathcal{E}(T)} |\mathcal{T}_k| + \sum_{k \notin \mathcal{E}(T)} \sum_{t \in \mathcal{T}_k} \|\varphi_{k,t}\|_{\Lambda_k^{-1}} \right] \quad (\text{using } \|\varphi\|_2 \leq BH \text{ and } \Lambda_k \succeq \lambda I) \\
 &= \beta \mathbb{E} [|\mathcal{E}(T)|] \frac{BH^2}{\sqrt{\lambda}} + \beta \mathbb{E} \left[ \sum_{k \notin \mathcal{E}(T)} \sum_{t \in \mathcal{T}_k} \|\varphi_{k,t}\|_{\Lambda_k^{-1}} \right] \quad (\text{using Wald's identity}) \\
 &= \beta \mathbb{E} [|\mathcal{E}(T)|] \frac{BH^2}{\sqrt{\lambda}} + \beta \mathbb{E} \left[ \sum_{k \notin \mathcal{E}(T)} \sum_{t \in \mathcal{T}_k} \left(1 \wedge \|\varphi_{k,t}\|_{\Lambda_k^{-1}}\right) \right],
 \end{aligned}$$

where in the last step we used the definition of  $\mathcal{E}(T)$ . We treat the first term separately in Lemma A.3, stated after this proof. This gives the following bound:

$$\mathbb{E} \left[ \sum_{t=1}^{T^+} \text{CB}_t(X_t, A_t) \right] \leq \frac{\beta d B H^2}{\sqrt{\lambda} \log(2)} \log \left( 1 + \frac{B^2 H^2 T}{\lambda d} \right) + \beta \mathbb{E} \left[ \sum_{k \notin \mathcal{E}(T)} \sum_{t \in \mathcal{T}_k} \left(1 \wedge \|\varphi_{k,t}\|_{\Lambda_k^{-1}}\right) \right].$$

Thus, we can focus on the second term in the right hand side. This term can be upper-bounded using the Cauchy–Schwarz inequality as

$$\sum_{k \notin \mathcal{E}(T)} \sum_{t \in \mathcal{T}_k} \left(1 \wedge \|\varphi_{k,t}\|_{\Lambda_k^{-1}}\right) \leq \sum_{k=1}^{K(T)} \sum_{t \in \mathcal{T}_k} \left(1 \wedge \|\varphi_{k,t}\|_{\Lambda_k^{-1}}\right) \leq \sqrt{T} \sqrt{\sum_{k=1}^{K(T)} \sum_{t \in \mathcal{T}_k} \left(1 \wedge \|\varphi_{k,t}\|_{\Lambda_k^{-1}}^2\right)}.$$

To proceed, we use the inequality  $(x \wedge |\mathcal{T}_k|) \leq \frac{|\mathcal{T}_k|}{\log(|\mathcal{T}_k|+1)} \log(1+x)$  that is valid for all  $x \geq 0$ . Setting  $C_k = \frac{|\mathcal{T}_k|}{\log(|\mathcal{T}_k|+1)}$ , this gives

$$\begin{aligned}
 \sum_{k=1}^{K(T)} \sum_{t \in \mathcal{T}_k} \left(1 \wedge \|\varphi_{k,t}\|_{\Lambda_k^{-1}}^2\right) &= \sum_{k=1}^{K(T)} \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \left(|\mathcal{T}_k| \wedge |\mathcal{T}_k| \|\varphi_{k,t}\|_{\Lambda_k^{-1}}^2\right) \leq \sum_{k=1}^{K(T)} \frac{C_k}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \log \left(1 + |\mathcal{T}_k| \|\varphi_{k,t}\|_{\Lambda_k^{-1}}^2\right) \\
 &\leq \max_k C_k \cdot \sum_{k=1}^{K(T)} \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \log \left(1 + |\mathcal{T}_k| \|\varphi_{k,t}\|_{\Lambda_k^{-1}}^2\right).
 \end{aligned}$$

The sum is handled separately in Lemma A.2 stated and proved right after this proof. Putting the result together with our previous calculations, we get

$$\sum_{k \notin \mathcal{E}(T)} \sum_{t \in \mathcal{T}_k} \left(1 \wedge \|\varphi_{k,t}\|_{\Lambda_k^{-1}}\right) \leq \sqrt{T} \sqrt{\max_k C_k} \sqrt{d \log \left(1 + \frac{B^2 H^2 T}{\lambda d}\right)}.$$

The only random quantity left in the upper-bound is the maximum over  $C_k$ . By concavity of the square-root function and Jensen's inequality, we get

$$\mathbb{E} \left[ \sum_{k \notin \mathcal{E}(T)} \sum_{t \in \mathcal{T}_k} \left(1 \wedge \|\varphi_{k,t}\|_{\Lambda_k^{-1}}\right) \right] \leq \sqrt{T} \sqrt{\mathbb{E} \left[ \max_k C_k \right]} \sqrt{d \log \left(1 + \frac{B^2 H^2 T}{\lambda d}\right)},$$

which we further upper bound by using Lemma A.4 as

$$\mathbb{E} \left[ \sum_{k \notin \mathcal{E}(T)} \sum_{t \in \mathcal{T}_k} \left(1 \wedge \|\varphi_{k,t}\|_{\Lambda_k^{-1}}\right) \right] \leq \sqrt{T} \sqrt{H(4 + 2 \log T)} \sqrt{d \log \left(1 + \frac{B^2 H^2 T}{\lambda d}\right)}.$$

We put together the two terms, and plug in the definition of  $\beta$  to get

$$\mathbb{E} \left[ \sum_{t=1}^{T^+} \text{CB}_t(X_t, A_t) \right] \leq C_1(T) + \sqrt{T} C_2(T),$$

where the two factors are defined as

$$C_1(T) = \left( H \sqrt{2 \left( \frac{d}{2} \log \left[ 1 + \frac{B^2 H^2 T}{\lambda d} \right] + \log T \right)} + \sqrt{\lambda} B \right) \frac{dBH^2}{\sqrt{\lambda} \log(2)} \log \left( 1 + \frac{B^2 H^2 T}{\lambda d} \right)$$

$$C_2(T) = \left( H \sqrt{2 \left( \frac{d}{2} \log \left[ 1 + \frac{B^2 H^2 T}{\lambda d} \right] + \log T \right)} + \sqrt{\lambda} B \right) \sqrt{H(4 + 2 \log T)} \sqrt{d \log \left( 1 + \frac{B^2 H^2 T}{\lambda d} \right)}.$$

The proof is then concluded by observing that  $C_1(T) + C_2(T) = \mathcal{O}(BH^{3/2}d \log(T)^{3/2})$ .  $\square$

**Lemma A.2.** *Following the same notations that in Section A.4*

$$\sum_{k=1}^K \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \log \left( 1 + |\mathcal{T}_k| \|\varphi_{k,t}\|_{\Lambda_k^{-1}}^2 \right) \leq d \log \left( 1 + \frac{B^2 H^2 T}{\lambda d} \right).$$

*Proof.* We will follow the steps of the proof of Lemma 19.4 in [Lattimore & Szepesvári \(2020\)](#). First, notice that  $\Lambda_k$  can be written as

$$\Lambda_{k+1} = \Lambda_k + \sum_{t \in \mathcal{T}_k} \varphi_{k,t} \varphi_{k,t}^\top = \Lambda_k^{1/2} \left( I + \sum_{t \in \mathcal{T}_k} \Lambda_k^{-1/2} \varphi_{k,t} \varphi_{k,t}^\top \Lambda_k^{-1/2} \right) \Lambda_k^{1/2}.$$

Taking the determinant of the above matrix, we get

$$\det(\Lambda_{k+1}) = \det(\Lambda_k) \det \left( I + \sum_{t \in \mathcal{T}_k} \Lambda_k^{-1/2} \varphi_{k,t} \varphi_{k,t}^\top \Lambda_k^{-1/2} \right).$$

Now, taking logarithms on both sides and using the concavity of  $\log \det$ , we obtain

$$\begin{aligned} \log \det(\Lambda_{k+1}) &= \log \det(\Lambda_k) + \log \det \left( I + \sum_{t \in \mathcal{T}_k} \Lambda_k^{-1/2} \varphi_{k,t} \varphi_{k,t}^\top \Lambda_k^{-1/2} \right) \\ &= \log \det(\Lambda_k) + \log \det \left( \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \left( I + |\mathcal{T}_k| \Lambda_k^{-1/2} \varphi_{k,t} \varphi_{k,t}^\top \Lambda_k^{-1/2} \right) \right) \\ &\geq \log \det(\Lambda_k) + \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \log \det \left( I + |\mathcal{T}_k| \Lambda_k^{-1/2} \varphi_{k,t} \varphi_{k,t}^\top \Lambda_k^{-1/2} \right) \\ &= \log \det(\Lambda_k) + \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \log \left( 1 + |\mathcal{T}_k| \|\varphi_{k,t}\|_{\Lambda_k^{-1}}^2 \right), \end{aligned}$$

where the inequality is Jensen's, and the final step follows from using the equality  $\det(I + vv^\top) = (1 + \|v\|_2^2)$  that holds for any  $v \in \mathbb{R}^d$ . Summing up for  $k$  gives

$$\log \det(\Lambda_{K(T)}) \geq \log \det(\Lambda_1) + \sum_{k=1}^K \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \log \left( 1 + |\mathcal{T}_k| \|\varphi_{k,t}\|_{\Lambda_k^{-1}}^2 \right),$$

and furthermore by the trace-determinant inequality we have

$$\log \left( \frac{\det \Lambda_{K(T)}}{\det \Lambda_0} \right) = \log \left( \frac{\det(\Lambda_{K(T)})}{\lambda^d} \right) \leq d \log \left( \frac{\text{tr}(\Lambda_{K(T)})}{\lambda d} \right).$$



Finally, the trace can be bounded as

$$\mathrm{tr}(\Lambda_{K(T)}) = \lambda d + \sum_{k=1}^{K(T)} \sum_{t \in \mathcal{T}_k} \|\varphi_{k,t}\|_2^2 \leq \lambda d + B^2 H^2 T.$$

Plugging this back into the previous inequality proves the claim.  $\square$

**Lemma A.3.** *The number of epochs that contain a feature vector with a norm larger than one is bounded as*

$$|\mathcal{E}(T)| \leq \frac{d}{\log(2)} \log \left( 1 + \frac{B^2 H^2 T}{\lambda d} \right).$$

A simpler version of this statement is given as Exercise 19.3 in [Lattimore & Szepesvári \(2020\)](#), and our proof below drew inspiration from the proof of Lemma 19 in [\(Ouhamma et al., 2022\)](#). We only have to deal with the challenge of randomized epoch schedules, which we do by similar arguments as in the proof of Lemma 4.4 above.

*Proof.* Let  $k \in [K(T)]$ . We define  $G_0 = \lambda I$  and  $G_{k+1} = G_k + \sum_{t \in \mathcal{T}_k} \varphi_{k,t} \varphi_{k,t}^\top \mathbb{1}_{\{t \in \mathcal{N}(T)\}}$ . We have the following decomposition:

$$\begin{aligned} G_{k+1} &= G_k^{1/2} \left( I + \sum_{t \in \mathcal{T}_k} \left( G_k^{-1/2} \varphi_{k,t} \right) \left( G_k^{-1/2} \varphi_{k,t} \right)^\top \mathbb{1}_{\{t \in \mathcal{N}(T)\}} \right) G_k^{1/2} \\ &= G_k^{1/2} \left( I + \sum_{t \in \mathcal{T}_k \cap \mathcal{N}(T)} \left( G_k^{-1/2} \varphi_{k,t} \right) \left( G_k^{-1/2} \varphi_{k,t} \right)^\top \right) G_k^{1/2}. \end{aligned}$$

Therefore, taking the log-determinant on both sides, we obtain

$$\log \det(G_{k+1}) = \log \det(G_k) + \log \det \left( I + \sum_{t \in \mathcal{T}_k \cap \mathcal{N}(T)} \left( G_k^{-1/2} \varphi_{k,t} \right) \left( G_k^{-1/2} \varphi_{k,t} \right)^\top \right).$$

If  $\mathcal{T}_k \cap \mathcal{N}(T) = \emptyset$ , or equivalently if  $k \notin \mathcal{E}(T)$  (i.e., there is no “bad” state-action pair in the epoch  $k$ ), the second term in the right-hand side is zero. Hence, summing over  $k \in [\mathcal{E}(T)]$ , we get

$$\log \det(G_{K(T)}) = \log \det(G_0) + \sum_{k \in \mathcal{E}(T)} \log \det \left( I + \sum_{t \in \mathcal{T}_k \cap \mathcal{N}(T)} \left( G_k^{-1/2} \varphi_{k,t} \right) \left( G_k^{-1/2} \varphi_{k,t} \right)^\top \right). \quad (10)$$

Using the concavity of  $\log \det$ , Jensen’s inequality gives us

$$\begin{aligned} \log \det(G_{K(T)}) &\geq \log \det(G_0) \\ &\quad + \sum_{k \in \mathcal{E}(T)} \frac{1}{|\mathcal{T}_k \cap \mathcal{N}(T)|} \sum_{t \in \mathcal{T}_k \cap \mathcal{N}(T)} \log \det \left( I + |\mathcal{T}_k \cap \mathcal{N}(T)| \left( G_k^{-1/2} \varphi_{k,t} \right) \left( G_k^{-1/2} \varphi_{k,t} \right)^\top \right) \\ &= \log \det(G_0) + \sum_{k \in \mathcal{E}(T)} \frac{1}{|\mathcal{T}_k \cap \mathcal{N}(T)|} \sum_{t \in \mathcal{T}_k \cap \mathcal{N}(T)} \log \left( 1 + |\mathcal{T}_k \cap \mathcal{N}(T)| \|\varphi_{k,t}\|_{G_k^{-1}}^2 \right), \end{aligned}$$

where the equality follows from the fact that  $\det(I + vv^\top) = (1 + \|v\|_2^2)$  that holds for any  $v \in \mathbb{R}^d$ . Then, we notice that  $G_k^{-1} \succeq \Lambda_k^{-1}$ , and thus we can further bound this expression as

$$\log \det(G_{K(T)}) \geq \log \det(G_0) + \sum_{k \in \mathcal{E}(T)} \frac{1}{|\mathcal{T}_k \cap \mathcal{N}(T)|} \sum_{t \in \mathcal{T}_k \cap \mathcal{N}(T)} \log \left( 1 + |\mathcal{T}_k \cap \mathcal{N}(T)| \|\varphi_{k,t}\|_{\Lambda_k^{-1}}^2 \right).$$

For  $k \in \mathcal{E}(T)$ ,  $t \in \mathcal{T}_k \cap \mathcal{N}(T)$ , we have  $|\mathcal{T}_k \cap \mathcal{N}(T)| \geq 1$ , and  $\|\varphi_{k,t}\|_{\Lambda_k^{-1}} \geq 1$  by definition of  $\mathcal{N}(T)$ . This implies that

$$\begin{aligned} \log \det(G_{K(T)}) &\geq \log \det(G_0) + \sum_{k \in \mathcal{E}(T)} \frac{1}{|\mathcal{T}_k \cap \mathcal{N}(T)|} \sum_{t \in \mathcal{T}_k \cap \mathcal{N}(T)} \log(2) \\ &\geq \log \det(G_0) + \log(2) |\mathcal{E}(T)|. \end{aligned}$$

Thus, we have

$$\begin{aligned} |\mathcal{E}(T)| &\leq \frac{1}{\log(2)} \log \left( \frac{\det(G_{K(T)})}{\det(G_0)} \right) \\ &= \frac{1}{\log(2)} \log \left( \frac{\det(G_{K(T)})}{\lambda^d} \right) && \text{(by the definition of } G_1) \\ &\leq \frac{d}{\log(2)} \log \left( \frac{\text{tr}(G_{K(T)})}{\lambda d} \right). && \text{(by the trace-determinant inequality)} \end{aligned}$$

Finally, the trace can be bounded as

$$\text{tr}(G_{K(T)}) = \lambda d + \sum_{k \in \mathcal{E}(T)} \sum_{t \in \mathcal{T}_k \cap \mathcal{N}(T)} \|\varphi_{k,t}\|_2^2 \mathbb{1}_{\mathcal{N}(T)}(t) \leq \lambda d + B^2 H^2 T.$$

The proof is concluded by putting this bound together with the previous inequality.  $\square$

**Lemma A.4.** *The random variables  $\{C_k\}_k$ , defined for all  $k$  by  $C_k = \frac{|\mathcal{T}_k|}{\log(1+|\mathcal{T}_k|)}$  where  $|\mathcal{T}_k|$  is a geometric random variable with parameter  $1 - \gamma$ , satisfy the following*

$$\mathbb{E} \left[ \max_k C_k \right] \leq \frac{4 + 2 \log T}{1 - \gamma}.$$

*Proof.* First, notice that  $\log(1 + |\mathcal{T}_k|) \geq \log 2$  so that  $C_k = \frac{|\mathcal{T}_k|}{\log(1+|\mathcal{T}_k|)} \leq \frac{|\mathcal{T}_k|}{\log 2}$ . Next, using the fact that the number of epochs is at most  $T$ , and observing that each  $|\mathcal{T}_k|$  is geometrically distributed with parameter  $1 - \gamma$ , we can bound  $\max_k |\mathcal{T}_k|$  by a maximum over  $T$  independent geometric random variables  $Z_1, \dots, Z_T$  with parameter  $1 - \gamma$ :

$$\begin{aligned} \mathbb{E} \left[ \max_k |\mathcal{T}_k| \right] &\leq \mathbb{E} \left[ \max_{j \in [T]} Z_j \right] \\ &= \sum_{i=0}^{\infty} \mathbb{P} \left[ \max_{j \in [T]} Z_j > i \right] && \text{(since each } Z_i \text{ is nonnegative)} \\ &\leq k + T \sum_{i=k}^{\infty} \mathbb{P}[Z_1 > i] && \text{(upper bounding the first } k > 0 \text{ terms by 1)} \\ &= k + T \sum_{i=k}^{\infty} \gamma^i && \text{(using that } Z_1 \text{ is geometric with parameter } 1 - \gamma) \\ &= k + \frac{T\gamma^k}{1 - \gamma}, \end{aligned}$$

where we have used the formula for the geometric sum in the last step. Now, setting  $k = \left\lceil \frac{\log T}{1 - \gamma} \right\rceil$ , we get

$$\begin{aligned} \mathbb{E} \left[ \max_k |\mathcal{T}_k| \right] &= k + T \frac{\gamma^k}{1 - \gamma} \leq \frac{1 + \log T}{1 - \gamma} + \frac{T \exp\left(\frac{\log \gamma}{1 - \gamma} \cdot \log T\right)}{1 - \gamma} \\ &\leq \frac{1 + \log T}{1 - \gamma} + \frac{T \exp(-\log T)}{1 - \gamma} = \frac{2 + \log T}{1 - \gamma}, \end{aligned}$$

where in the second line we have used the inequality  $\frac{\log \gamma}{1 - \gamma} \leq -1$  that holds for all  $\gamma \in (0, 1)$ . The proof is concluded by using that  $\log 2 > \frac{1}{2}$  and combining the above bound with the bound relating  $C_k$  to  $|\mathcal{T}_k|$ .  $\square$

## B. Standard Results

### B.1. Softmax policies and value functions

In this section, we recall a range of standard facts relating the softmax policies our algorithm uses and the associated value functions. These can be found in numerous papers, textbooks, and lecture notes—for concreteness, see Section 28.1 in [Lattimore & Szepesvári, 2020](#) as an example.

**Lemma B.1.** *Let  $\{V_k\}_{k \in [K]}$ ,  $\{\pi_k\}_{k \in [K]}$ , and  $\{Q_k\}_{k \in [K]}$  be the sequences of functions defined in Algorithm 1. Then, the following equalities are satisfied for all  $k \in [K]$  and  $x \in \mathcal{X}$ :*

$$V_k(x) = \max_{p \in \Delta(\mathcal{A})} \left\{ \langle p, Q_k(x, \cdot) \rangle - \frac{1}{\eta} \mathcal{D}_{\text{KL}}(p \| \pi_{k-1}(\cdot | x)) \right\}$$

$$\pi_k(\cdot | x) = \arg \max_{p \in \Delta(\mathcal{A})} \left\{ \langle p, Q_k(x, \cdot) \rangle - \frac{1}{\eta} \mathcal{D}_{\text{KL}}(p \| \pi_{k-1}(\cdot | x)) \right\}.$$

Furthermore, for all  $k \in [K]$  and  $x \in \mathcal{X}$ , we have

$$\sum_{i=1}^k V_i(x) = \max_{p \in \Delta(\mathcal{A})} \left\{ \left\langle p, \sum_{i=1}^k Q_i(x, \cdot) \right\rangle - \frac{1}{\eta} \mathcal{D}_{\text{KL}}(p \| \pi_0(\cdot | x)) \right\}.$$

*Proof.* First, we show that the maximum indeed takes the form claimed in the main paper and that the maximizer is given by a softmax policy. For simplicity, we drop the indices for now and consider the optimization problem

$$\sup_{p \in \Delta(\mathcal{A})} \left\{ \langle p, Q \rangle - \frac{1}{\eta} \mathcal{D}_{\text{KL}}(p \| p') \right\},$$

where  $Q \in \mathbb{R}^{\mathcal{A}}$ , and  $p' \in \Delta(\mathcal{A})$ . As the probability simplex is compact and  $(p \mapsto \langle p, Q \rangle - \frac{1}{\eta} \mathcal{D}_{\text{KL}}(p \| p'))$  is continuous, the supremum is attained at some  $p^* \in \Delta(\mathcal{A})$ . The Lagrangian function of this optimization problem is given for all  $p \in \mathbb{R}_+^{\mathcal{A}}$  and  $\alpha \in \mathbb{R}$  as

$$\mathcal{L}(p, \alpha) = \langle p, Q \rangle - \frac{1}{\eta} \mathcal{D}_{\text{KL}}(p \| p') + \alpha (\langle p, \mathbf{1} \rangle - 1).$$

Its partial derivative with respect to the primal variable  $p(a)$  is

$$\frac{\partial \mathcal{L}(p, \alpha)}{\partial p(a)} = Q(a) - \frac{1}{\eta} \left( \log \left( \frac{p(a)}{p'(a)} \right) + 1 \right) + \alpha.$$

Setting it to zero gives us the expression

$$p^*(a) = p'(a) \exp(\eta(Q(a) + \alpha) - 1).$$

Then, we use the constraint on  $p^*$  to find the value of  $\alpha$ . In particular,  $\langle p^*, \mathbf{1} \rangle = 1$  implies

$$\sum_{a \in \mathcal{A}} p'(a) \exp(\eta Q(a)) = \exp(1 - \eta\alpha),$$

from which we deduce that

$$\alpha = \frac{1}{\eta} \left( 1 - \log \left( \sum_{a \in \mathcal{A}} p'(a) \exp(\eta Q(a)) \right) \right).$$

Denoting  $V^* = \frac{1}{\eta} \log(\sum_{a \in \mathcal{A}} p'(a) \exp[\eta Q(a)])$ , we plug back the expression of  $\alpha$  into  $p^*$ :

$$p^*(a) = p'(a) \exp(\eta(Q(a) - V^*)).$$

From this, we can directly express the relative entropy between  $p^*$  and  $p'$  as

$$\mathcal{D}_{\text{KL}}(p^* \| p') = \sum_a p^*(a) \log \frac{p^*(a)}{p'(a)} = \sum_a p^*(a) (Q(a) - V^*) = \langle p^*, Q - V^* \mathbf{1} \rangle,$$

so that we can write

$$\langle p^*, Q \rangle - \frac{1}{\eta} \mathcal{D}_{\text{KL}}(p^* \| p') = \langle p^*, Q \rangle - \langle p^*, Q - V^* \mathbf{1} \rangle = V^*.$$

The first statement of the lemma then follows from applying this result to  $Q = Q_k(x, \cdot)$  and  $p' = \pi_{k-1}(\cdot|x)$ , for  $k \in [K]$ ,  $x \in \mathcal{X}$ . That is, for any state-action pair  $(x, a) \in \mathcal{X} \times \mathcal{A}$  and  $k \geq 1$ , denoting the maximum  $V_k$  and the maximizer  $\pi_k$ , we have that the following expressions are equivalent to the ones given in the statement of the lemma:

$$V_k(x) = \frac{1}{\eta} \log \left( \sum_{a \in \mathcal{A}} \pi_{k-1}(a|x) e^{\eta Q_k(x, a)} \right),$$

$$\pi_k(a|x) = \pi_{k-1}(a|x) \exp(\eta [Q_k(x, a) - V_k(x)]).$$

For the second statement, we start by denoting  $\bar{V}_k = \sum_{i=1}^k V_i$  and  $\bar{Q}_k = \sum_{i=1}^k Q_i$ , for  $k \geq 1$ , and show by induction that, for  $x \in \mathcal{X}$ , the following holds

$$\pi_k(\cdot|x) = \pi_0(\cdot|x) \exp(\eta [\bar{Q}_k(x, \cdot) - \bar{V}_k(x) \mathbf{1}]).$$

Let  $x \in \mathcal{X}$ . The case  $k = 1$  follows immediately from the previous statement with  $Q = Q_1(x, \cdot)$  and  $p' = \pi_0(\cdot|x)$ . Assume the previous equation holds at  $k$ . Using the first statement with  $Q = Q_{k+1}(x, \cdot)$  and  $p' = \pi_k(\cdot|x)$  we have, for  $a \in \mathcal{A}$ ,  $\pi_{k+1}(a|x) = \pi_k(a|x) e^{\eta [Q_{k+1}(x, a) - V_{k+1}(x)]}$ . Applying the inductive hypothesis, it gives

$$\begin{aligned} \pi_{k+1}(a|x) &= \pi_0(a|x) \exp(\eta [\bar{Q}_k(x, a) - \bar{V}_k(x)]) \exp(\eta [Q_{k+1}(x, a) - V_{k+1}(x)]) \\ &= \pi_0(a|x) \exp(\eta [\bar{Q}_{k+1}(x, a) - \bar{V}_{k+1}(x)]), \end{aligned}$$

which finishes the induction. Then, we move on to the actual statement. We have

$$\begin{aligned} V_k(x) &= \frac{1}{\eta} \log \left( \sum_{a \in \mathcal{A}} \pi_{k-1}(a|x) e^{\eta Q_k(x, a)} \right) && \text{(by the first statement)} \\ &= \frac{1}{\eta} \log \left( \sum_{a \in \mathcal{A}} \pi_0(a|x) e^{\eta (Q_k(x, a) + \bar{Q}_{k-1}(x, a) - \bar{V}_{k-1}(x))} \right) && \text{(by induction)} \\ &= \frac{1}{\eta} \log \left( \sum_{a \in \mathcal{A}} \pi_0(a|x) e^{\eta (\bar{Q}_k(x, a))} \right) - \bar{V}_{k-1}(x). \end{aligned}$$

Therefore, by definition of  $\bar{V}_{k-1}$ ,

$$\begin{aligned} \sum_{i=1}^k V_i(x) &= \frac{1}{\eta} \log \left( \sum_{a \in \mathcal{A}} \pi_0(a|x) e^{\eta (\bar{Q}_k(x, a))} \right) \\ &= \max_{p \in \Delta(\mathcal{A})} \left\{ \left\langle p, \sum_{i=1}^k Q_i(x, \cdot) \right\rangle - \frac{1}{\eta} \mathcal{D}_{\text{KL}}(p \| \pi_0(\cdot|x)) \right\}, \end{aligned}$$

which concludes the proof.  $\square$

## B.2. A self-normalized tail inequality

**Theorem B.2** (Theorem 14.7 in [de la Peña et al. \(2009\)](#), Theorem 2 in [Abbasi-Yadkori et al. \(2011\)](#)). *Let  $\{\eta_t\}_{t=1}^\infty$  be a real-valued stochastic process with corresponding filtration  $\{\mathcal{F}_t\}_{t=0}^\infty$ . Let  $\eta_t | \mathcal{F}_{t-1}$  be zero-mean and  $\sigma$ -subGaussian; i.e.  $\mathbb{E}[\eta_t | \mathcal{F}_{t-1}] = 0$ , and*

$$\forall \lambda \in \mathbb{R}, \mathbb{E} [e^{\lambda \eta_t} | \mathcal{F}_{t-1}] \leq e^{\frac{\lambda^2 \sigma^2}{2}}.$$

*Let  $\{\varphi_t\}_{t=0}^\infty$  be an  $\mathbb{R}^d$ -valued stochastic process where  $\varphi_t$  is  $\mathcal{F}_{t-1}$ -measurable. Assume  $\Lambda_0$  is a  $d \times d$  positive definite matrix, and let  $\Lambda_t = \Lambda_0 + \sum_{s=1}^t \varphi_s \varphi_s^\top$ . Then, for any  $\delta > 0$ , with probability at least  $1 - \delta$ , we have for all  $t \geq 0$ ,*

$$\left\| \sum_{s=1}^t \varphi_s \eta_s \right\|_{\Lambda_t^{-1}}^2 \leq 2\sigma^2 \log \left[ \frac{\det(\Lambda_t)^{1/2} \det(\Lambda_0)^{-1/2}}{\delta} \right].$$