Optimistic Planning by Regularized Dynamic Programming

Antoine Moulin [1] **Gergely Neu** [1]

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 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 recover known guarantees in tabular MDPs and to provide a computationally efficient algorithm for learning near-optimal policies in discounted linear mixture MDPs from a single stream of experience, and show 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 RL algorithms for a variety of problem settings. Following the work of Brafman & Tennenholtz (2002); Strehl et al. (2009) on optimistic exploration methods for RL in Markov decision processes (MDPs), a breakthrough was achieved by Jaksch, Ortner, and Auer (2010), whose UCRL2 algorithm was shown to achieve near-optimal regret guarantees in a broad class of tabular MDPs. In subsequent years, their work 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 (DP) 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 (Fruit et al., 2018; Qian et al., 2018). However, 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 an approximate Bellman operator that satisfies the necessary contraction properties.

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

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

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 MDPs. Instead of relying on a contraction argument (or an approximate version thereof), we propose to solve the 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 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 mixture MDPs and showing that it achieves a near-optimal regret bound of order $\sqrt{\left(B^2dH + d^2H^3 + \log|\mathcal{A}|\,H^4\right)T}$, 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 ε-optimal policy after about $\left(B^2dH + d^2H^3 + \log|\mathcal{A}|\,H^4\right)/\varepsilon^2$ iterations. Each policy update takes $\mathrm{poly}(d, H, T)$ iterations of regularized dynamic programming, each consisting of $\mathrm{poly}(d, H, T)$ elementary operations. This is to be contrasted with previous contributions on a similar[1] 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 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 presenting the notation at the end of this section and the background in Section 2, we introduce our algorithmic framework in Section 3. We provide generic performance guarantees and explain the key steps of the analysis in Section 4. The guarantees are instantiated in the context of tabular and linear mixture MDPs in Section 5. We conclude in Section 6 with a discussion of our contribution along with its limitations.

Notation. For a natural number $N > 0$, we denote $[N] = \{1, 2, \ldots, N\}$. For a real number M, we define the truncation operator Π_M that acts on functions f defined on a domain A via $\Pi_M f : x \mapsto \max\left(\min\left[f\left(x\right), M\right], 0\right)$. 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 $P \ll Q$, we define the relative entropy as $\mathcal{D}_{\mathrm{KL}}\left(P\|Q\right) = \mathbb{E}_{a\sim P}\left[\ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(a)\right)\right]$. For a distribution $P \in \Delta(A)$ and a bounded function $f \in \mathbb{R}^A$, we write $\langle P, f\rangle = \mathbb{E}_{a\sim P}\left[f(a)\right]$ 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[2], $\mathcal{A}$ is the finite action space, $r : \mathcal{X}\times\mathcal{A} \to [0,1]$ is the deterministic reward function assumed to be known[3], $P : \mathcal{X}\times\mathcal{A} \to \Delta\left(\mathcal{X}\right)$ is the transition probability distribution, $\gamma \in (0,1)$ is the discount factor, and $\nu_0 \in \Delta\left(\mathcal{X}\right)$ is the initial state distribution. The

[1] We provide a detailed discussion about the differences between our settings in Section 6.1.

[2] 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.

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

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 $t = 1, 2, \ldots$ 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 $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ from a state to a probability measure over actions. The value function and action-value function of a policy π 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

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

where $\mathbb{E}_\pi$ denotes the expectation with respect to the probability measure $\mathbb{P}_\pi$, generated by the interaction between the environment and the policy π. With some abuse of notation, we define the conditional expectation operator $P : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ as $(Pf)(x,a) = \sum_{x' \in \mathcal{X}} P(x'|x,a) f(x')$, for $f \in \mathbb{R}^{\mathcal{X}}$. Its adjoint $P^\intercal$ acts on distributions $\mu \in \Delta(\mathcal{X}\times\mathcal{A})$ as $(P^\intercal \mu)(x') = \sum_{x,a \in \mathcal{X}\times\mathcal{A}} P(x'|x,a)\mu(x,a)$. It returns the state distribution realized after starting from the state-action distribution μ and then taking a step forward in the MDP dynamics. With these, we can simply state the *Bellman equations* tying together the value functions as

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

We also introduce the operator $E : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ acting on functions $f \in \mathbb{R}^{\mathcal{X}}$ via the assignment $(Ef)(x,a) = f(x)$, and its adjoint via its action $E^\intercal \mu(x) = \sum_a \mu(x,a)$ on distributions $\mu \in \Delta(\mathcal{X}\times\mathcal{A})$. The operator E can be thought of as a "padding" operator over the action space that allows us to use vector notation, while $E^\intercal$ applied to a state-action distribution returns the corresponding marginal distribution of states. The adjoint $P^\intercal$ (resp. $E^\intercal$) is the operator such that, for any f, g, $\langle Pf, g\rangle = \langle f, P^\intercal g\rangle$ (resp. E, $E^\intercal$).

In a discounted MDP, a policy π 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 ν^π is a probability measure over $\mathcal{X}$. 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 μ^π, 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*:

$$E^\intercal \mu^\pi = \gamma P^\intercal \mu^\pi + (1-\gamma)\nu_0. \tag{1}$$

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 μ^* to denote an occupancy measure with maximal return and $\nu^* = E^\intercal \mu^*$ to denote the associated state-occupancy measure. Finally, given two policies π, π', we denote $\mathcal{D}_{\mathrm{KL}}(\pi\|\pi') = \left(\mathcal{D}_{\mathrm{KL}}(\pi(\cdot|x)\|\pi'(\cdot|x))\right)_{x\in\mathcal{X}}$, and we define $\mathcal{H}(\pi\|\pi') = \langle \nu^\pi, \mathcal{D}_{\mathrm{KL}}(\pi\|\pi')\rangle$, the conditional relative entropy[4].

In this paper, we will consider the setting of online learning in discounted MDPs, where the agent aims to produce an ε-optimal policy π_{out} satisfying $\langle \mu^* - \mu^{\pi_{\text{out}}}, r\rangle \le \varepsilon$ 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 ν_0, and that the learner follows a stationary policy π_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 ε-optimal. As an auxiliary performance measure, we will also consider the *expected regret* (or simply, *regret*)[5] of the learner defined as

$$\mathfrak{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}\left(\langle \mu^* - \mu^{\pi_t}, r\rangle\right)\right].$$

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 $1, \ldots, T$ and returning $\pi_{\text{out}} = \pi_I$ guarantees that

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

which can be made arbitrarily small if $\mathfrak{R}_T$ grows sublinearly 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

[4] Technically, this is the conditional relative entropy between the *occupancy measures* μ^π 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.

[5] In the related literature, it is more common to define regret as a random variable and bound it with high probability. Our algorithm is only suitable for bounding the expected regret, and thus we only define this quantity here; we defer further discussion to Section 6.

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 measures the suboptimality of the policies along the trajectory traversed by the learner. We defer a further discussion of the two settings to Section 6.1.

3. Algorithm

Our approach implements the principle of optimism in the face of uncertainty in discounted MDPs. 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-function and policy updates as follows. Starting with an initial estimate $V_0 = 0$ and an initial policy π_0, it calculates a sequence of updates for $k = 1, \dots, K$ as

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

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

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

Here, $\widehat{P}$ is a nominal transition model and CB_k is an exploration bonus defined to be large enough to ensure that $\gamma \widehat{P} V_k + \mathrm{CB}_k \geq \gamma P V_k$ and so that Q_{k+1} is an upper bound on the regularized Bellman update $r + \gamma P V_k$. 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 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 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 $\widehat{P}_k$ and perform one step of online mirror descent to produce the new policy π_k and the associated softmax value function V_k. We then update the exploration bonuses CB_k such that they satisfy, for all x, a

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

We will refer to exploration bonuses satisfying the above condition as *valid*. As we will see explicitly in Section 5, the model estimate and bonuses are computed using the data gathered so far, $\mathcal{D}_{T_k}$, where T_k denotes the first time index of epoch k. Finally, we apply an optimistic Bellman update to produce a state-action value estimate Q_{k+1}.

We highlight that the assignments in Algorithm 1 are only made symbolically for all x, a, and a practical implementation will not necessarily 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.

Finally, to make some of the arguments in Section 4 more convenient to state, we introduce some notation. We let $\mathcal{T}_k = \{T_k, T_k + 1, \dots, T_{k+1} - 1\}$ 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 observations are well-defined random variables, and are introduced to make sure that the last epoch does not require special treatment.

Algorithm 1 **RAVI-UCB**.

Inputs: Horizon T, learning rate $\eta > 0$, initial value V_0, initial policy π_0.
Initialize: $t = 1$, $Q_1 = EV_0$, $\mathcal{D}_1 = \emptyset$.
for $k = 1, \dots$ **do**
  $T_k = t$.
  $\widehat{P}_k = \texttt{TRANSITION-ESTIMATE}\left(\mathcal{D}_{T_k}\right)$.
  $V_k\left(x\right) = \frac{1}{\eta} \log\left(\sum_a \pi_{k-1}\left(a|x\right) e^{\eta Q_k(x,a)}\right)$.
  $\pi_k\left(a|x\right) = \pi_{k-1}\left(a|x\right) e^{\eta\left(Q_k(x,a) - V_k(x)\right)}$.
  $\mathrm{CB}_k = \texttt{BONUS}\left(\mathcal{D}_{T_k}\right)$.
  $Q_{k+1} = \Pi_H\Big[r + \mathrm{CB}_k + \gamma \widehat{P}_k V_k\Big]$.
  repeat
    Play $a_t \sim \pi_k\left(\cdot|x_t\right)$.
    Observe x_{t+1}.
    Update $\mathcal{D}_{t+1} = \texttt{ADD}\left(\mathcal{D}_t, \{(x_t, a_t, x_{t+1})\}\right)$.
    $t = t + 1$.
    With probability $1-\gamma$, reset to initial distribution: $x_t \sim \nu_0$ and **break**.
  until $t = T$
end for

4. Main Result & Analysis

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

Theorem 4.1. *Let $\{\pi_k\}_k$ and $\{\mathrm{CB}_k\}_k$ be the policies and exploration bonuses produced by **RAVI-UCB** over T timesteps, where the input is $\eta = \sqrt{2\log|\mathcal{A}|/(H^2T)}$, $V_0 = 0$ and any policy π_0. Suppose that the sequence of bonuses $\{\mathrm{CB}_k\}_k$ is valid in the sense of Equation (2). Then the policies $\{\pi_k\}_k$ satisfy the following bound:*

$$\mathfrak{R}_T \leq 2\mathbb{E}\left[\sum_{t=1}^{T^+} \mathrm{CB}_t\left(X_t, A_t\right)\right] + \sqrt{2H^4 \log|\mathcal{A}|\, T} + 2H^2.$$

We present the proof of Theorem 4.1 below. 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 setting considered. In particular, we first analyze **RAVI-UCB** using a generic exploration bonus that we will suppose to be "valid", and then show in Section 5 how to derive such valid exploration bonuses in the concrete settings of tabular MDPs and linear mixture MDPs.

4.1. Optimistic Planning

We first study the properties of our optimistic planning procedure, without making explicit references to the setting. For this general analysis, we will fix an epoch index k, assume that $\widehat{P}_k$ is some estimator of the transition kernel P and that the exploration bonus CB_k is valid in the sense of Equation (2). We provide the following inequality that will be useful for bounding the suboptimality gaps.

Lemma 4.2. *Let Q_{k+1} be the state-action value estimate produced by **RAVI-UCB** in epoch k, with any input, and assume the bonuses CB_k are valid in the sense of Equation (2). Then,*

$$r + \gamma P V_k \leq Q_{k+1} \leq r + 2\mathrm{CB}_k + \gamma P V_k,$$

where V_k is the value estimate defined in Algorithm 1.

Proof. We start by proving the lower-bound. 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 first case, we have $Q_{k+1}(x,a) = H$, which implies

$$Q_{k+1}(x,a) = H = 1 + \gamma H \geq r(x,a) + \gamma\big(PV_k\big)(x,a). \tag{3}$$

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 CB_k to show the following inequality:

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

where the first inequality is valid even when a truncation from below happens.

For the upper-bound, we proceed similarly and consider 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

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

from which the claim follows due to non-negativity of CB_k. As for the other case, we have

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

where the last step follows from the validity condition on CB_k. □

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

Lemma 4.3. *Let K be a fixed number of epochs, and let π_k and CB_k be the policy and exploration bonus produced by **RAVI-UCB** in epoch k, where the input is $V_0 = 0$, any policy π_0, and any $\eta > 0$. Suppose that $\{\mathrm{CB}_k\}_k$ is a sequence of valid exploration bonuses in the sense of Equation (2). Then, the sequence $\{\pi_k\}_k$ satisfies the following bound:*

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

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 π_k can be bounded in terms of $\langle\mu^{\pi_k}, \mathrm{CB}_k\rangle$. With this in mind, we begin by rewriting the performance gap of the output policy as follows:

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

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, Δ_k^*. We start by observing that $(1-\gamma)\,\nu_0 = E^\intercal \mu^* - \gamma P^\intercal \mu^*$, which allows us to write

$$\begin{aligned} \Delta_k^* &= \langle \mu^*, r \rangle - (1-\gamma) \langle \nu_0, V_k \rangle \\ &= \langle \mu^*, r + \gamma P V_k \rangle - \langle \mu^*, E V_k \rangle . \end{aligned} \tag{4}$$

In order to treat the first term in Equation (4), we use the lower-bound from Lemma 4.2 to obtain

$$\begin{aligned} \Delta_k^* &\le \langle \mu^*, Q_{k+1} - E V_k \rangle \\ &= \langle \mu^*, Q_{k+1} - E V_{k+1} \rangle + \langle \mu^*, E V_{k+1} - E V_k \rangle . \end{aligned}$$

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

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

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 C.1), one can show that, for all k,

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

Combining this with the previous inequality, we get

$$\sum_{k=1}^K \Delta_k^* \le \frac{1}{\eta} \mathcal{H}\left(\pi^* \| \pi_0\right) + \langle \mu^*, E V_{K+1} \rangle , \tag{5}$$

by definition of the conditional entropy and $V_1 = 0$.

We now move on to bounding Δ_k. Then, using the upper-bound of Lemma 4.2 to lower-bound r, we bound Δ_k as follows:

$$\begin{aligned} \Delta_k &= (1-\gamma) \langle \nu_0, V_k \rangle - \langle \mu^{\pi_k}, r \rangle \\ &\le (1-\gamma) \langle \nu_0, V_k \rangle - \big\langle \mu^{\pi_k}, Q_{k+1} - 2\mathrm{CB}_k - \gamma P V_k \big\rangle \\ &= \langle E^\intercal \mu^{\pi_k} - \gamma P^\intercal \mu^{\pi_k}, V_k \rangle \\ &\quad - \langle \mu^{\pi_k}, Q_{k+1} - \gamma P V_k \rangle + 2 \langle \mu^{\pi_k}, \mathrm{CB}_k \rangle , \end{aligned}$$

where we have used $(1-\gamma)\,\nu_0 = E^\intercal \mu^{\pi_k} - \gamma P^\intercal \mu^{\pi_k}$ in the third line. We can then rewrite the current upper-bound as

$$\begin{aligned} \Delta_k &\le \langle \mu^{\pi_k}, E V_k - Q_{k+1} \rangle + 2 \langle \mu^{\pi_k}, \mathrm{CB}_k \rangle \\ &= \langle \mu^{\pi_k}, E V_k \rangle - \langle \mu^{\pi_{k+1}}, Q_{k+1} \rangle \\ &\quad + \langle \mu^{\pi_{k+1}} - \mu^{\pi_k}, Q_{k+1} \rangle + 2 \langle \mu^{\pi_k}, \mathrm{CB}_k \rangle . \end{aligned}$$

To proceed, we use Lemma C.1 to note that

$$\langle \mu^{\pi_{k+1}}, Q_{k+1} \rangle = \big\langle E^\intercal \mu^{\pi_{k+1}}, V_{k+1} + \frac{1}{\eta} \mathcal{D}_{\mathrm{KL}}\left(\pi_{k+1} \| \pi_k\right) \big\rangle,$$

which allows us to continue as

$$\begin{aligned} \Delta_k &\le \langle \mu^{\pi_k}, E V_k \rangle - \langle \mu^{\pi_{k+1}}, E V_{k+1} \rangle \\ &\quad + \langle \mu^{\pi_{k+1}} - \mu^{\pi_k}, Q_{k+1} \rangle - \frac{1}{\eta} \mathcal{H}\left(\pi_{k+1} \| \pi_k\right) \\ &\quad + 2 \langle \mu^{\pi_k}, \mathrm{CB}_k \rangle . \end{aligned} \tag{6}$$

The last remaining difficulty is to control the second difference 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

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

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

$$\mathcal{D}_{\mathrm{KL}}\left(\mu^{\pi_{k+1}} \| \mu^{\pi_k}\right) \le H \cdot \mathcal{H}\left(\pi_{k+1} \| \pi_k\right).$$

Using these results, we obtain

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

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 $k = 1, \dots, K$,

$$\begin{aligned} \sum_{k=1}^K \Delta_k \le - \langle \mu^{\pi_{K+1}}, E V_{K+1} \rangle + \frac{\eta H^3}{2} K \\ + 2 \sum_{k=1}^K \langle \mu^{\pi_k}, \mathrm{CB}_k \rangle , \end{aligned} \tag{7}$$

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

$$\begin{aligned} \sum_{k=1}^K \left(\langle \mu^*, r \rangle - \langle \mu^{\pi_k}, r \rangle \right) \le 2 \sum_{k=1}^K \langle \mu^{\pi_k}, \mathrm{CB}_k \rangle + 2H \\ + \frac{1}{\eta} \mathcal{H}\left(\pi^* \| \pi_0\right) + \frac{\eta H^3}{2} K, \end{aligned}$$

where we used $\langle \mu^* - \mu^{\pi_{K+1}}, E V_{K+1} \rangle \le 2H$. □

4.2. The Epoch Schedule

The final part is to account for the effects of the randomized epoch schedule. Provided that the exploration bonuses are valid, we need to control the sum $\sum_{t=1}^T \langle \mu^{\pi_t}, \mathrm{CB}_t \rangle$. We relate it to a more easily tractable sum in the next lemma.

Lemma 4.4. *The sequence of policies selected by* ***RAVI-UCB*** *satisfies*

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

The proof is 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 π_k.

4.3. Putting Everything Together

The proof of Theorem 4.1 concludes by combining the above claims. In anticipation of Section 5, for our main assumption to be satisfied we let $\delta = 1/T$ and define the exploration bonuses as in Lemma 5.1 or Lemma 5.4. This implies the resulting exploration bonuses are valid with probability at least $1-\delta$, so on this event we can use Lemma 4.3 to bound the expected regret of **RAVI-UCB**. Setting π_0 as the uniform policy, we get

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

where we used that the expected epoch length is H and $\mathcal{H}\left(\pi^*\|\pi_0\right)\leq\log|\mathcal{A}|$. Noticing that $\mathbb{E}\left[K\right]=\left(1-\gamma\right)T$ and setting the learning rate $\eta=\sqrt{2\log|\mathcal{A}|/\left(H^2T\right)}$, the expected optimization error becomes

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

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

5. Applications

We now consider two classes of MDPs and show how to implement our algorithm and derive a regret bound.

5.1. Tabular MDPs

For didactic purposes, we first instantiate **RAVI-UCB** in the setting of tabular MDPs with small state and action spaces. As we will see, a simple application of our framework allows us to recover known guarantees in this setting. The algorithm can be found in Appendix B.1. Let $N_1(x,a)=1$ and $N_1'(x,a,x')=0$ denote the initial counts[6] for the tuples (x,a) and (x,a,x'). At epoch k, for $t\in\mathcal{T}_k$, we update $\mathcal{D}_{t+1}=\left(N_{t+1},N_{t+1}'\right)$ as $N_{t+1}(x,a)=N_t(x,a)+\mathbb{I}_{\{X_t=x,A_t=a\}}$ and $N_{t+1}'(x,a,x')=N_t'(x,a,x')+\mathbb{I}_{\{X_t=x,A_t=a,X_{t+1}=x'\}}$. We use $\widehat{P}_k\left(x'|x,a\right)=N_{T_k}(x,a,x')/N_{T_k}(x,a)$ as a model estimate, and given $\beta>0$, the exploration bonuses are defined as

$$\mathrm{CB}_k(x,a)=\frac{\beta}{\sqrt{N_{T_k}(x,a)}}. \tag{8}$$

The following lemma shows that an appropriate choice of the scaling parameter β ensures the validity of the exploration bonuses.

Lemma 5.1. *Let $\delta\in(0,1)$. Then, setting the coefficient $\beta=8H\sqrt{|\mathcal{X}|\log\left(|\mathcal{X}|\,|\mathcal{A}|\,T/\delta\right)}$ guarantees that, with probability $1-\delta$, the validity condition (2) is satisfied by CB_k as defined in Equation (8) for all k.*

Then, we can bound the bonuses as follows.

Lemma 5.2. *The sum of exploration bonuses generated by* ***RAVI-UCB*** *satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T^+}\mathrm{CB}_t\left(X_t,A_t\right)\right]=\mathcal{O}\left(\beta\sqrt{|\mathcal{X}|\,|\mathcal{A}|\,T}\right).$$

We refer the reader to previous works for the proofs of the above two lemmas (see, *e.g.*, Jaksch et al., 2010; Fruit et al., 2018). Combining the above two results gives a regret bound of order $|\mathcal{X}|\,H\sqrt{|\mathcal{A}|\,T}$, as expected.

5.2. Linear Mixture MDPs

We now focus on a class of MDPs commonly referred to as *linear mixture MDPs* (Modi et al., 2020; Ayoub et al., 2020) formally defined as follows.

Assumption 5.3 (Linear mixture MDP)**.** There exist a known feature map $\psi:\mathcal{X}\times\mathcal{A}\times\mathcal{X}\to\mathbb{R}^d$, 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 $(x,a)\in\mathcal{X}\times\mathcal{A}$, $V\in[0,H]^{\mathcal{X}}$,

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

Here, we suppose $\mathcal{M}$ satisfies Assumption 5.3. While remotely related to the notion of linear MDPs (Jin et al., 2020; Yang & Wang, 2019), linear mixture MDPs are a distinct class of models that cannot be captured in that framework, and have been widely studied in the past few years as linear MDPs—we refer to Zhou et al. (2021) for further discussion. As often assumed in the related literature, we assume the map $\varphi_k(x,a)=\sum_{x'}\psi(x,a,x')V_k\left(x'\right)$ can be computed (or approximated) efficiently. We provide a detailed discussion of all such computational matters in Section 6.2.

[6] We initialize N_1 at 1 to avoid divisions by zero.

The algorithm is in Appendix B.2. Let $\lambda > 0$ be a regularization parameter, $\Lambda_1 = \lambda I$, and $b_1 = 0$. At epoch k, for $t \in \mathcal{T}_k$, the data is stored as $\mathcal{D}_{t+1} = (\Lambda_{t+1}, b_{t+1})$ where $\Lambda_{t+1} = \Lambda_t + \varphi_k(x_t, a_t)\,\varphi_k(x_t, a_t)^\intercal$ and $b_{t+1} = b_t + \varphi_k(x_t, a_t)\,V_k(x_{t+1})$. $\widehat{P}_k = \sum_i \widehat{\theta}_{k,i}\psi_i$ is computed via a least-squares regression, where $\widehat{\theta}_k = \Lambda_{T_k}^{-1} b_{T_k}$. Given $\beta > 0$, the exploration bonuses are defined as

$$\mathrm{CB}_k(x,a) = \beta\,\|\varphi_k(x,a)\|_{\Lambda_{T_k}^{-1}}\,. \tag{9}$$

We now turn to the validity condition required by Lemma 4.3.

Lemma 5.4. *Let $\delta \in (0,1)$. Then, setting the coefficient $\beta = H\sqrt{2\left(\frac{d}{2}\log\left[1+\frac{TB^2H^2}{\lambda d}\right]+\log\frac{1}{\delta}\right)} + \sqrt{\lambda}B$ guarantees that, with probability $1-\delta$, the validity condition (2) is satisfied by CB_k as defined in Equation (9) for all k.*

The proof is in Appendix A.2. It relies on standard techniques regarding linear mixture MDPs (Zhou et al., 2021; Cai et al., 2020). One important property required is the boundedness of each V_k that is guaranteed by the truncation. Then, we can bound the sum of the exploration bonuses with the following lemma.

Lemma 5.5. *The sum of exploration bonuses generated by* ***RAVI-UCB*** *satisfies*

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

The proof (Appendix A.4) 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.

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

Theorem 5.6. *Suppose that* ***RAVI-UCB*** *is run with the uniform policy as π_0, $V_0 = 0$, $\lambda = 1$, a learning rate $\eta = \sqrt{2\log|\mathcal{A}|/(H^2T)}$, and an 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*

$$\mathfrak{R}_T = \widetilde{\mathcal{O}}\left(\sqrt{(d^2H^3 + B^2dH + H^4\log|\mathcal{A}|)\,T}\right).$$

$\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors of T, B, d, and H. A perhaps more useful result is the following, derived from an online-to-batch conversion. Suppose **RAVI-UCB** returns a policy $\pi_{\text{out}} = \pi_U$ with U being an epoch index chosen uniformly at random from the range of epochs. The following corollary provides a guarantee on the quality of this policy.

Corollary 5.7. *Let $\varepsilon > 0$. Then,* ***RAVI-UCB*** *run with the same parameters as before outputs a policy π_{out} satisfying $\mathbb{E}[\langle \mu^* - \mu^{\pi_{out}}, r\rangle] \le \varepsilon$ after T_ε steps, with*

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

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, but the latter is inherent to the online-to-batch conversion process used by our analysis. We will return to this point in Section 6.2.

6. Discussion

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

6.1. Results and Comparisons

There are many differences between our approach and 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 DP, 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) (see Theorem 16). Either way, all of

these analyses use tools from the analysis of mirror descent first developed by Martinet (1970), Rockafellar (1976), and Nemirovski & Yudin (1983) (see also Beck & Teboulle, 2003). Note that, as the guarantees of these regularization-based methods 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 $\sum_{t=1}^{T} \left(V^{*}\left(X_t\right) - V^{\pi_t}\left(X_t\right)\right)$ *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.

6.2. Limitations and Future Directions

On a related note, our method suffers from the limitation of requiring access to a reset action taking the agent back to the initial distribution ν_0 at any time. In general, this 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 these issues are inherent to learning in discounted MDPs.

Another limitation is that our guarantees only hold on expectation as opposed to high probability. In fact, several of our results can be strengthened to hold in this stronger sense, albeit at the cost of a more involved analysis. In particular, the only parts of our analysis that need to be changed are Lemmas 4.4 and 5.5, to deal with the randomized epoch schedule. The first of these can be handled via a 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 clarity. On the other hand, Corollary 5.7 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 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 3 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 mixture MDP assumption that we have used. While arguably well-studied in the past years, this model for linear function approximation has limitations that make it rather difficult to adapt to practical scenarios. The biggest is that learning algorithms in this model need access to an oracle to evaluate sums of the form $\sum_{x'} \psi\left(x, a, x'\right) V\left(x'\right)$, which can only be performed efficiently in special cases. Options include assuming that $\psi\left(x, a, \cdot\right)$ 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 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, requiring 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 mixture MDPs are still a rich family of models that in general is incomparable to linear MDPs, and can subsume many interesting models—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.

Acknowledgements

G. Neu was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 950180). Part of this work was done while the second author was visiting the Simons Institute for the Theory of Computing.

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 Tennenholtz, M. R-max: A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct): 213–231, 2002.

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

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.

Ouhamma, 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., Pirotta, 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.

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 4.3

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

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

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

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

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

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

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

A.2. Proof of Lemma 5.4

Let us fix $k \in [K]$, $t \in \{T_k, T_k+1, \ldots, T_{k+1}-1\}$, $\delta \in (0,1)$. We start by recalling the definition of the nominal transition model $\widehat{P}_k$ acting on functions V as $\left(\widehat{P}_k V\right)(x,a) = \left\langle \varphi_V(x,a), \widehat{\theta}_k \right\rangle$, 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

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

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

$$\begin{aligned}
(PV)(x,a) &= \sum_{x' \in \mathcal{X}} P(x'|x,a)\, V(x') && \text{(by definition of } P\text{)} \\
&= \sum_{x' \in \mathcal{X}} \langle \theta, \psi(x,a,x') \rangle\, V(x') && \text{(by Assumption 5.3)} \\
&= \left\langle \theta, \sum_{x' \in \mathcal{X}} \psi(x,a,x')\, V(x') \right\rangle \\
&= \langle \theta, \varphi_V(x,a) \rangle,
\end{aligned}$$

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

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

where we used the definition of Λ_{T_k} and Assumption 5.3 in the last line. Comparing the expressions for PV and $\widehat{P}_k V$, we obtain

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

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

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

where $\xi_k = \left\| \sum_{i=0}^{k-1} \sum_{j=T_i}^{T_{i+1}-1} \varphi_i(x_j,a_j)\left[V_i(x_{j+1}) - (PV_i)(x_j,a_j)\right] \right\|_{\Lambda_{T_k}^{-1}}$, and $b_k = \lambda \|\theta\|_{\Lambda_{T_k}^{-1}}$. The second term can be easily bounded as $|b_k| \le \sqrt{\lambda}\|\theta\|_2 \le \sqrt{\lambda} B$, using that $\lambda_{\min}(\Lambda_{T_k}) \ge \lambda$ 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 $\|\varphi_i(x_j,a_j)\|_2 \le BH$ and the true parameter as $\|\theta\|_2 \le B$ by Assumption 5.3. Therefore, we can apply the self-normalized concentration result in Theorem C.2 (stated in Appendix C.2), which guarantees that with probability at least $1-\delta$, the following bound holds for all $k \in [K]$:

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

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

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

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

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

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

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

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

$$\beta = H\sqrt{2\left(\frac{d}{2}\log\left[1 + \frac{TB^2H^2}{\lambda d}\right] + \log\frac{1}{\delta}\right)} + \sqrt{\lambda}B. \tag{10}$$

This concludes the proof. □

A.3. Proof of Lemma 4.4

For the sake of this proof, we slightly update our notation for $\mathcal{T}_k$ by setting $\mathcal{T}_{K(T)} = \left\{T_{K(T)}, T_{K(T)}+1, \dots, T^+\right\}$. 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

$$\sum_{t=1}^{T^+} \mathrm{CB}_t\left(X_t, A_t\right) = \sum_{k=1}^{K(T)} \sum_{t \in \mathcal{T}_k} \mathrm{CB}_t\left(X_t, A_t\right). \tag{11}$$

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 $\left(X_t, A_t\right)_{t \in \mathcal{T}_k}$ given $\mathcal{F}_{k-1}$ as follows. The initial state $X_{k,0}$ is drawn from ν_0, and then subsequently for each $i = 0, 1, \dots$, the actions are drawn as $A_{k,i} \sim \pi_k\left(\cdot | X_{k,i}\right)$ and follow-up states are drawn as $X_{k,i+1} \sim P\left(\cdot | X_{k,i}, A_{k,i}\right)$. Recalling the notational convention that $\mathrm{CB}_t = \mathrm{CB}_k$ for all $t \in \mathcal{T}_k$, we observe that for any k, we have

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

where in the third line we have observed that L_k follows a geometric law with parameter $1-\gamma$, and is independent of $\left(X_{k,i}, A_{k,i}\right)_i$. 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 μ^{π_k} up to the normalization constant $(1-\gamma)$, and finally concluded by observing that $\mathbb{E}\left[L_k | \mathcal{F}_{k-1}\right] = \frac{1}{1-\gamma}$.

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

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

□

A.4. Proof of Lemma 5.5

The proof is based on a classic "pigeonhole" argument often called the "elliptical potential lemma" (*e.g.*, Lemma 19.4 in Lattimore & Szepesvári, 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 4.3 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 $\varphi_{k,t} = \varphi_k\left(X_t, A_t\right)$, $\Lambda_k = \Lambda_{T_k}$. We also define $\mathcal{N}(T) = \left\{t \in [T] : \left\|\varphi_{k_t,t}\right\|_{\Lambda_{k_t}^{-1}} \ge 1\right\}$ 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:

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

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:

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

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

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

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

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

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

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

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

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

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

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

We put together the two terms, and plug in the definition of β to get

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

where the two factors are defined as

$$C_1\left(T\right) = \left(H\sqrt{2\left(\frac{d}{2}\log\left[1+\frac{B^2H^2T}{\lambda d}\right]+\log T\right)}+\sqrt{\lambda}B\right)\frac{dBH^2}{\sqrt{\lambda}\log\left(2\right)}\log\left(1+\frac{B^2H^2T}{\lambda d}\right)$$
$$C_2\left(T\right) = \left(H\sqrt{2\left(\frac{d}{2}\log\left[1+\frac{B^2H^2T}{\lambda d}\right]+\log T\right)}+\sqrt{\lambda}B\right)\sqrt{H\left(4+2\log T\right)}\sqrt{d\log\left(1+\frac{B^2H^2T}{\lambda d}\right)}.$$

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

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

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

Proof. We will follow the steps of the proof of Lemma 19.4 in Lattimore & Szepesvári (2020). First, notice that Λ_k can be written as

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

Taking the determinant of the above matrix, we get

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

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

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

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

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

and furthermore by the trace-determinant inequality we have

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

Finally, the trace can be bounded as

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

Plugging this back into the previous inequality proves the claim. □

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 5.5 above.

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

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

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

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

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

$$\log\det\left(G_{K(T)}\right) = \log\det(G_0) + \sum_{k \in \mathcal{E}(T)} \log\det\left(I + \sum_{t \in \mathcal{T}_k \cap \mathcal{N}(T)} \left(G_k^{-1/2} \varphi_{k,t}\right) \left(G_k^{-1/2} \varphi_{k,t}\right)^\intercal \right). \tag{12}$$

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

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

where the equality follows from the fact that $\det(I + vv^\intercal) = (1 + \|v\|_2^2)$ 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

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

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

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

Thus, we have

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

Finally, the trace can be bounded as

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

The proof is concluded by putting this bound together with the previous inequality. □

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 $1-\gamma$, satify 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 $Z_1, \ldots, Z_T$ with parameter $1-\gamma$:

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

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

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

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 $|\mathcal{T}_k|$. □

B. Applications: Algorithm Specifications

For clarity, we provide the complete algorithms in the context of tabular and linear mixture MDPs. The highlighted parts correspond to the instantiations of the functions TRANSITION-ESTIMATE, BONUS, and ADD from Algorithm 1.

B.1. Tabular MDPs

In tabular MDPs, we use the maximum likelihood estimates to compute $\widehat{P}_k$ and the classical count-based bonuses. Therefore, we only need to store and update the counts N_t and N'_t when interacting with the environment.

Algorithm 2 **RAVI-UCB** for tabular MDPs.

Inputs: Horizon T, learning rate $\eta > 0$, confidence parameter $\beta > 0$, value V_0, policy π_0.
Initialize: $t = 1$, $N'_1 = 0$, $N_1 = 1$, $Q_1 = EV_0$.
for $k = 1, \dots$ **do**
  $T_k = t$.
  $\widehat{P}_k(x'|x,a) = N'_{T_k}(x,a,x') / N_{T_k}(x,a)$.
  $V_k(x) = \frac{1}{\eta} \log\left(\sum_a \pi_{k-1}(a|x)\, e^{\eta Q_k(x,a)}\right)$.
  $\pi_k(a|x) = \pi_{k-1}(a|x)\, e^{\eta(Q_k(x,a) - V_k(x))}$.
  $\mathrm{CB}_k(x,a) = \beta / \sqrt{N_{T_k}(x,a)}$.
  $Q_{k+1} = \Pi_H\left[r + \mathrm{CB}_k + \gamma \widehat{P}_k V_k\right]$.
  repeat
    Play $a_t \sim \pi_k(\cdot|x_t)$, and observe x_{t+1}.
    Update $N'_{t+1}(x_t, a_t, x_{t+1}) = N'_t(x_t, a_t, x_{t+1}) + 1$, and $N_{t+1}(x_t, a_t) = N_t(x_t, a_t) + 1$.
    $t = t + 1$.
    With probability $1 - \gamma$, reset to initial distribution: $x_t \sim \nu_0$ and **break**.
  until $t = T$
end for

B.2. Linear Mixture MDPs

In linear mixture MDPs, $\widehat{P}_k$ is computed via a least-squares regression, and we use elliptical bonuses for CB_k. Thus, we only need to store and update the empirical covariance matrix Λ_t and the vector b_t when interacting with the environment.

Algorithm 3 **RAVI-UCB** for linear mixture MDPs.

Inputs: Horizon T, learning rate $\eta > 0$, confidence parameter $\beta > 0$, regularization parameter λ, value V_0, policy π_0.
Initialize: $t = 1$, $\Lambda_1 = \lambda I$, $b_1 = 0$, $Q_1 = EV_0$.
for $k = 1, \dots$ **do**
  $T_k = t$.
  $\widehat{\theta}_k = \Lambda_{T_k}^{-1} b_{T_k}$.
  $\widehat{P}_k = \sum_i \widehat{\theta}_{k,i} \psi_i$.
  $V_k(x) = \frac{1}{\eta} \log\left(\sum_a \pi_{k-1}(a|x)\, e^{\eta Q_k(x,a)}\right)$.
  $\pi_k(a|x) = \pi_{k-1}(a|x)\, e^{\eta(Q_k(x,a) - V_k(x))}$.
  $\varphi_k(x,a) = \sum_{x'} \psi(x,a,x')\, V_k(x')$.
  $\mathrm{CB}_k(x,a) = \beta \left\|\varphi_k(x,a)\right\|_{\Lambda_{T_k}^{-1}}$.
  $Q_{k+1} = \Pi_H\left[r + \mathrm{CB}_k + \gamma \widehat{P}_k V_k\right]$.
  repeat
    Play $a_t \sim \pi_k(\cdot|x_t)$, and observe x_{t+1}.
    Update $\Lambda_{t+1} = \Lambda_t + \varphi_k(x_t, a_t)\, \varphi_k(x_t, a_t)^\intercal$, and $b_{t+1} = b_t + \varphi_k(x_t, a_t)\, V_k(x_{t+1})$.
    $t = t + 1$.
    With probability $1 - \gamma$, reset to initial distribution: $x_t \sim \nu_0$ and **break**.
  until $t = T$
end for

C. Standard Results

C.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 C.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 $x \in \mathcal{X}$:*

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

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

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

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

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

where $Q \in \mathbb{R}^{\mathcal{A}}$, and $p' \in \Delta(\mathcal{A})$. As the probability simplex is compact and $\left(p \mapsto \langle p, Q\rangle - \frac{1}{\eta}\mathcal{D}_{\mathrm{KL}}(p\|p')\right)$ 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

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

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

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

Then, we use the constraint on p^* to find the value of α. 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 $V^* = \frac{1}{\eta}\log\left(\sum_{a\in\mathcal{A}} p'(a)\exp[\eta Q(a)]\right)$, we plug back the expression of α into p^*:

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

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

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

so that we can write

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

The first statement of the lemma then follows from applying this result to $Q = Q_k\left(x,\cdot\right)$ and $p' = \pi_{k-1}\left(\cdot|x\right)$, 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 π_k, we have that the following expressions are equivalent to the ones given in the statement of the lemma:

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

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

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

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

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

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

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

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

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

which concludes the proof. □

C.2. A Self-Normalized Tail Inequality

Theorem C.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 σ-subGaussian; i.e. $\mathbb{E}\left[\eta_t|\mathcal{F}_{t-1}\right] = 0$, and*

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

Let $\{\varphi_t\}_{t=0}^\infty$ be an $\mathbb{R}^d$-valued stochastic process where φ_t is $\mathcal{F}_{t-1}$-measurable. Assume Λ_0 is a $d\times d$ positive definite matrix, and let $\Lambda_t = \Lambda_0 + \sum_{s=1}^t \varphi_s\varphi_s^\intercal$. Then, for any $\delta > 0$, with probability at least $1-\delta$, we have for all $t \geq 0$,

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