Optimistically Optimistic Exploration for Provably Efficient Infinite-Horizon Reinforcement and Imitation Learning

Antoine Moulin
Universitat Pompeu Fabra
antoine.moulin@upf.edu

Gergely Neu
Universitat Pompeu Fabra
gergely.neu@gmail.com

Luca Viano
EPFL
luca.viano@epfl.ch

Abstract

We study the problem of reinforcement learning in infinite-horizon discounted linear Markov decision processes (MDPs), and propose the first computationally efficient algorithm achieving near-optimal regret guarantees in this setting. Our main idea is to combine two classic techniques for optimistic exploration: additive exploration bonuses applied to the reward function, and artificial transitions made to an absorbing state with maximal return. We show that, combined with a regularized approximate dynamic-programming scheme, the resulting algorithm achieves a regret of order $\tilde{\mathcal{O}}(\sqrt{d^3(1-\gamma)^{-7/2}T})$, where T is the total number of sample transitions, $\gamma \in (0,1)$ is the discount factor, and d is the feature dimensionality. The results continue to hold against adversarial reward sequences, enabling application of our method to the problem of imitation learning in linear MDPs, where we achieve state-of-the-art results.

1 Introduction

Since the breakthrough work of Jin et al. [2019], the class of linear Markov decision processes (MDPs) has become a standard model for theoretical analysis of reinforcement learning (RL) algorithms under linear function approximation. This work demonstrated the possibility of constructing computationally and statistically efficient methods for large-scale RL, and pioneered an analysis technique that influenced the entire field of RL theory. Hundreds of follow-up papers have studied variations of this model, studying extensions such as learning with adversarial rewards [Neu and Olkhovskaya, 2021, Zhong and Zhang, 2024, Sherman et al., 2023b, Dai et al., 2023, Sherman et al., 2023a, Cassel and Rosenberg, 2024, Liu et al., 2023], without rewards [Wang et al., 2020, Wagenmaker et al., 2022, Hu et al., 2022], or with unknown features Agarwal et al. [2020], Uehara et al. [2021], Zhang et al. [2022], Mhammedi et al. [2024], Modi et al. [2024]. The linearity constraint itself has been relaxed in a variety of ways Zanette et al. [2020], Cai et al. [2020], Du et al. [2021], Weisz et al. [2024], Golowich and Moitra [2024], Wu et al. [2024b]. However, practically all of these developments retained one major limitation of the original work of Jin et al. [2019]: it only applies to finite-horizon MDPs. Generalizations to the more challenging (and practically much more popular) infinite-horizon MDP models have so far remained very limited, yielding only highly impractical methods or suboptimal performance guarantees [Wei et al., 2021]. In this paper, we propose an efficient algorithm that successfully addresses this long-standing open problem.

We consider the problem of learning a nearly optimal policy in γ-discounted MDPs [Puterman, 1994], under the linear MDP assumption first proposed by Jin et al. [2019] (see also Yang and Wang, 2019). We consider an interaction protocol where a learning agent interacts with the environment in a sequence of K episodes of geometrically distributed length, and aims to pick a sequence of policies such that its regret against the best fixed policy is as small as possible. Our algorithm achieves a regret

18th European Workshop on Reinforcement Learning (EWRL 2025).

bound of order $H\sqrt{d^3T} + H^{7/4}\sqrt{dT\log|\mathcal{A}|}$, where d is the feature dimensionality, $H = \frac{1}{1-\gamma}$ is the effective horizon, $\mathcal{A}$ is the action space, and T is the number of interactions. This implies a bound on the sample complexity of learning an ε-optimal policy of the order $\frac{H^3d^3+H^{7/2}d\log|\mathcal{A}|}{\varepsilon^2}$. The algorithm returns a single softmax policy that is fully described in terms of a d-dimensional parameter vector and a d^2-dimensional feature-covariance matrix. This constitutes the first sample-complexity result of the optimal order $1/\varepsilon^2$ achieved by a computationally efficient algorithm. The regret guarantees are also shown to hold if the reward function changes adversarially over time, and we additionally provide an extension of our method for the setting of imitation learning.

On the technical side, our main contribution is the development of a new optimistic exploration mechanism that combines two classic ideas from two different eras of RL theory. First, following the recipe of Jin et al. [2019], we make use of additive UCB-style exploration bonuses which have been successfuly used for several decades in both bandit problems [Lai and Robbins, 1985, Auer et al., 2002, Auer, 2002, Dani et al., 2008, Abbasi-Yadkori et al., 2011] and reinforcement learning [Kaelbling et al., 1996, Strehl and Littman, 2008, Jaksch et al., 2010, Azar et al., 2017]. Second (and more importantly), we adapt another classic (but apparently recently less well-known) idea underlying the `Rmax` algorithm of Brafman and Tennenholtz [2002] (see also Szita and Szepesvári, 2010 and Chapter 8 in Kakade, 2003). Roughly speaking, this technique amounts to replacing the standard empirical model estimate with a fixed optimistic estimate in state-action pairs that are very poorly explored. This addresses the notorious problem of empirical estimates in linear MDPs that they tend to have extremely high variance in under-explored states, which can only be offset with very large additive exploration bonuses. Our `Rmax`-style scheme counteracts these large bonuses by effectively swapping out the possibly over-optimistic estimates obtained via additive bonuses with more reasonably sized estimates. Besides Brafman and Tennenholtz [2002], our algorithm design and analysis is also strongly inspired by the recent work of Cassel and Rosenberg [2024] who proposed a slightly limited variant of the same exploration mechanism for finite-horizon MDPs.

2 Preliminaries

In this section, we first provide the general definitions that will repeatedly appear throughout the paper, and then go on to describe a set of ideas that will be heavily featured in our algorithm design and analysis. We finally describe the concrete learning setting in detail at the end of the section.

2.1 Markov decision processes

A Markov decision process (MDP) with reward function r is defined by the tuple $\mathcal{M}(r) = (\mathcal{X}, \mathcal{A}, r, P, \gamma, \nu_0)$, where $\mathcal{X}$ is the (possibly infinite) state space, $\mathcal{A}$ is the finite action space, $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0,1]$ is the reward function assigning rewards to each state-action-next-state transition, $P : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$ is the transition kernel, $\gamma \in (0,1)$ is the discount factor, and $\nu_0 \in \Delta(\mathcal{X})$ is the initial-state distribution. For convenience, we will assume that $\mathcal{X}$ is countable but note that this can be lifted at the expense of making the measure-theoretic notation much heavier. The MDP $\mathcal{M}(r)$ models a sequential decision-making problem between a decision-making *agent* and its *environment*. The interaction starts with the environment drawing the random initial state $X_0 \sim \nu_0$, whereafter in each time step $t = 0, 1, 2, \ldots$, the following steps are repeated: the agent observes state $X_t \in \mathcal{X}$, takes action $A_t \in \mathcal{A}$, and consequently the environment generates the next state $X_{t+1} \sim P(\cdot|X_t, A_t)$, resulting in reward $R_t = r(X_t, A_t, X_{t+1})$. With a slight abuse of notation, we denote the mean reward of a state-action pair $(x,a) \in \mathcal{X} \times \mathcal{A}$ by $r(x,a) = \mathbb{E}_{X' \sim P(\cdot|x,a)}[r(x,a,X')]$.

A *stationary state-feedback policy* (or, in short, *policy*) is a randomized behavior rule $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ that determines the action taken in each time step t as $A_t \sim \pi(\cdot|X_t)$. The *action-value function* of a policy π in $\mathcal{M}$ is defined for any state-action pair (x,a) as

$$Q^{\pi}_{P,r}(x,a) = \mathbb{E}_{P,\pi}\left[\sum_{\tau=0}^{\infty} \gamma^{\tau} r(X_\tau, A_\tau) \middle| (X_0, A_0) = (x,a)\right],$$

where $\mathbb{E}_{P,\pi}$ denotes the expectation with respect to the random sequence of states and actions generated by the transition kernel P and the policy π. The *value function* of π at state x is defined as $V^{\pi}_{P,r}(x) = \mathbb{E}_{A\sim\pi(\cdot|x)}\left[Q^{\pi}_{P,r}(x,A)\right]$. With some abuse of notation, we define the conditional expectation operator $P : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ via its action $\left(Pf\right)(x,a) = \mathbb{E}_{X'\sim P(\cdot|x,a)}[f(X')]$ for any function $f \in \mathbb{R}^{\mathcal{X}}$ and state-action pair (x,a). Its adjoint $P^{\intercal}$ is the operator that acts on distributions

$\mu \in \Delta(\mathcal{X} \times \mathcal{A})$ as $P^\intercal \mu = \mathbb{E}_{(X,A)\sim\mu}[P(\cdot|X,A)]$. With this notation, the value functions can be shown to satisfy the *Bellman equations* written as

$$Q^\pi_{P,r} = r + \gamma P V^\pi_{P,r}\,.$$

For convenience, we also introduce the operator $E : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ defined via $\left(Ef\right)(x,a) = f(x)$ and whose adjoint acts on state-action distributions as $\left(E^\intercal \mu\right)(x) = \sum_{a\in\mathcal{A}} \mu(x,a)$. When interacting with an MDP, any stationary policy π induces a unique *state-occupancy measure* denoted as $\nu(\pi) \in \Delta(\mathcal{X})$ and a state-action occupancy measure $\mu(\pi) \in \Delta(\mathcal{X}\times\mathcal{A})$ defined (with an unusual but helpful abuse of notation) as

$$\nu(\pi,\cdot) = (1-\gamma)\sum_{\tau=0}^{\infty}\gamma^\tau \mathbb{P}_{P,\pi}[X_\tau \in \cdot] \quad \text{and} \quad \mu(\pi,\cdot) = (1-\gamma)\sum_{\tau=0}^{\infty}\gamma^\tau \mathbb{P}_{P,\pi}[(X_\tau,A_\tau)\in\cdot]\,.$$

2.2 Optimistically augmented Markov decision processes

A key concept in our algorithm design is that of *optimistically augmented Markov decision processes* (OA-MDPs), inspired by the construction of Brafman and Tennenholtz [2002]. The OA-MDP associated with $\mathcal{M}(r)$ is defined on the augmented state space $\mathcal{X}^+ = \mathcal{X} \cup \{x^+\}$, where x^+ is an artificial *heaven* state appended to the original set of states. The transition dynamics are defined via a perturbation of the original transition function, governed by the *ascension function* $p^+ : \mathcal{X}^+ \times \mathcal{A} \to [0,1]$. In particular, the transition kernel from state-action pair x, a to x' is defined as

$$P^+(\cdot|x,a) = \big(1-p^+(x,a)\big)P(\cdot|x,a) + p^+(x,a)\,\mathbb{I}_{\{x^+\in\cdot\}}\,.$$

In words, the sequence of states in the augmented MDP follows the dynamics of the original process, except that the process *ascends* to heaven with probability $p^+(X_t,A_t)$ in round t. The augmented reward function is the same for all triples in the original MDP x, a, x', and ascension to heaven results in maximal reward $r(x,a,x^+) = R_{\max}$. The resulting state-action reward function is then

$$r^+(x,a) = \mathbb{E}_{X'\sim P^+(\cdot|x,a)}[r(x,a,X')] = \big(1-p^+(x,a)\big)r(x,a) + p^+(x,a)\,R_{\max}\,.$$

Once the process enters x^+, it remains there forever (*i.e.*, $P^+(\{x^+\}|x^+,a) = 1$ for all actions a) and obtains maximal reward $R_{\max}$ in each round. Without loss of generality (and for notational convenience), we will assume throughout the state x^+ also exists in the original MDP $\mathcal{M}(r)$, but is not reachable either via regular transitions ($P(\{x^+\}|x,a) = 0$) or initialization ($\nu_0(\{x^+\}) = 0$). We will also follow the convention that $p^+(x^+,a) = 0$ for all actions a. We will refer to the optimistically augmented MDP as $\mathcal{M}^+(r,p^+)$, and illustrate the relation of the two processes in Figure 1.

Our algorithm and its analysis will feature a sequence of ascension functions denoted by p^+_k, and the associated transition function will be denoted by P^+_k. Within the augmented MDP induced by p^+_k, we denote the value functions of a policy π in $\mathcal{M}^+(r,p^+_k)$ as $V^\pi_{P^+_k,r^+}$ and $Q^\pi_{P^+_k,r^+}$. Likewise, we will use $\nu^+_k(\pi)$ and $\mu^+_k(\pi)$ to refer to the occupancy measures of π in $\mathcal{M}^+(r,p^+_k)$. It is easy to see that for any policy π, the value functions satisfy $V^\pi_{P^+_k,r^+} \geq V^\pi_{P,r}$ and $Q^\pi_{P^+_k,r^+} \geq Q^\pi_{P,r}$, which explains why we call the resulting MDP "optimistic". Furthermore, for all non-heaven states x, we have $\nu^+_k(\pi,x) \leq \nu(\pi,x)$ and $\mu^+_k(\pi,x,a) \leq \mu^+(\pi,x,a)$. Our analysis will heavily rely on these facts (which will be proved formally later).

2.3 Online learning in linear MDPs

We consider a variation of the MDP setup described above, incorporating two modifications: *i)* periodic resets of the state evolution to the initial-state distribution ν_0, and *ii)* the ability of the environment to change the reward function adversarially after each reset. This is a natural adaptation[1] of the well-explored setting of online learning in adversarial MDPs to the discounted-reward case we study in this work. More precisely, we consider the following sequential interaction process between the learning agent and its environment. The interaction proceeds through T time steps, organized into K episodes (of random length) as follows. The initial state drawn as $X_0 \sim \nu_0$ and then the following steps are repeated in every consecutive round $t = 0, 1, \ldots, T$:

[1] See Section 6 for a discussion of the role of resets and related online-learning settings.

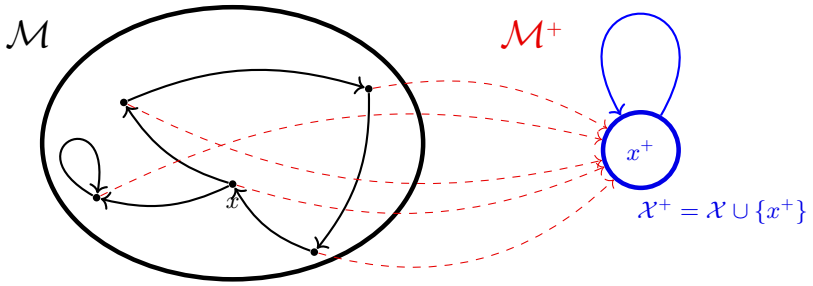

Figure 1: Illustration of the MDP $\mathcal{M}$ in black and its extension in blue. The MDP $\mathcal{M}^+$ contains the additional red dashed edges that allow ascension to heaven.

- The agent observes the state $X_t \in \mathcal{X}$,
- the agent chooses an action $A_t \in \mathcal{A}$,
- the environment generates the next state $X'_{t+1} \sim P(\cdot|X_t, A_t)$,
- the environment selects a reward function r_t,
- the agent receives a reward $r_t(X_t, A_t) \in [0, 1]$ and observes the function r_t,
- with probability γ, the process moves to the next state $X_{t+1} = X'_{t+1}$, otherwise a new episode begins and the process is reset to the initial-state distribution as $X_{t+1} \sim \nu_0$.

Without significant loss of generality, we will restrict the environment to update the reward function only at the end of each episode, and use r_k to refer to the reward function within episode k. Other than this restriction, the environment is free to choose the rewards in an adaptive (and possibly adversarial) way. The objective for the agent is to select a sequence of policies π_k so as to minimize its *pseudo-regret* over K episodes with respect to an arbitrary comparator policy $\pi^\star : \mathcal{X}^+ \to \Delta(\mathcal{A})$, given by

$$\mathfrak{R}_K = \sum_{k=1}^{K} \left\langle \nu_0, V^{\pi^\star}_{P,r_k} - V^{\pi_k}_{P,r_k} \right\rangle = \frac{1}{1-\gamma} \sum_{k=1}^{K} \left\langle \mu(\pi^\star) - \mu(\pi_k), r_k \right\rangle ,$$

where the second equality follows from the definition of occupancy measures. Since the learning agent can only learn about the transition function via interaction with the environment, it needs to address the classic dilemma of exploration versus exploitation. Clearly, this setup generalizes the more standard problem formulation where $r_k = r$ holds for all episodes k. In this case, $\frac{\mathfrak{R}_K}{K}$ corresponds to the expected suboptimality of the average policy played by the agent.

In later sections, we will consider the following structural assumption on the transitions.

Assumption 1 (Linear MDP) *A discounted MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, r, P, \gamma, \nu_0)$ is a* linear MDP *if there exist a known feature map $\varphi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, an unknown map $m : \mathcal{X} \to \mathbb{R}^d$ and an unknown vector $w \in \mathbb{R}^d$ such that for any triplet $(x, a, x') \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$,*

$$P(x'|x, a) = \langle \varphi(x, a), m(x') \rangle , \quad r(x, a) = \langle \varphi(x, a), w \rangle .$$

We will also use the operators $\Phi : \mathbb{R}^d \to \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ such that $(\Phi\theta)(x, a) = \langle \theta, \varphi(x, a) \rangle$ holds for any $\theta \in \mathbb{R}^d$ and $M : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^d$ such that $(Mf)_i = \int_{\mathcal{X}} f(x) \mathrm{d}m_i(x')$. Thus, we can write the transition operator and the reward function as $P = \Phi M$ and $r = \Phi w$. Moreover, we assume that $\|w\| \le W_{\max}$ and that for all $x, a \in \mathcal{X} \times \mathcal{A}$, the features have bounded norm, i.e. $\|\varphi(x, a)\| \le B$.

Further notation. We will use $\pi_{\texttt{unif}}$ to denote both the uniform probability distribution over $\mathcal{A}$ and the policy that plays uniformly at random at every state. $\Delta(A)$ denotes the simplex over a discrete set A. Given two distributions $p, q \in \Delta(\mathcal{Z})$ on the countable set $\mathcal{Z}$, we denote the Kullback–Leibler divergence as $\mathcal{D}_{\mathrm{KL}}(p\|q) = \sum_{z\in\mathcal{Z}} \log\left(\frac{p(z)}{q(z)}\right) p(z)$ and we use the convention that $\mathcal{D}_{\mathrm{KL}}(p\|q) = +\infty$ whenever there exists an element $z \in \mathcal{Z}$ such that $q(z) = 0$ and $p(z) > 0$. For a distribution $p \in \Delta(\mathcal{Z})$ and a function $f \in \mathbb{R}^{\mathcal{Z}}$, we will use the notation $\langle p, f \rangle = \mathbb{E}_{Z\sim p}[f(Z)]$.

3 Algorithm

Our algorithm implements the principle of *optimism in the face of uncertainty* (OFU), by combining two classic ideas for optimistic exploration in reinforcement learning. We refer to these two separate mechanisms as two *degrees of optimism*, with *first-degree optimism* defined using the idea of exploration bonuses added to the rewards, and *second-degree optimism* leveraging the notion of optimistically augmented MDPs defined in Section 2.2. These two incentives for exploration are respectively inspired by the upper-confidence-bound (UCB) methods popularized by Azar et al. [2017], and the classic `Rmax` algorithm of Brafman and Tennenholtz [2002]. These two mechanisms are combined with the regularized approximate dynamic programming method of Moulin and Neu [2023], called "regularized approximate value iteration with upper confidence bounds" (`RAVI-UCB`).

3.1 Overview

We begin by describing each element of our solution in general terms, and provide the pseudocode of the resulting algorithm (specifically tailored to linear MDPs) as Algorithm 1. In recognition of the influence of the two algorithms mentioned above, we refer to our method as `Rmax-RAVI-UCB`.

Regularized dynamic programming. If the transition kernel P were known, the learner could achieve low regret by deploying the following regularized value iteration (RVI) scheme:

$$Q_{k+1} = r_k + \gamma P V_k, \quad V_{k+1}(x) = \max_{u \in \Delta_{\mathcal{A}}} \left\{ \langle u, Q_{k+1}(x, \cdot) \rangle - \frac{1}{\eta} \mathcal{D}_{\mathrm{KL}} \left(u \| \pi_k(\cdot|x) \right) \right\}, \tag{1}$$

and update its policies as $\pi_{k+1}(x|a) \propto \pi_k(x|a) e^{\eta Q_{k+1}(x,a)}$ for some positive learning-rate parameter $\eta > 0$. As observed by Moulin and Neu [2023], the regularization in the policy updates is helpful for controlling the difference between consecutive policies and occupancy measures, thus addressing a major challenge that one faces when analyzing approximate DP methods in infinite-horizon MDPs. Unfortunately, P is unknown and needs to be estimated. The estimation error introduced in this process is taken care of by the two degrees of optimistic adjustments we describe next.

First-degree optimism. Our method will make use of a (possibly implicitly defined) sequence of estimates of the transition operator $\widehat{P}_k : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$, and an associated sequence of *exploration bonuses* $\mathrm{CB}_k : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. We say that the sequence of bonuses is *valid* if it satisfies $\left| \left((\widehat{P}_k - P) V_k \right)(x, a) \right| \leq \mathrm{CB}_k(x, a)$ holds for all value-function estimates V_k calculated by the algorithm, simultaneously for all x, a and k. Using this property, a key idea in our algorithm is to use $\widehat{P}_k V_k + \mathrm{CB}_k$ as an upper confidence bound on $P V_k$, therefore providing an optimistic estimate of the ideal value-function updates (1).

Second-degree optimism. Unfortunately, relying only on first-degree optimism as defined above may result in value estimates that grow without bounds, thus leading to unstable policy updates. In order to prevent this unbounded growth, we employ the idea of optimistically augmented MDPs (defined in Section 2.2), with the ascension function defined as $p_k^+(x, a) = \sigma\left(\alpha\, \mathrm{CB}_k\left(x, a\right) - \omega\right)$, where $\sigma : z \mapsto \frac{e^z}{1+e^z}$ is the sigmoid function and $\alpha > 0$ and $\omega > 0$ are positive hyperparameters. Technically, this is implemented by defining our action-value updates for each x, a as

$$Q_{k+1}\left(x, a\right) = \left(1 - p_k^+(x, a)\right) \left(r_k\left(x, a\right) + \mathrm{CB}_k\left(x, a\right) + \gamma \widehat{P}_k V_k\left(x, a\right) \right) + p_k^+(x, a)\, \frac{R_{\max}}{1 - \gamma}\,.$$

This adjustment makes sure that the value estimates remain bounded, thanks to the multiplicative effect of the ascension function that effectively trades off the possibly huge values of $\mathrm{CB}_k + \gamma \widehat{P}_k V_k$ by the constant upper bound $R_{\max}/(1-\gamma)$ in highly uncertain state-action pairs. Additionally, supposing that CB_k is a valid sequence of exploration bonuses, one can verify that the inequality $Q_{k+1} \geq r_k^+ + \gamma P_k^+ V_k$ holds elementwise—that is, the estimates Q_k provide upper bounds on the ideal action-value updates (1) defined in the optimistically augmented MDP.

3.2 Technical details

To complete the outline given above, we specify the missing technical details specific to linear MDPs.

Model estimation and bonus design. For estimating the transition model and defining the exploration bonuses, we follow the classic approach of Jin et al. [2019] (see also Neu and Pike-Burke, 2020). Specifically, we define the least-squares model estimate $\widehat{P}_k = \Phi \widehat{M}_k$ as the operator that maps a function $v \in \mathbb{R}^{\mathcal{X}}$ to $\Phi \widehat{M}_k v \in \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$. The vector $\widehat{M}_k v \in \mathbb{R}^d$ is the solution to the least-squares regression problem with features $\{\varphi(X_t, A_t)\}_{t=1}^{T_k - 1}$ and targets $\{v(X_t)\}_{t=2}^{T_k}$, where T_k denotes the beginning of episode k. The problem (with target $v = V_k$) admits the closed form solution given in line 25 of Algorithm 1. The matrix $\Lambda_{T_k} = \sum_{t=1}^{T_k} \varphi(X_t, A_t)\varphi(X_t, A_t)^\intercal + I$ used in the computation of the least-squares model estimate is called the *empirical covariance matrix*. Finally, we define the *exploration bonuses* for each $(x, a) \in \mathcal{X} \times \mathcal{A}$ as $\mathrm{CB}_k(x, a) = \beta \left\|\varphi(x, a)\right\|_{\Lambda_{t_e}^{-1}} = \beta \sqrt{\left\langle \varphi(x, a), \Lambda_{t_e}^{-1} \varphi(x, a) \right\rangle}$, where $\beta > 0$ will be chosen during the analysis to be large enough to guarantee bonus validity, and t_e is the time step marking the beginning of the eth epoch (see below). The state x^+ is given special treatment: for all actions a, we fix $\mathrm{CB}_k(x^+, a) = 0$, $p_k^+(x^+, a) = 0$, and $Q_k(x^+, a) = V_k(x^+) = \frac{R_{\max}}{1-\gamma}$.

Bookkeeping. In order to turn the above ideas into a tractable algorithm amenable to theoretical analysis, a few additional bookkeeping steps are necessary. The most important of these is the introduction of an *epoch schedule*, which is instrumental in keeping the complexity of the exploration bonuses and the policies low. Using a classic trick from Abbasi-Yadkori et al. [2011], a new epoch is started every time that there is a significant reduction in the uncertainty of the model estimates (as measured by the determinant of the empirical covariance matrix of the features in the case of linear MDPs—see line 16 in Algorithm 1). When a epoch change is triggered, the exploration bonuses are recomputed and the policy is reset to a uniform policy.

3.3 Discussion

Before moving to the the analysis, we highlight some important features of our method.

Computational and storage complexity. Due to its design outlined above (and detailed in Algorithm 1), `Rmax-RAVI-UCB` produces a sequence of policies that are simply parameterized by a d-dimensional vector and a d^2-dimensional covariance matrix. Specifically, this is made possible by keeping the bonus function fixed within each epoch and resetting the policy to uniform at the beginning of each new epoch. Therefore, storing the policies in memory and drawing actions in each new state can be both done efficiently. This should be contrasted with most other known algorithms making use of softmax policies, which crucially rely on clipped value-function estimates that cannot be stored or sampled from efficiently (as they require storing the entire history of parameter vectors and exploration bonuses). Examples of such methods include Cai et al. [2020], Zhong and Zhang [2024], Sherman et al. [2023b], Moulin and Neu [2023]. This not only makes the implementation of these algorithms impractical, but also results in suboptimal regret guarantees due to the excessive complexity of the policy and value-function classes. This major improvement is made possible in our algorithm by the incorporation of second-degree optimism (inspired by both Brafman and Tennenholtz, 2002 and Cassel and Rosenberg, 2024), which obviates the need for explicit clipping of the value estimates and keeps these bounded via alternative means. All other elements in our algorithm (such as estimating the value estimates via least-squares regression) are standard, and match the complexity of other efficient methods for online learning in linear MDPs Jin et al. [2019], Wang et al. [2021], He et al. [2023].

Relation with existing algorithms. Being a combination of `Rmax` and `RAVI-UCB`, our algorithm can recover these two extremes and several other known methods by an appropriate choice of hyperparameters. Setting $\mathrm{CB}_k = 0$, we recover algorithms which leverage only *second-degree* optimism. For example, in the particular case of $\alpha = \infty$ and tabular features, we recover `Rmax` up to some very minor changes [Brafman and Tennenholtz, 2002, Szita and Szepesvári, 2010]. On the other hand, using the ascension function $p_k^+(x, a) = \mathbb{1}\left\{ r_k(x, a) + \mathrm{CB}_k(x, a) + \gamma \widehat{P}_k V_k(x, a) \geq \frac{R_{\max}}{1-\gamma} \right\}$ essentially recovers the standard truncation rule applied by most related methods (under the condition that the exploration bonuses be valid). In particular, under this choice of p_k^+, our algorithm reduces to `RAVI-UCB`. Setting the regularization parameter $\eta = \infty$ in the resulting method recovers optimistic value iteration methods such as Azar et al. [2017], Jin et al. [2019].

4 Main Result and Analysis

The following theorem states our main result about the performance of `Rmax-RAVI-UCB`.

Theorem 1 *Suppose that Assumption 1 holds, and that Algorithm 1 is executed with parameters specified in Appendix B.7 for a fixed number K of episodes. Then, with probability at least $1-\delta$,*

$$\mathfrak{R}_K = \tilde{\mathcal{O}}\left(\sqrt{d^3H^3K} + \sqrt{dH^{9/2}K\log\left(|\mathcal{A}|\right)}\right).$$

For the commonly studied case where the reward function r is fixed, this result can be easily translated to a bound on the sample complexity of producing an ε-optimal policy as well, as stated below.

Corollary 1 *Let I be drawn uniformly from $\{1,2,\ldots,K\}$. Then, policy π_I produced by Algorithm 1 (with the same parameter tuning as in Theorem 1) satisfies $\mathbb{E}_I\left[V^{\pi^*}_{P,r} - V^{\pi_I}_{P,r}\right] \leq \varepsilon$, if the number of episodes satisfies*

$$K = \Omega\left(\frac{H^3d^3 + H^{9/2}d\log|\mathcal{A}|}{\varepsilon^2}\right).$$

As the length of each episode is geometrically distributed with expectation H, the number of interaction steps satisfies $\mathbb{E}[T] = HK$ when the number of episodes K is fixed. Taking this into account, both results can be restated in terms of the number of sample transitions T. Likewise, similar results can be proved when treating the sample size T as fixed and letting K be the smallest (random) number of episodes covering the sample budget.

In the remaining part of this section, we describe the main steps constituting the proof of Theorem 1. The analysis will make crucial use of the notion of optimistically augmented MDPs defined in Section 2.2. Specifically, we define an augmented MDP for each episode k as $\mathcal{M}^+_k = \mathcal{M}^+(r_k, p^+_k)$ and use the shorthand $\mathcal{M}_k = \mathcal{M}(r_k)$ for the true MDP with reward function r_k. Letting $\mu^+_k(\pi)$ denote the occupancy measure induced by policy π in $\mathcal{M}^+_k$, the first step in our analysis is to rewrite the regret as follows:

$$\begin{aligned}\sum_{k=1}^{K}\left\langle\mu\left(\pi^\star\right)-\mu\left(\pi_k\right), r_k\right\rangle = &\sum_{k=1}^{K}\underbrace{\left\langle\mu\left(\pi^\star\right)-\mu\left(\pi_k\right), r_k - r^+_k\right\rangle}_{=\texttt{reward-bias}_k} + \sum_{k=1}^{K}\underbrace{\left\langle\mu\left(\pi^\star\right)-\mu^+_k\left(\pi^\star\right), r^+_k\right\rangle}_{=-\texttt{model-bias}_k(\pi^\star)} \\ &+ \underbrace{\sum_{k=1}^{K}\left\langle\mu^+_k\left(\pi^\star\right)-\mu^+_k\left(\pi_k\right), r^+_k\right\rangle}_{=\mathfrak{R}^+_K} + \sum_{t=1}^{T}\underbrace{\left\langle\mu^+_k\left(\pi_k\right)-\mu\left(\pi_k\right), r^+_k\right\rangle}_{=\texttt{model-bias}_k(\pi_t)}.\end{aligned}$$

Here, $\mathfrak{R}^+_K$ corresponds to the (normalized) regret of `RAVI-UCB` in the sequence of optimistically augmented MDPs $(\mathcal{M}^+_k)$, and the other terms account for the difference between this term and the regret in the original problem. Some of these are easy to handle by exploiting the optimistic nature of our augmentation technique, whereas others require a more careful tuning of the ascension functions and exploration bonuses.

A large part of the analysis will be based on the condition that the exploration bonuses $\{\mathrm{CB}_k\}_k$ used for performing optimistic value iteration are *valid* estimates of the uncertainty we have on the model, in the sense that the following event holds

$$\mathcal{E}_{\text{valid}} = \left\{\forall(x,a)\in\mathcal{X}\times\mathcal{A}, \forall k\in[K], \left|\left(P-\widehat{P}_k\right)V_k(x,a)\right| \leq \mathrm{CB}_k(x,a)\right\}. \tag{2}$$

For the sake of clearly presenting the main ideas of our analysis, we will work under the condition that the event $\mathcal{E}_{\text{valid}}$ holds, and we will show later in the context of linear MDPs that this is indeed true with high probability. Furthermore, we will work under the condition that for any $k\in[K]$, there exists $Q_{\max} > 0$ such that $\|Q_k\|_\infty \leq Q_{\max}$. We will show this is true under an appropriate choice of p^+_k, and give an expression for $Q_{\max}$. Finally, we assume the episode lengths (L_k) are all less than some $L_{\max}$ and refer to this event as $\mathcal{E}_L$. We will show that this also holds with high probability.

4.1 Controlling the bias due to the augmented rewards

The first lemma controls the bias introduced by the second-degree optimism introduced into the reward function in terms of the ascension functions.

Lemma 1 *For any choice of p_k^+, we have* $\texttt{reward-bias}_k \leq R_{\max} \langle \mu(\pi_k), p_k^+ \rangle$.

The proof essentially follows from the definition of r_k^+, and is found in Appendix B.1.

4.2 Controlling the bias due to the augmented transitions

Next, we provide a bound on the bias due to introducing second-degree optimism in the transition function. It is easy to see that playing in the augmented MDP $\mathcal{M}_k^+$ always yields higher discounted return due to the presence of the heaven state x^+. On the other hand, one can intuitively see that the difference in the bias term can be upper bounded by the amount of time spent in heaven x^+. The following lemma formalizes this claim by providing a bound on $\texttt{model-bias}_k(\pi) = \langle \mu_k^+(\pi) - \mu(\pi), r_k^+ \rangle$ for any policy π.

Lemma 2 *Let π be any policy and p_k^+ be any ascension function at episode k. Then,*

$$0 \leq \texttt{model-bias}_k(\pi) \leq \frac{R_{\max}}{1-\gamma} \langle \mu(\pi), p_k^+ \rangle .$$

Both bounds are proved through a coupling argument, provided in Appendix B.2. Applying the lower bound to $\pi^\star$ and the upper bound to π_k, we obtain

$$-\sum_{k=1}^{K} \texttt{model-bias}_k(\pi^\star) \leq 0, \text{ and } \sum_{k=1}^{K} \texttt{model-bias}_k(\pi_k) \leq \frac{R_{\max}}{1-\gamma} \sum_{k=1}^{K} \langle \mu(\pi_k), p_k^+ \rangle . \quad (3)$$

4.3 Regret analysis in the optimistically augmented MDP

To control the main term $\mathfrak{R}_{\mathrm{K}}^+$, we adapt the analysis of RAVI-UCB due to Moulin and Neu [2023] with some appropriate changes. The key idea is to define an estimate $\widehat{P}_k^+$ of the optimistically augmented transition operator P_k^+ associated with $\mathcal{M}_k^+$, via its action on a function $v \in \mathbb{R}^{\mathcal{X}}$:

$$\left(\widehat{P}_k^+ v\right)(x,a) = (1 - p_k^+(x,a)) \cdot \left(\widehat{P}_k v\right)(x,a) + p_k^+(x,a) \cdot \frac{R_{\max}}{1-\gamma} .$$

Then, it is easy to verify that the validity of the exploration bonuses (Eq. 2) implies that the scaled bonuses $(1-p_k^+) \odot \mathrm{CB}_k$ satisfy the following analogous validity condition in the augmented MDP:

$$\left| \left(P_k^+ - \widehat{P}_k^+\right) V_k(x,a) \right| \leq (1 - p_k^+(x,a)) \, \mathrm{CB}_k(x,a) .$$

With these insights, our algorithm can be seen as an instantiation of RAVI-UCB on the sequence of optimistically augmented MDPs $(\mathcal{M}_k^+)$, and thus it can be analyzed by following the steps of Moulin and Neu [2023]. In particular, the following lemma (an adaptation of Lemma 4.3 of Moulin and Neu [2023], proved here in Appendix B.3) gives a bound on the regret in the augmented MDP.

Lemma 3 *Suppose that the bonuses $\{\mathrm{CB}_k\}$ are valid in the sense of Equation 2 and that for any k, $\|Q_k\|_\infty \leq Q_{\max}$. Then, the sequence of policies output by Algorithm 1 satisfies*

$$\mathfrak{R}_{\mathrm{K}}^+ \leq \frac{E(K) \log |\mathcal{A}|}{\eta} + 4 Q_{\max} E(K) + \frac{2 Q_{\max}^2 \eta K}{\sqrt{1-\gamma}} + 2 \sum_{k=1}^{K} \langle \mu(\pi_k), (1-p_k^+) \odot \mathrm{CB}_k \rangle . \quad (4)$$

4.4 Choosing the ascension functions

It remains to verify that our choice of the probabilities p_k^+ is such that the terms appearing in Lemmas 1 and Equation (3) are small, yet the value of $Q_{\max}$ also remains bounded. In particular, we show in Lemma 10 that our choice satisfies

$$p_k^+(x,a) \leq 2\alpha^2 \, \mathrm{CB}_k(x,a)^2 + 2e^{-\omega} . \quad (5)$$

Furthermore, the following lemma (proved in in Appendix B.4) shows a suitable choice of $Q_{\max}$.

Lemma 4 *Suppose the bonuses* $\{\mathrm{CB}_k\}_{k\in[K]}$ *are valid in the sense of Equation 2 and the ascension functions are chosen as in Line 22 of Algorithm 1. Then, for any* $k \in [K]$*, the iterate* Q_k *satisfies* $\|Q_k\|_\infty \le Q_{\max}$ *with* $Q_{\max} = \frac{R_{\max}+2\omega/\alpha}{1-\gamma}$.

4.5 Exploration bonuses

The final technical step is to verify the validity of the exploration bonuses and to bound their cumulative size. The following lemma addresses the latter question.

Lemma 5 *Suppose Assumption 1 and event* $\mathcal{E}_L$ *hold and denote* $T = L_{\max}K$*. Then, with probability at least* $1-\delta$*, the policies* $\{\pi_k\}$ *and bonuses* $\{\mathrm{CB}_k\}$ *satisfy*

$$\sum_{k=1}^{K} \langle \mu(\pi_k), \mathrm{CB}_k \rangle \le 4(1-\gamma)\beta B \sqrt{dT \log\left(1 + \frac{B^2 T}{d}\right)} + 4\beta B \log\left(\frac{2K}{\delta}\right)^2,$$

and

$$\sum_{k=1}^{K} \langle \mu(\pi_k), \mathrm{CB}_k^2 \rangle \le 8(1-\gamma)\beta^2 B^2 d \log\left(1 + \frac{B^2 T}{d}\right) + 4\beta^2 B^2 \log\left(\frac{2K}{\delta}\right)^2,$$

For the proof, see Appendix B.5. Finally, it remains to show that the events $\mathcal{E}_{\text{valid}}$ and $\mathcal{E}_L$ hold which is done in the following lemma, whose proof can be found in Appendix B.6.

Lemma 6 *Let* $\beta = 8Q_{\max} d \log\left(c\alpha W_{\max} R_{\max} B^{9/2} Q_{\max}^4 L_{\max}^{5/2} K^{7/2} d^{5/2} \delta^{-1}\right)$*. Then, the event* $\mathcal{E}_{valid} \cap \mathcal{E}_L$ *holds with probability* $1-2\delta$.

4.6 Putting everything together

Theorem 1 then follows from applying Lemmas 1-6, using Equation (5), and bounding the total number of epochs (Lemma 9). The full details are provided in Appendix B.7.

5 Application: Imitation Learning from features alone

In this section, we show an application of the results presented in Section 4, making crucial use of the fact that our main result in Theorem 1 allows adaptively chosen reward functions.

5.1 Setting and motivation

We consider a linear MDP with an unknown reward function r_{true} defined in terms of the feature map $\varphi_r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{d_r}$ as $r_{\text{true}}(x,a) = \langle \varphi_r(x,a), w_{\text{true}} \rangle$ or with the operator notation $r_{\text{true}} = \Phi_r w_{\text{true}}$. The transition function is defined using the feature map $\varphi_P : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{d_P}$ via $P(x'|x,a) = \langle \varphi_P(x,a), m(x') \rangle$. It is easy to see that this is a linear MDP in terms of the concatenated feature map of dimension $d = d_r + d_P$. We consider a learning problem that we call *imitation learning from features alone*, where a learner receives as input a data set of feature vectors $\mathcal{D}_{\pi_E} = \{\varphi_r(X_E^i, A_E^i)\}_{i=1}^{\tau_E}$ where $X_E^i, A_E^i \sim \mu(\pi_E)$ generated by an *expert policy* π_E by interacting with the MDP $\mathcal{M}(r_{\text{true}})$. The learner is tasked with producing an ε-suboptimal policy π^{out} that satisfies $\mathbb{E}\left[\langle \nu_0, V_{r_{\text{true}}}^{\pi_E} - V_{r_{\text{true}}}^{\pi^{\text{out}}} \rangle\right] \le \varepsilon$. The learner has no knowledge of the reward function r_{true} apart from knowing a function class including it, but has access to the feature maps and can interact with the MDP.

This framework captures the important special case where the rewards depend only on the state but not on the action, by choosing a feature map φ_r that only depends on the states. In this case, the data set taken as input is significantly less informative than a full record of states and actions $\mathcal{D}_{\pi_E}^{\mathrm{x,a}} = \{(X_E^i, A_E^i)\}_{i=1}^{\tau_E}$. However, observing expert actions has been found restrictive in various practical scenarios, which gives a strong motivation to study the setting described above. For instance in robotics, a features only dataset describing a robotic manipulation task can be collected easily via cameras and sensors [Torabi et al., 2018, Zhu et al., 2020, Yang et al., 2019, Torabi et al., 2019]. On the other hand, collecting actions on top of features is more challenging as it requires knowledge of the

internal dynamics of the observed robots. Another example of imitation learning from features alone is learning to drive from a video which does not show the driver's actions but only the movements of the car induced by those actions. Finally, notice that imitation learning from states only, studied for example in Sun et al. [2019], is a particular case of our setting for $\varphi_r(x, a) = \mathbf{e}_x$. In this case, the expert dataset consists of states sampled from the expert state occupancy measure.

5.2 Algorithm and sample complexity guarantees

In the following, we propose a provably efficient algorithm for imitation learning from features alone. Our algorithm design and analysis is driven by the following decomposition of the regret, defined as $\mathfrak{R}_{\mathrm{K}}^{\mathrm{IL}} = \sum_{k=1}^{K} \left\langle \nu_0, V_{P,r_{\mathrm{true}}}^{\pi_E} - V_{P,r_{\mathrm{true}}}^{\pi_k} \right\rangle$, in terms of an appropriately chosen sequence of reward functions $r_1, \ldots, r_K$:

$$(1-\gamma)\,\mathfrak{R}_{\mathrm{K}}^{\mathrm{IL}} = \sum_{k=1}^{K} \langle r_k, \mu(\pi_E) - \mu(\pi_k) \rangle + \sum_{k=1}^{K} \langle \Phi_r^{\intercal} \mu(\pi_k) - \Phi_r^{\intercal} \mu(\pi_E), w_k - w_{\mathrm{true}} \rangle \,.$$

The first term in this decomposition corresponds to the regret of our online learning algorithm for adversarial MDPs, and can be controlled by invoking `Rmax-RAVI-UCB` on the sequence of rewards (r_k). The second term in the decomposition corresponds to the regret of another online learning algorithm picking a sequence of reward functions, aiming to minimize the sequence of linear loss functions $\Phi_r^{\intercal} \mu(\pi_k) - \Phi_r^{\intercal} \mu(\pi_E)$ (or at least do as well as the fixed comparator w_{true}). This objective can be achieved by running a standard online learning method such as projected online gradient descent (OGD, Zinkevich, 2003), using a sequence of unbiased loss estimates that can be computed efficiently using the observed feature vectors. The full algorithm is specified as Algorithm 2 in Appendix A and it is shown to satisfy the following guarantees.

Theorem 2 *Algorithm 2, when run for $K = \tilde{\mathcal{O}}\left(d^3 H^{9/2} \varepsilon^{-2} \log(|\mathcal{A}|)\right)$ iterations with an expert dataset of size $\tau_E = \tilde{\mathcal{O}}\left(W_{\max}^2 H^2 \varepsilon^{-2}\right)$, outputs an ε-suboptimal policy.*

Notably, the above is the first known bound for this setting that achieves a scaling ε^{-2} with the precision parameter for both the number of MDP interactions K and expert samples τ_E. We provide a detailed comparison with existing imitation learning theory works in Appendix C, whereas the complete technical details supporting the above theorem and a matching worst-case lower bound is provided in Appendix D.

5.3 Lower bounds

The upper bound of Theorem 2 depends both on the number of interaction steps K and the expert samples τ_E. It is natural to ask if these dependences can be improved in the setting we consider. We address both of these questions in the negative in a set of lower bounds described below.

Lower bound on the number of MDP interactions K. Theorem 5 in Appendix E proves that for any imitation learning from features alone algorithm, even in the setting $\tau_E = \infty$, there exists an MDP and an expert policy where at least $K = \Omega\left(\frac{dH^2}{\varepsilon^2}\right)$ interactions are needed to output a ε-optimal policy. This lower bound shows that the upper bound provided for K in Theorem 2 achieves the optimal scaling in ε and can be improved at most by a factor $d^2 H^{5/2}$. More importantly, this lower bound marks a clear separation between standard and features only imitation learning: purely offline learning ($K = 0$) is impossible in imitation learning from features alone while it is possible in standard imitation learning where the experts actions are observed (see, e.g., Foster et al., 2024).

In the construction of the lower bound, we consider a two state MDP with a reward function that depends only on the state variable and we set $\varphi_r(x, a) = \mathbf{e}_x$ which prevents observing expert actions. In this MDP, for $\tau_E = \infty$, the learner observes the expert state occupancy measure exactly and therefore, the "good" state that achieves the maximum of the expert state occupancy measure. However, since the learner does not know the MDP dynamics, interactions with the MDP are needed to find out the action that allows to maximize the learner state occupancy measure in the "good" state. Following standard techniques in MDP and bandits lower bound, we can ensure that the amount of MDP interactions is at least $\Omega\left(\varepsilon^{-2}\right)$.

Lower bound on the number of expert samples τ_E. Theorem 6 in Appendix E proves that for any algorithm for imitation learning from features alone, even for $K = \infty$, there exists an MDP and an expert policy where learning an ε-optimal policy requires at least $\tau_E = \Omega\left(W_{\max}^2 H^2 \varepsilon^{-2}\right)$ expert samples. This lower bound proves that Algorithm 2 scales optimally with all the problem parameters and in the precision ε. Moreover, this lower bound highlights again a clear distinction with standard imitation learning. On the one hand, standard imitation learning can be reduced to a supervised classification problem when the optimal policy is deterministic and the actions are observed in the dataset. As a consequence, the classic lower bound for supervised classification of order $\mathcal{O}\left(\varepsilon^{-1}\right)$ [Shalev-Shwartz and Ben-David, 2014] holds and it is matched by a purely offline behavioural cloning [Rajaraman et al., 2020]. On the other hand, in our lower bound construction we choose again $\varphi_r(x, a) = \mathbf{e}_x$ to make the expert actions unobservable for the learner and we can prove a larger lower bound of order $\mathcal{O}\left(\varepsilon^{-2}\right)$ which holds even if the expert policy is deterministic.

For the proof, our construction is again a two state MDP (a "good" state with high reward and a "bad" state with lower reward). The expert policy is chosen to be the optimal one. The transition dynamics and the initial distribution are chosen in a way that the expert state occupancy meausure is only marginally higher in the "good" state. That is, the expert occupancy measure equals roughly $1/2 + \varepsilon$ in the "good" state and $1/2 - \varepsilon$ in the "bad" state. By standard arguments, we then conclude that the learner needs at least $\Omega\left(\varepsilon^{-2}\right)$ samples from the expert to identify the "good" state.

6 Concluding remarks

We close by discussing a few open problems and potential improvements to our results.

On the objective function. We have focused on a relatively under-studied objective function: the discounted return from a fixed initial-state distribution. This is different from the objectives studied by other works such as Liu and Su [2020], He et al. [2021], Zhou et al. [2021], but arguably more natural when one is interested in learning algorithms that produce a single near-optimal policy at the end of an interactive learning period (which is the case in most practical applications one can think of). It is easy to see that resets to the initial state are absolutely necessary in this setting, unless one wants to make strong assumptions about the transition dynamics. A more exciting question is if our algorithm can be adapted to the significantly more challenging setting of undiscounted infinite-horizon reinforcement learning where existing methods [Wei et al., 2021, Hong et al., 2024, He et al., 2024] either obtain suboptimal regret bounds or leverage oracles whose computationally efficient implementation is unknown. So far, our attempts towards tackling this problem have remained unsuccessful. We believe that significant new ideas are necessary for solving this major open problem, but also that the techniques we introduce in this paper will be part of an eventual solution.

On second-degree optimism. Our key technical contribution draws inspiration from two sources: the `Rmax` algorithm of Brafman and Tennenholtz [2002], and the very recent work of Cassel and Rosenberg [2024]. While many of our technical tools are directly imported from the latter work, the concept of optimistically augmented MDPs and the connection with `Rmax` has arguably brought about a new level of understanding that can be potentially valuable for future work. It has certainly proved useful for our setting, where the notion of "contracted sub-MDP" used in the analysis of Cassel and Rosenberg [2024] cannot be meaningfully interpreted and used for analysis. We hope our work can bring some fresh attention to older (but apparently still powerful) ideas from the past of RL theory such as the `Rmax` trick.

On the tightness of the bounds. We find it very likely that our performance guarantees can be improved to some extent in terms of their dependence on H and d. In fact, we have made no attempt to optimize the scaling with respect to these parameters, and actually believe that the $H^{9/4}$ factor in the regret bound can be improved relatively easily. Specifically, we find it very likely that performing several value-function and policy updates at the end of each episode can reduce this factor—but we opted to keep the algorithm simple and the paper easy to read. We invite future researchers to verify this conjecture. Likewise, we believe that the dependence on the dimension d can be improved by using more sophisticated estimators and concentration inequalities (as done in the finite-horizon setting by He et al., 2023, Agarwal et al., 2023), but leave working out the (possibly non-trivial) details for another paper.

Acknowledgments

The authors wish to thank Asaf Cassel and Aviv Rosenberg for sharing further insights about their work at the virtual RL theory seminars, and Volkan Cevher for initial discussions about this project. This project has received funding from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 950180). This work is funded (in part) through a PhD fellowship of the Swiss Data Science Center, a joint venture between EPFL and ETH Zurich. Luca Viano acknowledges travel support from ELISE (GA no 951847).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Alekh Agarwal, Yujia Jin, and Tong Zhang. Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 987–1063. PMLR, 2023.

Sanjeev Arora, Simon Du, Sham Kakade, Yuping Luo, and Nikunj Saunshi. Provable representation learning for imitation learning via bi-level optimization. In *International Conference on Machine Learning*, pages 367–376. PMLR, 2020.

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

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning Journal*, 47(2-3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL http://dx.doi.org/10.1023/A:1013689704352.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pages 263–272. PMLR, 2017.

Ronen I Brafman and Moshe Tennenholtz. R-max — a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47:119–137, 1979.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning (ICML)*, 2020.

Asaf Cassel and Aviv Rosenberg. Warm-up free policy optimization: Improved regret in linear markov decision processes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=1c9XHlHTs7.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret, 2019. URL https://arxiv.org/abs/1902.06223.

Yan Dai, Haipeng Luo, Chen-Yu Wei, and Julian Zimmert. Refined regret for adversarial mdps with linear function approximation. *arXiv preprint arXiv:2301.12942*, 2023.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In Rocco A. Servedio and Tong Zhang, editors, *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 355–366. Omnipress, 2008.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

Dylan J Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. *arXiv preprint arXiv:2407.15007*, 2024.

Noah Golowich and Ankur Moitra. Linear bellman completeness suffices for efficient online reinforcement learning with few actions. *arXiv preprint arXiv:2406.11640*, 2024.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted MDPs. *Advances in Neural Information Processing Systems*, 34:22288–22300, 2021.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR, 2023.

Jianliang He, Han Zhong, and Zhuoran Yang. Sample-efficient learning of infinite-horizon average-reward mdps with general function approximation. *arXiv preprint arXiv:2404.12648*, 2024.

Kihyuk Hong, Yufan Zhang, and Ambuj Tewari. Provably efficient reinforcement learning for infinite-horizon average-reward linear mdps. *arXiv preprint arXiv:2405.15050*, 2024.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. 2012.

Pihe Hu, Yu Chen, and Longbo Huang. Towards minimax optimal reward-free reinforcement learning in linear mdps. In *The Eleventh International Conference on Learning Representations*, 2022.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, August 2010. ISSN 1532-4435.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation, 2019.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

Angeliki Kamoutsi, Goran Banjac, and John Lygeros. Efficient performance bounds for primal-dual reinforcement learning from demonstrations. In *International Conference on Machine Learning (ICML)*, 2021.

Rahul Kidambi, Jonathan Chang, and Wen Sun. Mobile: Model-based imitation learning from observation alone. *Advances in Neural Information Processing Systems*, 34:28598–28611, 2021.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Jonathan Wilder Lavington, Sharan Vaswani, and Mark Schmidt. Improved policy optimization for online imitation learning. In *Conference on Lifelong Learning Agents*, pages 1146–1173. PMLR, 2022.

Filippo Lazzati, Mirco Mutti, and Alberto Maria Metelli. How to scale inverse rl to large state spaces? a provably efficient approach. *arXiv preprint arXiv:2406.03812*, 2024.

Yichen Li and Chicheng Zhang. On efficient online imitation learning via classification. *Advances in Neural Information Processing Systems*, 35:32383–32397, 2022.

David Lindner, Andreas Krause, and Giorgia Ramponi. Active exploration for inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 35:5843–5853, 2022.

Haolin Liu, Chen-Yu Wei, and Julian Zimmert. Towards optimal regret in adversarial linear mdps with bandit feedback. *arXiv preprint arXiv:2310.11550*, 2023.

Shuang Liu and Hao Su. Regret bounds for discounted MDPs. *arXiv preprint arXiv:2002.05138*, 2020.

Zhihan Liu, Yufeng Zhang, Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Learning from demonstration: Provably efficient adversarial policy imitation with linear function approximation. In *International Conference on Machine Learning (ICML)*, 2022.

Zak Mhammedi, Adam Block, Dylan J Foster, and Alexander Rakhlin. Efficient model-free exploration in low-rank mdps. *Advances in Neural Information Processing Systems*, 36, 2024.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *Journal of Machine Learning Research*, 25(6):1–76, 2024.

Antoine Moulin and Gergely Neu. Optimistic planning by regularized dynamic programming. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25337–25357. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/moulin23a.html.

Gergely Neu and Julia Olkhovskaya. Online learning in MDPs with linear function approximation and bandit feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, April 1994.

Nived Rajaraman, Lin Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33:2914–2924, 2020.

Nived Rajaraman, Yanjun Han, Lin Yang, Jingbo Liu, Jiantao Jiao, and Kannan Ramchandran. On the value of interaction and function approximation in imitation learning. *Advances in Neural Information Processing Systems*, 34:1325–1336, 2021.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8210–8219. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/rosenberg20a.html.

S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Contextual bandits and imitation learning with preference-based active queries. *Advances in Neural Information Processing Systems*, 36, 2024a.

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Selective sampling and imitation learning via online regression. *Advances in Neural Information Processing Systems*, 36, 2024b.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Lior Shani, Tom Zahavy, and Shie Mannor. Online apprenticeship learning. *arXiv:2102.06924*, 2021.

Uri Sherman, Alon Cohen, Tomer Koren, and Yishay Mansour. Rate-optimal policy optimization for linear markov decision processes. *arXiv preprint arXiv:2308.14642*, 2023a.

Uri Sherman, Tomer Koren, and Yishay Mansour. Improved regret for efficient online reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 31117–31150. PMLR, 2023b.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Wen Sun, Arun Venkatraman, Geoffrey J Gordon, Byron Boots, and J Andrew Bagnell. Deeply aggrevated: Differentiable imitation learning for sequential prediction. In *International conference on machine learning*, pages 3309–3318. PMLR, 2017.

Wen Sun, Anirudh Vemula, Byron Boots, and Drew Bagnell. Provably efficient imitation learning from observation alone. In *International conference on machine learning*, pages 6036–6045. PMLR, 2019.

Gokul Swamy, Sanjiban Choudhury, J Andrew Bagnell, and Steven Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pages 10022–10032. PMLR, 2021.

Gokul Swamy, Nived Rajaraman, Matt Peng, Sanjiban Choudhury, J Bagnell, Steven Z Wu, Jiantao Jiao, and Kannan Ramchandran. Minimax optimal online imitation learning via replay estimation. *Advances in Neural Information Processing Systems*, 35:7077–7088, 2022.

István Szita and Csaba Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1031–1038, 2010.

Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018.

Faraz Torabi, Garrett Warnell, and Peter Stone. Recent advances in imitation learning from observation. *arXiv preprint arXiv:1905.13566*, 2019.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

Luca Viano, Angeliki Kamoutsi, Gergely Neu, Igor Krawczuk, and Volkan Cevher. Proximal point imitation learning. *Advances in Neural Information Processing Systems*, 35:24309–24326, 2022.

Luca Viano, Stratis Skoulakis, and Volkan Cevher. Imitation learning in discounted linear MDPs without exploration assumptions. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=DChQpB4AJy.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, pages 22430–22456. PMLR, 2022.

Ruosong Wang, Simon S Du, Lin Yang, and Russ R Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *Advances in Neural Information Processing Systems*, 34:13524–13536, 2021.

Chen-Yu Wei, Mehdi Jafarnia-Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward MDPs with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015, 2021.

Gellért Weisz, András György, and Csaba Szepesvári. Online rl in linearly q^π-realizable mdps is as easy as in linear mdps if you learn what to ignore. *Advances in Neural Information Processing Systems*, 36, 2024.

Feiyang Wu, Jingyang Ke, and Anqi Wu. Inverse reinforcement learning with the average reward criterion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=YFSrf8aciU.

Runzhe Wu, Yiding Chen, Gokul Swamy, Kianté Brantley, and Wen Sun. Diffusing states and matching scores: A new framework for imitation learning. *arXiv preprint arXiv:2410.13855*, 2024a.

Runzhe Wu, Ayush Sekhari, Akshay Krishnamurthy, and Wen Sun. Computationally efficient rl under linear bellman completeness for deterministic dynamics. *arXiv preprint arXiv:2406.11810*, 2024b.

Tian Xu, Ziniu Li, Yang Yu, and Zhi-Quan Luo. Understanding adversarial imitation learning in small sample regime: A stage-coupled analysis. *arXiv preprint arXiv:2208.01899*, 2022.

Tian Xu, Ziniu Li, Yang Yu, and Zhi-Quan Luo. Provably efficient adversarial imitation learning with unknown transitions. *arXiv preprint arXiv:2306.06563*, 2023.

Chao Yang, Xiaojian Ma, Wenbing Huang, Fuchun Sun, Huaping Liu, Junzhou Huang, and Chuang Gan. Imitation learning from observations by minimizing inverse dynamics disagreement. *Advances in neural information processing systems*, 32, 2019.

Lin Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning (ICML)*, 2019.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

Siliang Zeng, Mingyi Hong, and Alfredo Garcia. Structural estimation of markov decision processes in high-dimensional state space with finite-time guarantees. *arXiv preprint arXiv:2210.01282*, 2022a.

Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Maximum-likelihood inverse reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 35:10122–10135, 2022b.

Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Understanding expertise through demonstrations: A maximum likelihood framework for offline inverse reinforcement learning. *arXiv preprint arXiv:2302.07457*, 2023.

Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient reinforcement learning in block mdps: A model-free representation learning approach. In *International Conference on Machine Learning*, pages 26517–26547. PMLR, 2022.

Han Zhong and Tong Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *Advances in Neural Information Processing Systems*, 36, 2024.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted MDPs with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021.

Zhuangdi Zhu, Kaixiang Lin, Bo Dai, and Jiayu Zhou. Off-policy imitation learning from observations. *Advances in neural information processing systems*, 33:12402–12413, 2020.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In Tom Fawcett and Nina Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, pages 928–936. AAAI Press, 2003.

Contents of Appendix

A Omitted pseudocodes

This section includes the pseudocode for `Rmax-RAVI-UCB`. Each of the steps is explained in details in Section 3.

Algorithm 1 `Rmax-RAVI-UCB` for Linear MDPs.

Inputs: Number of resets K, learning rate $\eta > 0$, exploration coefficient $\beta > 0$, threshold $\omega > 0$, slope sigmoid $\alpha > 0$.
Initialize: $X_1 \sim \nu_0$, $\pi_1 = \pi_{\texttt{unif}}$, $Q_1 = 0$, $\mathcal{D}_1 = \emptyset$, $\Lambda_1 = I$, $t = 1$, $e = 0$.
for $k = 1, \ldots, K$ **do**
  `# interact with the environment`
  The adversary adaptively chooses r_k, i.e. $r_k = \textsc{RewardUpdate}\left(\{\pi_\ell\}_{\ell=1}^{k}, \{r_\ell\}_{\ell=1}^{k-1}\right)$.
  while true do
    Observe the state X_t and play an action $A_t \sim \pi_k(\cdot \mid X_t)$.
    Receive reward $r_k(X_t, A_t)$ and observe the function r_k.
    With probability $1 - \gamma$, reset to initial distribution: $X_{t+1} \sim \nu_0$ set $T_k = t$ and **break** .
    Otherwise observe the next state $X_{t+1} \sim P(\cdot \mid X_t, A_t)$.
    Update $\Lambda_{t+1} = \Lambda_t + \varphi(X_t, A_t)\,\varphi(X_t, A_t)^{\mathsf{T}}$.
    $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(X_t, A_t, X_{t+1})\}$.
    $t = t + 1$.
  end while
  `# initialize new epoch`
  if $t = T_1$ **or** $\det \Lambda_{T_k} \geq 2 \det \Lambda_{t_e}$ **then**
    $e = e + 1$.
    Set $k_e = k$ and $t_e = t$.
    Reset the policy $\pi_k = \pi_{\texttt{unif}}$.
  end if
  For any $(x, a) \in \mathcal{X} \times \mathcal{A}$, $\mathrm{CB}_k(x, a) = \beta \|\varphi(x, a)\|_{\Lambda_{t_e}^{-1}}$, and $\mathrm{CB}_k(x^+, a) = 0$.
  For any $(x, a) \in \mathcal{X} \times \mathcal{A}$, $p_k^+(x, a) = \sigma(\alpha\,\mathrm{CB}_k(x, a) - \omega)$, and $p_k^+(x^+, a) = 0$.
  `# optimistic regularized value iteration`
  $r_k^+ = (1 - p_k^+) \odot r_k + p_k^+ \cdot R_{\max}$.
  $\widehat{MV_k} = \Lambda_{T_k}^{-1} \sum_{(x,a,x') \in \mathcal{D}_{T_k}} \varphi(x, a)\, V_k(x')$.
  $\widehat{P_k^+ V_k} = (1 - p_k^+) \odot \Phi \widehat{MV_k} + p_k^+ \cdot V_k(x^+)$, and $\widehat{P_k^+ V_k}(x^+, \cdot) = \frac{R_{\max}}{1-\gamma}$.
  $Q_{k+1} = r_k^+ + (1 - p_k^+) \odot \mathrm{CB}_k + \gamma \widehat{P_k^+ V_k}$.
  $V_{k+1}(x) = \frac{1}{\eta} \log\left(\sum_a \pi_k(a \mid x)\, e^{\eta Q_{k+1}(x,a)}\right)$.
  $\pi_{k+1} = \pi_k \odot e^{\eta(Q_{k+1} - EV_{k+1})}$.
end for
Output: π_I, with $I \sim \mathcal{U}([K])$.

Next, we include the pseudocode for our imitation learning algorithms built on `Rmax-RAVI-UCB`. At line 4, the learner computes an estimate of the expert features expectation computing an elementwise empirical average of the features in the dataset $\mathcal{D}_{\pi_E}$. Such an estimate is leveraged in the online gradient descent (OGD) update given by the function at lines 5-7. This function instantiates the general RewardUpdate routine given in Algorithm 1. That is, after each policy update in `Rmax-RAVI-UCB`, the reward player estimates the feature expectation of the current policy π_k as the plug in estimator $\varphi_r(X_k, A_k)$ with X_k, A_k sampled from the occupancy measure $\mu(\pi_k)$. Notice that for the reinforcement learning applications, the adversarial reward sequence is generated online observing the policies. Therefore, for this application it is important that the guarantees in Theorem 1 holds against adaptive adversaries.

Algorithm 2 `FRA-IL` (Feature Rmax Adversarial Imitation Learning)

Inputs:
(1) a features dataset $\mathcal{D}_{\pi_E} = \left\{\varphi_r\left(X_E^i, A_E^i\right)\right\}_{i=1}^{\tau_E}$ where for any $i \in [\tau_E]$, $X_E^i, A_E^i \sim \mu\left(\pi_E\right)$,
(2) read access to $\varphi_P\left(x, a\right)$ for all $x, a \in \mathcal{X} \times \mathcal{A}$,
(3) trajectory access to $\mathcal{M} \setminus r_{\text{true}}$, and
(4) the reward weights class $\mathcal{W}$ such that $w_{\text{true}} \in \mathcal{W}$ and $\|w\| \leq W_{\max}$ for all $w \in \mathcal{W}$.
Set $K, \eta, \beta, \omega, \alpha$ as in Theorem 1.
Set $\eta_r = {}^{W_{\max}}/_{B\sqrt{K}}$.
Estimate $\widehat{\lambda\left(\pi_E\right)} = \frac{1}{\left|\mathcal{D}_{\pi_E}\right|} \sum_{i=1}^{\tau_E} \varphi_r\left(X_E^i, A_E^i\right)$.
Function $\textsc{Ogd}(\mu\left(\pi_k\right), w_{k-1})$
Sample $X_k, A_k \sim \mu\left(\pi_k\right)$.
    return $w_k = \Pi_{\mathcal{W}}\left(w_{k-1} + \eta_r\left(\widehat{\lambda\left(\pi_E\right)} - \varphi_r\left(X_k, A_k\right)\right)\right)$.
Output: `Rmax-RAVI-UCB` $\left(K, \eta, \beta, \omega, \alpha, \textsc{RewardUpdate} = \textsc{Ogd}\right)$.

B Omitted proofs from Section 4

B.1 Proof of Lemma 1 (reward bias)

Lemma 1 *For any choice of p_k^+, we have $\texttt{reward-bias}_k \leq R_{\max} \langle \mu(\pi_k), p_k^+ \rangle$.*

Proof 1 *First note that for any action a, the rewards are equal, i.e. $r_k(x^+, a) = r_k^+(x^+, a)$. For the other states, plugging the definition of r_k^+ gives*

$$\begin{aligned}
\texttt{reward-bias}_k &= \langle \mu(\pi^\star) - \mu(\pi_k), r_k - r_k^+ \rangle \\
&= \langle \mu(\pi^\star) - \mu(\pi_k), p_k^+ \odot (r_k - R_{\max}\mathbf{1}) \rangle \\
&\leq -\langle \mu(\pi_k), p_k^+ \odot (r_k - R_{\max}\mathbf{1}) \rangle \\
&\leq R_{\max} \langle \mu(\pi_k), p_k^+ \rangle ,
\end{aligned}$$

where the first inequality follows from $r_k - R_{\max}\mathbf{1} \preceq 0$ and $\mu(\pi^\star) \succeq 0$, and the second inequality is due to $r_k \succeq 0$.

B.2 Proof of Lemma 2 (model bias)

Lemma 2 *Let π be any policy and p_k^+ be any ascension function at episode k. Then,*

$$0 \leq \texttt{model-bias}_k(\pi) \leq \frac{R_{\max}}{1-\gamma} \langle \mu(\pi), p_k^+ \rangle .$$

Proof 2 *Let us consider a process $(X_\tau, A_\tau)_{\tau \in \mathbb{N}}$ generated by the policy π in the real MDP, i.e., such that $X_0 \sim \nu_0$, and for any $\tau \in \mathbb{N}$, $A_\tau \sim \pi(\cdot \mid X_\tau)$, and $X_{\tau+1} \sim P(\cdot \mid X_\tau, A_\tau)$. We denote $\left(X_{k,\tau}^+, A_{k,\tau}^+\right)_{\tau \in \mathbb{N}}$ its coupled process in the optimistic MDP at episode k generated as follows. At the first stage we set $X_{k,0}^+ = X_0$, then for any $\tau \geq 1$, the coupled process evolves as follows*

$$X_{k,\tau+1}^+, A_{k,\tau+1}^+ = \begin{cases} X_{\tau+1}, A_{\tau+1} & \text{w.p.} \quad 1 - p_k^+(X_\tau, A_\tau) \quad \text{if} \quad X_{k,\tau}^+, A_{k,\tau}^+ = X_\tau, A_\tau \\ x^+, a & \text{w.p.} \quad p_k^+(X_\tau, A_\tau) \quad \text{if} \quad X_{k,\tau}^+, A_{k,\tau}^+ = X_\tau, A_\tau \\ x^+, a & \text{if} \quad X_{k,\tau}^+, A_{k,\tau}^+ \neq X_\tau, A_\tau \end{cases} ,$$

where a can be any action. Then, we can rewrite the bias term as

$$\begin{aligned}
\texttt{model-bias}_k(\pi) &= (1-\gamma)\, \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \left(r_k^+\left(X_{k,\tau}^+, A_{k,\tau}^+\right) - r_k^+(X_\tau, A_\tau) \right) \right] \\
&= (1-\gamma)\, \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{I}_{\left\{ \left(X_{k,\tau}^+, A_{k,\tau}^+\right) \neq (X_\tau, A_\tau) \right\}} \left(r_k^+\left(X_{k,\tau}^+, A_{k,\tau}^+\right) - r_k^+(X_\tau, A_\tau) \right) \right] .
\end{aligned}$$

By definition, the state-action pairs $\left(X_{k,\tau}^+, A_{k,\tau}^+\right)$ and (X_τ, A_τ) differ when the coupled process goes to heaven, i.e. $X_{k,\tau}^+ = x^+$. Noting that $r_k(x^+, a) = R_{\max}$ for any action $a \in \mathcal{A}$, we further get

$$\begin{aligned}
\texttt{model-bias}_k(\pi) &= (1-\gamma)\, \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{I}_{\left\{ \left(X_{k,\tau}^+, A_{k,\tau}^+\right) \neq (X_\tau, A_\tau) \right\}} \left(r_k^+\left(x^+, A_{k,\tau}^+\right) - r_k^+(X_\tau, A_\tau) \right) \right] \\
&= (1-\gamma)\, \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{I}_{\left\{ \left(X_{k,\tau}^+, A_{k,\tau}^+\right) \neq (X_\tau, A_\tau) \right\}} \left(R_{\max} - r_k^+(X_\tau, A_\tau) \right) \right] ,
\end{aligned}$$

and $\texttt{model-bias}_k(\pi) \geq 0$ follows from $r_k^+ \preceq R_{\max}$. For the upper bound, we can instead use $r_k \succeq 0$ and continue as follows

$$\begin{aligned}
\texttt{model-bias}_k(\pi) &\leq (1-\gamma)\, R_{\max} \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{I}_{\left\{ \left(X_{k,\tau}^+, A_{k,\tau}^+\right) \neq (X_\tau, A_\tau) \right\}} \right] \\
&= (1-\gamma)\, \gamma R_{\max} \sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{P}\left[\left(X_{k,\tau+1}^+, A_{k,\tau+1}^+\right) \neq (X_{\tau+1}, A_{\tau+1}) \right] ,
\end{aligned}$$

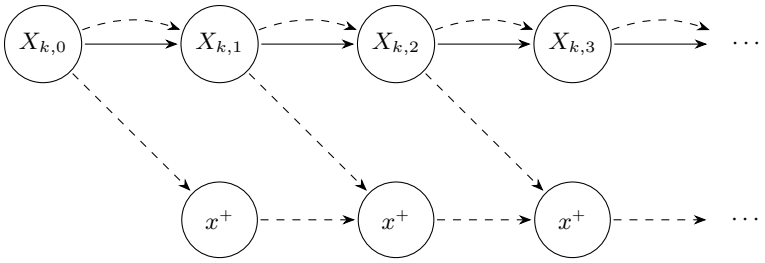

Figure 2: The thick arrows represent the transitions of the process in the original MDP, while the dashed ones correspond to the utopian one.

where we used $\mathbb{P}\left[\left(X_{k,0}^+, A_{k,0}^+\right) \neq (X_0, A_0)\right] = 0$ *by definition. Then, as illustrated in Figure 2, two cases can happen. Either the coupled process was still in the original MDP and transitioned to heaven, either it was already in heaven. Denoting for any* $\tau \geq 0$ *the event* $\mathcal{E}_{\text{split}}(\tau) = \left\{\left(X_{k,\tau}^+, A_{k,\tau}^+\right) \neq (X_\tau, A_\tau)\right\}$ *and* $\mathcal{E}_{\text{split}}^c(\tau)$ *its complementary, we have*

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}_{\text{split}}(\tau+1)\right] &= \mathbb{P}\left[\mathcal{E}_{\text{split}}(\tau+1) \mid \mathcal{E}_{\text{split}}(\tau)\right]\mathbb{P}\left[\mathcal{E}_{\text{split}}(\tau)\right] \\ &\quad + \mathbb{P}\left[\mathcal{E}_{\text{split}}(\tau+1) \,\middle|\, \mathcal{E}_{\text{split}}^c(\tau)\right]\mathbb{P}\left[\mathcal{E}_{\text{split}}^c(\tau)\right] .\end{aligned}$$

If the coupled process is already in the heaven state x^+*, then it stays there. Otherwise, it can transition there with probability* $\mathbb{E}\left[p_k^+(X_\tau, A_\tau)\right]$*, thus*

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}_{\text{split}}(\tau+1)\right] &= \mathbb{P}\left[\mathcal{E}_{\text{split}}(\tau)\right] + \mathbb{E}\left[p_k^+(X_\tau, A_\tau)\right]\mathbb{P}\left[\mathcal{E}_{\text{split}}^c(\tau)\right] \\ &\leq \mathbb{P}\left[\mathcal{E}_{\text{split}}(\tau)\right] + \mathbb{E}\left[p_k^+(X_\tau, A_\tau)\right] \\ &\leq \sum_{u=0}^{\tau}\mathbb{E}\left[p_k^+(X_u, A_u)\right] ,\end{aligned}$$

by induction. Therefore, we get

$$\begin{aligned}\texttt{model-bias}_k(\pi) &\leq (1-\gamma)\gamma R_{\max}\sum_{\tau=0}^{\infty}\gamma^\tau\sum_{u=0}^{\tau}\mathbb{E}\left[p_k^+(X_u, A_u)\right] \\ &= (1-\gamma)\gamma R_{\max}\mathbb{E}\left[\sum_{u=0}^{\infty}\sum_{\tau=0}^{\infty}\gamma^\tau\mathbb{I}_{\{u\leq\tau\}}p_k^+(X_u, A_u)\right] \\ &= (1-\gamma)\gamma R_{\max}\mathbb{E}\left[\sum_{u=0}^{\infty}\sum_{\tau=u}^{\infty}\gamma^\tau p_k^+(X_u, A_u)\right] \\ &= \gamma R_{\max}\mathbb{E}\left[\sum_{u=0}^{\infty}\gamma^u p_k^+(X_u, A_u)\right] ,\end{aligned}$$

and the conclusion follows from the definition of $\mu(\pi)$ *and* $\gamma < 1$*.*

B.2.1 Alternative Proof of Lemma 2

We also provide an alternative proof based on a simulation lemma.

Proof 3 *By the flow constraints associated to* $\mu_k^+(\pi)$ *and after rearranging, we get*

$$\begin{aligned}\texttt{model-bias}_k(\pi) &= (1-\gamma)\left\langle \nu_0, V_{P_k^+, r_k^+}^\pi - V_{P, r_k^+}^\pi\right\rangle \\ &= \left\langle E^\intercal\mu_k^+(\pi) - \gamma\left(P_k^+\right)^\intercal\mu_k^+(\pi), V_{P_k^+, r_k^+}^\pi - V_{P, r_k^+}^\pi\right\rangle \\ &= \left\langle \mu_k^+(\pi), EV_{P_k^+, r_k^+}^\pi - EV_{P, r_k^+}^\pi - \gamma P_k^+ V_{P_k^+, r_k^+}^\pi + \gamma P_k^+ V_{P, r_k^+}^\pi\right\rangle .\end{aligned}$$

Applying Lemma 6 to both $V^{\pi}_{P_k^+,r_k^+}$ and V^{π}_{P,r_k^+} and Bellman's equation for $Q^{\pi}_{P_k^+,r_k^+}$, we further have

$$\begin{aligned}\texttt{model-bias}_k(\pi) &= \left\langle \mu_k^+(\pi), Q^{\pi}_{P_k^+,r_k^+} - Q^{\pi}_{P,r_k^+} - \gamma P_k^+ V^{\pi}_{P_k^+,r_k^+} + \gamma P_k^+ V^{\pi}_{P,r_k^+} \right\rangle \\ &= \left\langle \mu_k^+(\pi), r_k^+ + \gamma P_k^+ V^{\pi}_{P,r_k^+} - Q^{\pi}_{P,r_k^+} \right\rangle ,\end{aligned}$$

Plugging the definition of P_k^+ and this time using Bellman's equation for Q^{π}_{P,r_k^+}, we obtain

$$\begin{aligned}\texttt{model-bias}_k(\pi) &= \left\langle \mu_k^+(\pi), r_k^+ + \gamma(1-p_k^+) \odot PV^{\pi}_{P,r_k^+} + p_k^+ \odot \mathbf{e}_{x^+} V^{\pi}_{P,r_k^+} - Q^{\pi}_{P,r_k^+} \right\rangle \\ &= \left\langle \mu_k^+(\pi), p_k^+ \odot \left(\mathbf{e}_{x^+} V^{\pi}_{P,r_k^+} - \gamma PV^{\pi}_{P,r_k^+} \right) \right\rangle \\ &= \left\langle \mu_k^+(\pi), p_k^+ \odot \left(\frac{R_{\max}}{1-\gamma}\mathbf{1} - \gamma PV^{\pi}_{P,r_k^+} \right) \right\rangle , \end{aligned} \tag{6}$$

where the last equality is due to having $\left(\mathbf{e}_{x^+} V^{\pi}_{P,r_k^+}\right)(x,a) = V^{\pi}_{P,r_k^+}(x^+) = \frac{R_{\max}}{1-\gamma}$ for any state-action pair (x,a). The lower bound follows from noticing that $PV^{\pi}_{P,r_k^+} \preceq \frac{1}{1-\gamma}\mathbf{1} \preceq \frac{R_{\max}}{1-\gamma}\mathbf{1}$,

$$\begin{aligned}\texttt{model-bias}_k(\pi) &\geq \left(\frac{R_{\max}}{1-\gamma} - \frac{\gamma R_{\max}}{1-\gamma} \right) \cdot \langle \mu_k^+(\pi), p_k^+ \rangle \\ &= R_{\max} \cdot \langle \mu_k^+(\pi), p_k^+ \rangle \\ &\geq 0 .\end{aligned}$$

Moving to the upper bound, from Equation 6 and $PV^{\pi}_{P,r_k^+} \succeq 0$, we get

$$\begin{aligned}\texttt{model-bias}_k(\pi) &= \left\langle \mu_k^+(\pi), p_k^+ \odot \left(\frac{R_{\max}}{1-\gamma}\mathbf{1} - \gamma PV^{\pi}_{P,r_k^+} \right) \right\rangle \\ &\leq \frac{R_{\max}}{1-\gamma} \langle \mu_k^+(\pi), p_k^+ \rangle \\ &\leq \frac{R_{\max}}{1-\gamma} \langle \mu(\pi), p_k^+ \rangle ,\end{aligned}$$

where the last inequality follows from Lemma 8.

B.3 Proof of Lemma 3 (optimistic regret)

In order to prove Lemma 3, we first need the following result that shows that the functions Q_k are optimistic estimates of an ideal sequence of dynamic-programming updates computed in the augmented MDPs. The statement is is an adaptation of Lemma 4.2 of Moulin and Neu [2023], and its complete proof is provided below.

Lemma 7 *Suppose that the bonuses CB_k are valid for the MDP $\mathcal{M}_k$ in the sense of Equation 2. Then, for any k and any state-action pair $(x,a) \in \mathcal{X}^+ \times \mathcal{A}$, the iterates Q_k satisfy*

$$r_k^+ + \gamma P_k^+ V_k \leq Q_{k+1} \leq r_k^+ + 2(1-p_k^+) \odot \mathrm{CB}_k + \gamma P_k^+ V_k .$$

Proof 4 *For x^+ and any action a, it is straightforward to check that both inequalities are equalities. Let $(x,a) \in \mathcal{X} \times \mathcal{A}$. We have*

$$\begin{aligned} r_k^+(x,a) + \gamma P_k^+ V_k(x,a) &= r_k^+(x,a) + \gamma \left(P_k^+ - \widehat{P}_k^+ \right) V_k(x,a) + \gamma \widehat{P}_k^+ V_k(x,a) \\ &\leq Q_{k+1}(x,a) \\ &\leq r_k^+(x,a) + 2(1-p_k^+(x,a))\,\mathrm{CB}_k(x,a) + \gamma P_k^+ V_k(x,a) , \end{aligned}$$

where both inequalities use the fact that

$$\left| \left(P_k^+ - \widehat{P}_k^+ \right) V_k(x,a) \right| \leq (1-p_k^+(x,a))\,\mathrm{CB}_k(x,a) ,$$

which is implied by the event $\mathcal{E}_{\mathrm{valid}}$ in Equation 2.

Lemma 3 *Suppose that the bonuses $\{\mathrm{CB}_k\}$ are valid in the sense of Equation 2 and that for any k, $\|Q_k\|_\infty \leq Q_{\max}$. Then, the sequence of policies output by Algorithm 1 satisfies*

$$\mathfrak{R}_{\mathrm{K}}^{+} \leq \frac{E\left(K\right)\log|\mathcal{A}|}{\eta} + 4Q_{\max}E\left(K\right) + \frac{2Q_{\max}^2\eta K}{\sqrt{1-\gamma}} + 2\sum_{k=1}^{K}\left\langle \mu\left(\pi_k\right), \left(1-p_k^+\right)\odot \mathrm{CB}_k\right\rangle . \quad (4)$$

Proof 5 *We decompose $\mathfrak{R}_{\mathrm{K}}^{+}$ as follows*

$$\mathfrak{R}_{\mathrm{K}}^{+} = \sum_{k=1}^{K}\Big(\underbrace{\left(\left\langle \mu_k^+\left(\pi^\star\right), r_k^+\right\rangle - \left(1-\gamma\right)\left\langle \nu_0, V_k\right\rangle\right)}_{=\Delta_k^\star} + \underbrace{\left(1-\gamma\right)\left\langle \nu_0, V_k\right\rangle - \left\langle \mu_k^+\left(\pi_k\right), r_k^+\right\rangle}_{=\Delta_k}\Big),$$

where we defined $\Delta_k^\star$ and Δ_k. We start with the first term. Using the flow constraint with $\mu_k^+\left(\pi^\star\right)$,

$$\begin{aligned}\Delta_k^\star &= \left\langle \mu_k^+\left(\pi^\star\right), r_k^+\right\rangle - \left\langle E^\intercal\mu_k^+\left(\pi^\star\right) - \gamma\left(P_k^+\right)^\intercal\mu_k^+\left(\pi^\star\right), V_k\right\rangle \\ &= \left\langle \mu_k^+\left(\pi^\star\right), r_k^+ + \gamma P_k^+V_k - EV_{k+1}\right\rangle + \left\langle \mu_k^+\left(\pi^\star\right), EV_{k+1} - EV_k\right\rangle .\end{aligned}$$

Using the lower bound on Q_{k+1} from Lemma 7, we have

$$\Delta_k^\star \leq \left\langle \mu_k^+\left(\pi^\star\right), Q_{k+1} - EV_{k+1}\right\rangle + \left\langle \mu_k^+\left(\pi^\star\right), EV_{k+1} - EV_k\right\rangle ,$$

where the term in $x = x^+$ is equal to zero. Summing over $k \in [K] = \bigcup_{e\in[1,E(K)]}[k_e, k_{e+1}-1]$,

$$\sum_{k=1}^{K}\Delta_k^\star \leq \sum_{e=1}^{E(K)}\left\langle \mu_{k_e}^+\left(\pi^\star\right), \sum_{k\in\mathcal{K}_e}\left(Q_{k+1} - EV_{k+1}\right)\right\rangle + \sum_{e=1}^{E(K)}\left\langle E^\intercal\mu_{k_e}^+\left(\pi^\star\right), V_{k_{e+1}} - V_{k_e}\right\rangle_{\mathcal{X}} ,$$

where the sum within each epoch of the second term telescoped. By Moulin and Neu, 2023, Lemma C.1, we have for any state $x \in \mathcal{X}$

$$\begin{aligned}\sum_{k\in\mathcal{K}_e} V_{k+1}\left(x\right) &= \max_{p\in\Delta(\mathcal{A})}\left\langle p, \sum_{k\in\mathcal{K}_e} Q_{k+1}\left(x,\cdot\right)\right\rangle - \frac{1}{\eta}\mathcal{D}_{KL}\left(p\|\pi_{\boldsymbol{unif}}\right) \\ &\geq \left\langle \pi^\star\left(\cdot\mid x\right), \sum_{k\in\mathcal{K}_e} Q_{k+1}\left(x,\cdot\right)\right\rangle - \frac{1}{\eta}\mathcal{D}_{KL}\left(\pi^\star\left(\cdot\mid x\right)\|\pi_{\boldsymbol{unif}}\right) ,\end{aligned}$$

where we used $\pi_{k_e} = \pi_{\boldsymbol{unif}}$ in the first equality and denoted $\mathcal{K}_e$ the set of episodes in epoch e. Multiplying the previous inequality by $\nu_{k_e}^+\left(\pi^\star, x\right)$, summing over $x \in \mathcal{X}$, and noting that $\mu_{k_e}^+\left(\pi^\star\right) = \nu_{k_e}^+\left(\pi^\star\right)\odot\pi^\star$, we obtain

$$\begin{aligned}\sum_{e=1}^{E(K)}\left\langle \mu_{k_e}^+\left(\pi^\star\right), \sum_{k\in\mathcal{K}_e}\left(Q_{k+1} - EV_{k+1}\right)\right\rangle &\leq \frac{1}{\eta}\sum_{e=1}^{E(K)}\left\langle \nu_{k_e}^+\left(\pi^\star\right), \mathcal{D}_{KL}\left(\pi^\star\|\pi_{\boldsymbol{unif}}\right)\right\rangle \\ &\leq \frac{E\left(K\right)\log|\mathcal{A}|}{\eta} .\end{aligned}$$

The second term can be bounded with Hölder's inequality,

$$\sum_{e=1}^{E(K)}\left\langle E^\intercal\mu_{k_e}^+\left(\pi^\star\right), V_{k_{e+1}} - V_{k_e}\right\rangle \leq \sum_{e=1}^{E(K)}\left\|\nu_{k_e}^+\left(\pi^\star\right)\right\|_1\left\|V_{k_{e+1}} - V_{k_e}\right\|_\infty \leq 2E\left(K\right)Q_{\max} ,$$

where we used $\nu_{k_e}^+\left(\pi^\star\right) \in \Delta\left(\mathcal{X}^+\right)$ and $\|V_k\|_\infty \leq Q_{\max}$ which follows from $\|Q_k\|_\infty \leq Q_{\max}$. Therefore, we get

$$\sum_{k=1}^{K}\Delta_k^\star \leq \frac{E\left(K\right)\log|\mathcal{A}|}{\eta} + 2E\left(K\right)Q_{\max} .$$

Moving to Δ_k, we apply the flow constraints to $\mu_k^+\left(\pi_k\right)$ to get

$$\begin{aligned}\Delta_k &= \left\langle E^\intercal\mu_k^+\left(\pi_k\right) - \gamma\left(P_k^+\right)^\intercal\mu_k^+\left(\pi_k\right), V_k\right\rangle - \left\langle \mu_k^+\left(\pi_k\right), r_k^+\right\rangle \\ &= \left\langle \mu_k^+\left(\pi_k\right), EV_k\right\rangle - \left\langle \mu_k^+\left(\pi_k\right), r_k^+ + \gamma P_k^+V_k\right\rangle \\ &\leq \left\langle \mu_k^+\left(\pi_k\right), EV_k - Q_{k+1}\right\rangle + 2\left\langle \mu_k^+\left(\pi_k\right), \left(1-p_k^+\right)\odot\mathrm{CB}_k\right\rangle \\ &\leq \left\langle \mu_k^+\left(\pi_k\right), EV_k - Q_{k+1}\right\rangle + 2\left\langle \mu\left(\pi_k\right), \left(1-p_k^+\right)\odot\mathrm{CB}_k\right\rangle ,\end{aligned}$$

where the first inequality follows from the upper bound on Q_{k+1} in Lemma 7 and the term in $x = x^+$ being equal to zero, and the second inequality is due to Lemma 8. Next, noticing $\langle \mu_k^+(\pi_{k+1}), Q_{k+1}\rangle = \left\langle \nu_k^+(\pi_{k+1}), V_{k+1} + \frac{1}{\eta}\mathcal{D}_{KL}(\pi_{k+1}\|\pi_k)\right\rangle$,

$$\begin{aligned}\langle \mu_k^+(\pi_k), EV_k - Q_{k+1}\rangle &= \langle \nu_k^+(\pi_k), V_k\rangle - \langle \mu_k^+(\pi_{k+1}), Q_{k+1}\rangle \\ &\quad + \langle \mu_k^+(\pi_{k+1}), Q_{k+1}\rangle - \langle \mu_k^+(\pi_k), Q_{k+1}\rangle \\ &= \langle \nu_k^+(\pi_k), V_k\rangle - \langle \nu_k^+(\pi_{k+1}), V_{k+1}\rangle \\ &\quad - \frac{1}{\eta}\langle \nu_k^+(\pi_{k+1}), \mathcal{D}_{KL}(\pi_{k+1}\|\pi_k)\rangle \\ &\quad + \langle \mu_k^+(\pi_{k+1}) - \mu_k^+(\pi_k), Q_{k+1}\rangle\ .\end{aligned}$$

We sum over k and look at the different terms separately. First, we get

$$\begin{aligned}\sum_{k=1}^{K}\left(\langle \nu_k^+(\pi_k), V_k\rangle - \langle \nu_k^+(\pi_{k+1}), V_{k+1}\rangle\right) &= \sum_{e=1}^{E(K)}\left(\left\langle \nu_{k_e}^+(\pi_{k_e}), V_{k_e}\right\rangle - \left\langle \nu_{k_e}^+\left(\pi_{k_{e+1}}\right), V_{k_{e+1}}\right\rangle\right) \\ &\le 2Q_{\max}E(K)\ .\end{aligned}$$

For the third term, we have

$$\sum_{k=1}^{K}\langle \mu_k^+(\pi_{k+1}) - \mu_k^+(\pi_k), Q_{k+1}\rangle = \sum_{e=1}^{E(K)}\sum_{k\in\mathcal{K}_e}\left\langle \mu_{k_e}^+(\pi_{k+1}) - \mu_{k_e}^+(\pi_k), Q_{k+1}\right\rangle\ .$$

Successively applying Hölder's inequality, Pinsker's inequality and Lemma 7,

$$\begin{aligned}\left\langle \mu_{k_e}^+(\pi_{k+1}) - \mu_{k_e}^+(\pi_k), Q_{k+1}\right\rangle &\le Q_{\max}\left\|\mu_{k_e}^+(\pi_{k+1}) - \mu_{k_e}^+(\pi_k)\right\|_1 \\ &\le Q_{\max}\sqrt{2\mathcal{D}_{KL}\left(\mu_{k_e}^+(\pi_{k+1})\,\|\mu_{k_e}^+(\pi_k)\right)} \\ &\le Q_{\max}\sqrt{\frac{2}{1-\gamma}\left\langle \nu_{k_e}^+(\pi_{k+1}), \mathcal{D}_{KL}(\pi_{k+1}\|\pi_k)\right\rangle}\,.\end{aligned}$$

For any $x\in\mathcal{X}$, the KL divergence between π_{k+1} and π_k in state x can be bounded as

$$\begin{aligned}&\mathcal{D}_{KL}(\pi_{k+1}\|\pi_k)(x) \\ &\quad= \sum_{a\in\mathcal{A}}\pi_{k+1}(a\mid x)\left(\eta Q_{k+1}(x,a) - \log\left(\sum_{b\in\mathcal{A}}\pi_k(b\mid x)\exp\left[\eta Q_{k+1}(x,b)\right]\right)\right) \\ &\quad= \eta\sum_{a\in\mathcal{A}}\pi_{k+1}(a\mid x)Q_{k+1}(x,a) - \log\left(\sum_{b\in\mathcal{A}}\pi_k(b\mid x)\exp\left[\eta Q_{k+1}(x,b)\right]\right) \\ &\quad\le \eta\sum_{a\in\mathcal{A}}\left[\pi_{k+1}(a\mid x) - \pi_k(a\mid x)\right]Q_{k+1}(x,a) \\ &\quad\le \eta Q_{\max}\left\|\pi_{k+1}(\cdot\mid x) - \pi_k(\cdot\mid x)\right\|_1 \\ &\quad\le \eta Q_{\max}\sqrt{2\mathcal{D}_{KL}(\pi_{k+1}\|\pi_k)(x)}\,,\end{aligned}$$

where the first inequality follows from Jensen's and the convexity of $-\log$, the second inequality is by Hölder's and the boundedness of Q_k, and the last inequality is due to Pinkser's. Dividing by $\sqrt{\mathcal{D}_{KL}(\pi_{k+1}\|\pi_k)(x)}$ and squaring the inequality, we get

$$\mathcal{D}_{KL}(\pi_{k+1}\|\pi_k)(x) \le 2\eta^2 Q_{\max}^2\,.$$

Plugging this back into the previous inequality and summing over $k\in[K]$, we get

$$\sum_{k=1}^{K}\langle \mu_k^+(\pi_{k+1}) - \mu_k^+(\pi_k), Q_{k+1}\rangle \le \frac{2Q_{\max}^2\eta K}{\sqrt{1-\gamma}}\,.$$

The remaining term is nonpositive. The sum of the Δ_k terms is thus bounded by

$$\sum_{k=1}^{K}\Delta_k \le 2Q_{\max}E(K) + \frac{2Q_{\max}^2\eta K}{\sqrt{1-\gamma}} + 2\sum_{k=1}^{K}\langle \mu(\pi_k), (1-p_k^+)\odot \mathrm{CB}_k\rangle\ .$$

Finally, combining the bounds on $\sum_{k=1}^{K}\Delta_k^\star$ and $\sum_{k=1}^{K}\Delta_k$, we get

$$\mathfrak{R}_{\mathrm{K}}^{+} \leq \frac{E\left(K\right)\log\left|\mathcal{A}\right|}{\eta} + 4Q_{\max}E\left(K\right) + \frac{2Q_{\max}^{2}\eta K}{\sqrt{1-\gamma}} + 2\sum_{k=1}^{K}\left\langle \mu\left(\pi_k\right), \left(1-p_k^{+}\right)\odot \mathrm{CB}_k\right\rangle \,.$$

B.4 Proof of Lemma 4 (choice of $Q_{\max}$)

Lemma 4 *Suppose the bonuses $\{\mathrm{CB}_k\}_{k\in[K]}$ are valid in the sense of Equation 2 and the ascension functions are chosen as in Line 22 of Algorithm 1. Then, for any $k\in[K]$, the iterate Q_k satisfies $\left\|Q_k\right\|_\infty \leq Q_{\max}$ with $Q_{\max} = \frac{R_{\max}+2\omega/\alpha}{1-\gamma}$.*

Proof 6 *We want to find $Q_{\max}$ such that for any k, $\left\|Q_k\right\|_\infty \leq Q_{\max}$. Since V_k is a log-sum-exp of Q_k, we have $\left\|V_k\right\|_\infty \leq \left\|Q_k\right\|_\infty$. Next, we proceed by induction to find a suitable value of $Q_{\max}$. Let $k\in\mathbb{N}^\star$ and assume $\left\|Q_k\right\|_\infty \leq Q_{\max}$. For any (x,a),*

$$\begin{aligned}
\left|Q_{k+1}\left(x,a\right)\right| &\leq r_k^{+}\left(x,a\right) + \left(1-p_k^{+}\left(x,a\right)\right)\mathrm{CB}_k\left(x,a\right) + \gamma\left|\widehat{P}_k^{+}V_k\left(x,a\right)\right| \\
&\leq r_k^{+}\left(x,a\right) + 2\left(1-p_k^{+}\left(x,a\right)\right)\mathrm{CB}_k\left(x,a\right) + \gamma P_k^{+}V_k\left(x,a\right) \\
&\leq R_{\max} + 2\left(1-p_k^{+}\left(x,a\right)\right)\mathrm{CB}_k\left(x,a\right) + \gamma Q_{\max}\,,
\end{aligned}$$

where the second inequality follows from the validity of the bonuses (and corresponds to the upper bound on Q_{k+1} from Lemma 7), and the third inequality is due to the inductive assumption and the boundedness of the rewards. Plugging the definition of the probabilities p_k^{+}, we further get

$$\begin{aligned}
\left|Q_{k+1}\left(x,a\right)\right| &\leq R_{\max} + 2\sigma\left(\omega - \alpha\,\mathrm{CB}_k\left(x,a\right)\right)\mathrm{CB}_k\left(x,a\right) + \gamma Q_{\max} \\
&\leq R_{\max} + \gamma Q_{\max} + 2\sup_{z\geq 0}\left\{\sigma\left(\omega-\alpha z\right)z\right\} \\
&\leq R_{\max} + \gamma Q_{\max} + \frac{2\omega}{\alpha}\,,
\end{aligned}$$

the last inequality is simply a property of the sigmoid function and is showm in Lemma 11. For the induction to work at time $k+1$, we need to set $Q_{\max}$ such that $Q_{\max} = R_{\max} + \gamma Q_{\max} + \frac{2\omega}{\alpha}$, that is

$$Q_{\max} = \frac{R_{\max} + 2\omega/\alpha}{1-\gamma}\,.$$

The initial case is also true since $\left\|Q_1\right\|_\infty \leq \frac{R_{\max}}{1-\gamma} \leq Q_{\max}$.

B.5 Proof of Lemma 5 (bound on bonuses)

We now control the sum of bonuses. For any episode k, we will denote $\mathcal{T}_k$ the set of timesteps in episode k.

Lemma 5 *Suppose Assumption 1 and event $\mathcal{E}_L$ hold and denote $T = L_{\max}K$. Then, with probability at least $1-\delta$, the policies $\{\pi_k\}$ and bonuses $\{\mathrm{CB}_k\}$ satisfy*

$$\sum_{k=1}^{K}\left\langle \mu\left(\pi_k\right), \mathrm{CB}_k\right\rangle \leq 4\left(1-\gamma\right)\beta B\sqrt{dT\log\left(1+\frac{B^2T}{d}\right)} + 4\beta B\log\left(\frac{2K}{\delta}\right)^2\,,$$

and

$$\sum_{k=1}^{K}\left\langle \mu\left(\pi_k\right), \mathrm{CB}_k^2\right\rangle \leq 8\left(1-\gamma\right)\beta^2B^2d\log\left(1+\frac{B^2T}{d}\right) + 4\beta^2B^2\log\left(\frac{2K}{\delta}\right)^2\,,$$

To prove Lemma 5, we need the following result.

Lemma 1 *Suppose $\mathcal{E}_L$ holds. Let $\{f_k\}_{k\in[K]} \subset \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ be a sequence of functions with values in $[0,M]$ almost surely. Then, with probability at least $1-\delta$ the schedule and policies produced by Algorithm 1 satisfy*

$$\sum_{k=1}^{K}\left\langle \mu\left(\pi_k\right), f_k\right\rangle \leq 2\left(1-\gamma\right)\sum_{k=1}^{K}\sum_{t\in\mathcal{T}_k} f_k\left(X_t,A_t\right) + 4M\log\left(\frac{2K}{\delta}\right)^2\,.$$

Proof 7 *We denote $\mathcal{F}_{k-1}$ the σ-field generated by the history up to the end of episode $k-1$. We have,*

$$\begin{aligned}
\langle \mu(\pi_k), f_k \rangle &= (1-\gamma)\,\mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau f_k(X_\tau, A_\tau) \,\middle|\, \mathcal{F}_{k-1}\right] \\
&= (1-\gamma)\,\mathbb{E}\left[\sum_{\tau=0}^{\infty} \mathbb{I}_{\{\tau<L_k\}} f_k(X_\tau, A_\tau) \,\middle|\, \mathcal{F}_{k-1}\right] \\
&= (1-\gamma)\,\mathbb{E}\left[\sum_{\tau=0}^{L_k-1} f_k(X_\tau, A_\tau) \,\middle|\, \mathcal{F}_{k-1}\right] \\
&= (1-\gamma)\,\mathbb{E}\left[\sum_{\tau\in\mathcal{T}_k} f_k(X_\tau, A_\tau) \,\middle|\, \mathcal{F}_{k-1}\right].
\end{aligned}$$

Plugging it back in the previous display,

$$\sum_{k=1}^{K} \langle \mu(\pi_k), f_k \rangle = (1-\gamma) \sum_{k=1}^{K} \mathbb{E}\left[\sum_{t\in\mathcal{T}_k} f_k(X_t, A_t) \,\middle|\, \mathcal{F}_{k-1}\right].$$

Since we assume $\mathcal{E}_L$ holds, for any k we have that $\sum_{t\in\mathcal{T}_k} f_k(X_t, A_t)$ takes values in $[0, ML_{\max}]$ almost surely. Using the concentration inequality from Lemma 15, we get

$$\begin{aligned}
\sum_{k=1}^{K} \langle \mu(\pi_k), f_k \rangle &\le 2(1-\gamma) \sum_{k=1}^{K} \sum_{t\in\mathcal{T}_k} f_k(X_t, A_t) + 4(1-\gamma) ML_{\max} \log\left(\frac{2K}{\delta}\right) \\
&\le 2(1-\gamma) \sum_{k=1}^{K} \sum_{t\in\mathcal{T}_k} f_k(X_t, A_t) + 4M \log\left(\frac{2K}{\delta}\right)^2,
\end{aligned}$$

where we used that $L_{\max} = \frac{\log(K/\delta)}{1-\gamma}$.

We now prove Lemma 5. With a slight abuse of notation, we use the convention that for any epoch e and any t in epoch e, the bonuses at time step t are $\mathrm{CB}_t = \mathrm{CB}_{t_e}$. Noting that the bonuses CB_t take values in $[0, \beta B]$, we apply Lemma 1 to get

$$\sum_{k=1}^{K} \langle \mu(\pi_k), \mathrm{CB}_k \rangle \le 2(1-\gamma) \sum_{t=1}^{T} \mathrm{CB}_t(X_t, A_t) + 4\beta B \log\left(\frac{2K}{\delta}\right)^2,$$

where we used $\mathrm{CB}_t \succeq 0$, and $T_{K+1} \le T$ which follows from the event $\mathcal{E}_L$. Likewise, applying Lemma 1 to $\mathrm{CB}_t^2 \in \left[0, \beta^2 B^2\right]$, we obtain a similar bound for the term $\sum_{k=1}^{K} \left\langle \mu(\pi_k), \mathrm{CB}_k^2 \right\rangle$,

$$\sum_{k=1}^{K} \left\langle \mu(\pi_k), \mathrm{CB}_k^2 \right\rangle \le 2(1-\gamma) \sum_{t=1}^{T} \mathrm{CB}_t(X_t, A_t)^2 + 4\beta^2 B^2 \log\left(\frac{2K}{\delta}\right)^2.$$

By Cauchy-Schwartz's inequality, we have $\sum_{t=1}^{T} \mathrm{CB}_t(X_t, A_t) \le \sqrt{T}\sqrt{\sum_{t=1}^{T} \mathrm{CB}_t(X_t, A_t)^2}$, so we can focus on the latter sum. By definition of the bonuses for linear MDPs, we have

$$\sum_{t=1}^{T} \mathrm{CB}_t(X_t, A_t)^2 = \beta^2 \sum_{e=1}^{E(K)} \sum_{k\in\mathcal{K}_e} \sum_{t\in\mathcal{T}_k} \|\varphi(X_t, A_t)\|^2_{\Lambda_{t_e}^{-1}}.$$

Since the covariance matrix only contains the data until the beginning of the epoch, there is a delay with $\varphi(X_t, A_t)$ which is further ahead. To compensate for this, note that for any $t \in [t_e, t_{e+1} - 1]$, we have $\det \Lambda_t \le 2 \det \Lambda_{t_e}$ due to the update condition in Algorithm 1, so by Lemma 13

$$\|\varphi(X_t, A_t)\|^2_{\Lambda_{t_e}^{-1}} \le \frac{\det\left(\Lambda_{t_e}^{-1}\right)}{\det\left(\Lambda_t^{-1}\right)} \|\varphi(X_t, A_t)\|^2_{\Lambda_t^{-1}} \le 2\|\varphi(X_t, A_t)\|^2_{\Lambda_t^{-1}}.$$

We plug this back into the previous inequality and apply Lemma 16 to obtain[2]

$$\sum_{t=1}^{T} \mathrm{CB}_t\left(X_t, A_t\right)^2 \leq 2\beta^2 \sum_{t=1}^{T} \left\|\varphi\left(X_t, A_t\right)\right\|^2_{\Lambda_t^{-1}} \leq 4\beta^2 B^2 \log\left(\frac{\det \Lambda_T}{\det \Lambda_0}\right) .$$

Using the definition of Λ_0, Λ_T, the trace-determinant inequality, and the assumption $\|\varphi(\cdot,\cdot)\|_2 \leq B$, we finally get

$$\begin{aligned}\sum_{t=1}^{T} \mathrm{CB}_t\left(X_t, A_t\right)^2 &\leq 4\beta^2 B^2 d \log\left(\frac{d + \sum_{t=1}^{T} \left\|\varphi\left(X_t, A_t\right)\right\|_2^2}{d}\right) \\ &\leq 4\beta^2 B^2 d \log\left(1 + \frac{B^2 T}{d}\right) .\end{aligned}$$

The conclusion follows from plugging this back into the inequalities of interest.

B.6 Proof of Lemma 6 (good event holds)

Before stating the proof of Lemma 6, we need to define some auxiliary quantities and state two intermediate results. First recall that $\{L_k\}_{k=1}^K$ denote the number of steps between consecutive resets and that for any $k \geq 2$, $L_k = T_k - T_{k-1}$, and $L_1 = T_1$. We need to prove the episodes are not too long, *i.e.* $\mathcal{E}_L = \{\forall k \in [K], L_k \leq L_{\max}\}$ holds with high probability, where $L_{\max} = H \log(K/\delta)$. This is done in Lemma 2. Then, we define the event $\mathcal{E}_{V,\mathrm{alg}}$ on the iterates generated by Algorithm 1

$$\mathcal{E}_{V,\mathrm{alg}} = \left\{\forall k \in [K], \left\|MV_k - \widehat{MV_k}\right\|_{\Lambda_{T_k}} \leq \beta\right\} ,$$

where $\widehat{MV_k} = \Lambda_{T_k}^{-1} \sum_{(x,a,x') \in \mathcal{D}_{T_k}} \varphi(x,a) V_k(x')$. To prove $\mathcal{E}_{V,\mathrm{alg}}$ holds with high probability, we need to resort to a standard uniform covering argument first introduced by Jin et al., 2019. To do so, let us denote with $p^+_{\Lambda,\beta,\alpha} = \sigma(\alpha\beta \|\varphi(\cdot,\cdot)\|_\Lambda - w) = 1 - \sigma(-\alpha\beta \|\varphi(\cdot,\cdot)\|_\Lambda + \omega)$ an ascension function parametrized by the matrix Λ, the scalar β and the sigmoid slope α. Then, we define the following class of functions on $\mathcal{X} \times \mathcal{A}$

$$\begin{aligned}\mathcal{Q} = \Bigg\{ & Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R} \quad \text{s.t.} \\ & Q = \left(1 - p^+_{\Lambda,\beta,\alpha}\right) \odot (\Phi\theta + \beta \|\varphi(\cdot,\cdot)\|_\Lambda) + p^+_{\Lambda,\beta,\alpha} \cdot \frac{R_{\max}}{1-\gamma}, \\ & \beta = \widetilde{\mathcal{O}}\left(Q_{\max} d\right), \quad \alpha = 2\omega, \quad \lambda_{\max}(\Lambda) \leq 1, \quad \lambda_{\min}(\Lambda) \geq \frac{1}{2KBL_{\max}}, \\ & \|\theta\| \leq W_{\max} + Q_{\max} L_{\max} KB, \|Q\|_\infty \leq Q_{\max} \Bigg\} \cup \{\mathbf{0}\} ,\end{aligned}$$

where $Q_{\max} = H\left(R_{\max} + \frac{2\omega}{\alpha}\right)$ and we included the function 0 to make sure $Q_1 \in \mathcal{Q}$. Furthermore, denote for any $\eta > 0$ the function $f_\eta : \mathbb{R}^{\mathcal{X}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{X}}$ defined for $Q \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ as $f_\eta(Q) = \frac{1}{\eta} \log \sum_{a\in\mathcal{A}} \exp(\eta Q(\cdot, a))$. We then define the following function class in $\mathbb{R}^{\mathcal{X}}$

$$\mathcal{V} = \left\{V : \mathcal{X} \to \mathbb{R} \quad \text{s.t.} \quad \exists \{Q_\ell\}_{\ell=1}^K, \{\bar{Q}_\ell\}_{\ell=1}^K \subset \mathcal{Q}, V = f_\eta \circ \left(\sum_{\ell=1}^K Q_\ell\right) - f_\eta \circ \left(\sum_{\ell=1}^K \bar{Q}_\ell\right)\right\}, \tag{7}$$

as well as the event

$$\mathcal{E}_{\mathcal{V}} = \left\{\forall V \in \mathcal{V}, \forall k \in [K], \left\|MV - \widehat{MV}\right\|_{\Lambda_{T_k}} \leq \beta\right\} ,$$

where $\widehat{MV} = \Lambda_{T_k}^{-1} \sum_{(x,a,x') \in \mathcal{D}_{T_k}} \varphi(x,a) V(x')$. Finally, we define the event that the iterates of Algorithm 1 are in the function class $\mathcal{V}$

$$\mathcal{E}_{\mathrm{in}} = \{\forall k \in [K], V_k \in \mathcal{V}\} .$$

What remains is to show that the iterates of the algorithm belong to $\mathcal{V}$, and that the event $\mathcal{E}_{\mathcal{V}}$ holds with high probability. This is done in Lemmas 3 and 4, respectively. We can now prove Lemma 6.

[2]Note that this is where the linear dependency in B appears, but this can be removed by setting $\lambda = 1/B^2$.

Lemma 6 *Let* $\beta = 8Q_{\max} d \log\left(c\alpha W_{\max} R_{\max} B^{9/2} Q_{\max}^4 L_{\max}^{5/2} K^{7/2} d^{5/2} \delta^{-1}\right)$. *Then, the event* $\mathcal{E}_{valid} \cap \mathcal{E}_L$ *holds with probability* $1 - 2\delta$.

Proof 8 *For any episode* $k \in [K]$ *and state-action pair* $(x, a) \in \mathcal{X} \times \mathcal{A}$, *we have by Cauchy-Schwartz's inequality*

$$\left|PV_k(x,a) - \widehat{PV_k}(x,a)\right| \leq \left\|MV_k - \widehat{MV_k}\right\|_{\Lambda_{T_k}} \|\varphi(x,a)\|_{\Lambda_{T_k}^{-1}} .$$

This inequality shows that the event $\mathcal{E}_{V,\text{alg}}$ *implies the event* $\mathcal{E}_{\text{valid}}$, *i.e.*

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}_{\text{valid}} \cap \mathcal{E}_L\right] &= \mathbb{P}\left[\mathcal{E}_{\text{valid}} \mid \mathcal{E}_L\right] \mathbb{P}\left[\mathcal{E}_L\right] \\ &\geq \mathbb{P}\left[\mathcal{E}_{V,\text{alg}} \mid \mathcal{E}_L\right] \mathbb{P}\left[\mathcal{E}_L\right] \\ &\geq \mathbb{P}\left[\mathcal{E}_{V,\text{alg}} \cap \mathcal{E}_{\text{in}} \mid \mathcal{E}_L\right](1-\delta) ,\end{aligned}$$

where in the last inequality we used the monotonicity of $\mathbb{P}$ *and Lemma 2. Then, conditioned on the event* $\mathcal{E}_{\text{in}}$ *that the iterates are in the function class* $\mathcal{V}$, *the event* $\mathcal{E}_{\mathcal{V}}$ *implies* $\mathcal{E}_{V,\text{alg}}$, *that is*

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}_{V,\text{alg}} \cap \mathcal{E}_{\text{in}} \mid \mathcal{E}_L\right] &= \mathbb{P}\left[\mathcal{E}_{V,\text{alg}} \mid \mathcal{E}_{\text{in}} \cap \mathcal{E}_L\right] \mathbb{P}\left[\mathcal{E}_{\text{in}} \mid \mathcal{E}_L\right] \\ &\geq \mathbb{P}\left[\mathcal{E}_{\mathcal{V}} \mid \mathcal{E}_{\text{in}} \cap \mathcal{E}_L\right] \mathbb{P}\left[\mathcal{E}_{\text{in}} \mid \mathcal{E}_L\right] \\ &= \mathbb{P}\left[\mathcal{E}_{\mathcal{V}} \cap \mathcal{E}_{\text{in}} \mid \mathcal{E}_L\right] .\end{aligned}$$

Finally, by Lemma 3 we have $\mathbb{P}\left[\mathcal{E}_{\text{in}} \mid \mathcal{E}_{\mathcal{V}}, \mathcal{E}_L\right] = 1$ *thus*

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}_{\mathcal{V}} \cap \mathcal{E}_{\text{in}} \mid \mathcal{E}_L\right] &= \mathbb{P}\left[\mathcal{E}_{\text{in}} \mid \mathcal{E}_{\mathcal{V}}, \mathcal{E}_L\right] \mathbb{P}\left[\mathcal{E}_{\mathcal{V}} \mid \mathcal{E}_L\right] \\ &= \mathbb{P}\left[\mathcal{E}_{\mathcal{V}} \mid \mathcal{E}_L\right] \\ &\geq 1 - \delta ,\end{aligned}$$

where the last inequality follows from Lemma 4. In conclusion, we get

$$\mathbb{P}\left[\mathcal{E}_{\text{valid}} \cap \mathcal{E}_L\right] \geq (1-\delta)^2 \geq 1 - 2\delta .$$

We now show the episodes are not too long.

Lemma 2 *Let* $\delta \in (0,1)$ *and define* $L_{\max} = H \log(K/\delta)$. *Then, the event* $\mathcal{E}_L$ *holds with probability at least* $1 - \delta$.

Proof 9 *For any* k *and by definition of the cumulative density function of the geometric distribution with parameter* $1-\gamma$, *we have that* $\mathbb{P}\left[L_k \leq L_{\max}\right] = 1 - \gamma_{\max}^L$. *Therefore,* $\mathbb{P}\left[L_k \leq L_{\max}\right] \geq 1 - \delta/K$ *for* $L_{\max} \geq \frac{\log\left(\frac{\delta}{K}\right)}{\log(1/\gamma)}$. *Lower bounding the denominator as* $\log(1/\gamma) \geq 1 - \gamma$, *we have that for* $L_{\max} = \frac{\log(K/\delta)}{1-\gamma}$ *and a union bound over* $k \in [K]$, *we have that* $\mathbb{P}\left[\mathcal{E}_L\right] \geq 1 - \delta$.

Lemma 3 *Assume the events* $\mathcal{E}_{\mathcal{V}}$ *and* $\mathcal{E}_L$ *hold. Then, for all* $k \in [K]$, *it holds that* $V_k \in \mathcal{V}$, *i.e.* $\mathcal{E}_{\text{in}}$ *holds.*

Proof 10 *The bound is proven by induction over* $k \in [K]$. *The base case holds by initialization since* $Q_0 = \mathbf{0}$ *is in* $\mathcal{Q}$. *For the induction step, we assume that for all* $\ell \in [k]$, $Q_\ell \in \mathcal{Q}, V_\ell \in \mathcal{V}$ *and we show that* $Q_{k+1} \in \mathcal{Q}$ *and* $V_{k+1} \in \mathcal{V}$.

By definition of the function classes $\mathcal{Q}$ *and* $\mathcal{V}$ *it holds that* $\|Q_k\|_\infty, \|V_k\|_\infty \leq Q_{\max}$. $\mathcal{E}_{\mathcal{V}}$ *together with the induction assumption imply that the bonuses are valid at time* k, *meaning that the derivations from Lemma 4 guarantee that* $\|Q_{k+1}\|_\infty \leq Q_{\max}$. *Moreover, denote* θ_{k+1} *the vector used to represent* Q_{k+1}, *defined as*

$$\theta_{k+1} = w_k + \gamma \widehat{MV_k} = w_k + \gamma \Lambda_{T_k}^{-1} \sum_{(x,a,x') \in \mathcal{D}_{T_k}} \varphi(x,a) V_k(x') .$$

It remains to show that θ_{k+1} *satisfies the norm constraint defined in* $\mathcal{Q}$. *By the triangular inequality and plugging the various assumptions, we have*

$$\begin{aligned}\|\theta_{k+1}\| &\leq \|w_k\| + \gamma \left\| (\Lambda_{T_k})^{-1} \sum_{(x,a,x') \in \mathcal{D}_{T_k}} \varphi(x,a) V_k(x') \right\| \\ &\leq W_{\max} + \gamma \lambda_{\max}\left((\Lambda_{T_k})^{-1}\right) |\mathcal{D}_{T_k}| \, \|V_k\|_\infty \max_{x,a} \|\varphi(x,a)\|_2 \\ &\leq W_{\max} + K L_{\max} Q_{\max} B ,\end{aligned}$$

where we also used $\gamma < 1$ in the last inequality. This proves that $Q_{k+1} \in \mathcal{Q}$. Therefore, we have that $Q_\ell \in \mathcal{Q}$ for $\ell \in [k+1]$. We now show that $V_{k+1} \in \mathcal{V}$. Let $x \in \mathcal{X}$ and k_e be the initial index of the epoch e such that $k \in \mathcal{K}_e$. By Moulin and Neu, 2023, Lemma C.1, the sum of V iterates is equal to a log-sum-exp function of the sum of Q iterates. Thus,

$$\begin{aligned} V_{k+1}(x) &= \sum_{i=k_e}^{k+1} V_i(x) - \sum_{j=k_e}^{k} V_j(x) \\ &= f_\eta\left(\sum_{i=k_e}^{k+1} Q_i\right)(x) - f_\eta\left(\sum_{j=k_e}^{k} Q_j\right)(x) \,. \end{aligned}$$

Since $\mathbf{0} \in \mathcal{Q}$ and $Q_\ell \in \mathcal{Q}$ for $\ell \in [k+1]$, we can pad with zeros the two sums inside the exponentials and conclude that $V_{k+1}(x)$ can be written as the difference between two log-sum-exp functions of the sum of K functions in $\mathcal{Q}$. Thus $V_{k+1} \in \mathcal{V}$ and this concludes the induction.

Lemma 4 *Assume the event $\mathcal{E}_L$ holds, and set β as*

$$\beta = 8Q_{\max} d \log(c\alpha W_{\max} R_{\max} B^{9/2} Q_{\max}^4 L_{\max}^{5/2} K^{7/2} d^{5/2} \delta^{-1}) \,.$$

where $c = 60 \cdot 26$. Then, the event $\mathcal{E}_\mathcal{V}$ holds with probability $1 - \delta$.

Proof 11 *Under the event $\mathcal{E}_L$, invoking standard concentration results for Linear MDPs (see Lemmas D.3 and D.4 in Jin et al. [2019]), we have that with probability $1 - \delta$ it holds that*

$$\begin{aligned} &\left\| MV - (\Lambda_{T_k})^{-1} \sum_{(x,a,x') \in \mathcal{D}_{T_k}} \varphi(x,a) V(x') \right\|_{\Lambda_{T_k}} \\ &\qquad \le Q_{\max} \sqrt{2d \log\left(\frac{1 + KL_{\max}B}{\delta}\right) + 4 \log \mathcal{N}_\epsilon + 8K^2 L_{\max}^2 B^2 \epsilon^2} \,, \end{aligned}$$

where $\mathcal{N}_\epsilon$ is the ϵ-covering number of the class $\mathcal{V}$. In particular, for $\epsilon = (KL_{\max}B)^{-1}$, we can invoke Lemma 5 to obtain

$$\begin{aligned} \log \mathcal{N}_\epsilon &\le 4d^2 \log\left(4(W_{\max}B + Q_{\max}L_{\max}KB^2 + 3\sqrt{d} + \beta B + R_{\max})\sqrt{K^5 L_{\max}^3}\alpha\beta B^{5/2}\right) \\ &\le 4d^2 \log\left(20(3W_{\max}\beta Q_{\max}L_{\max}KB^2\sqrt{d}R_{\max})\sqrt{K^5 L_{\max}^3}\alpha\beta B^{5/2}\right) \\ &\le 4d^2 \log\left(60 W_{\max} R_{\max} \beta^2 \alpha B^{9/2} Q_{\max}^2 L_{\max}^{5/2} K^{7/2}\sqrt{d}\right) \,. \end{aligned}$$

Plugging in, we have that

$$\begin{aligned} &\left\| MV - (\Lambda_{T_k})^{-1} \sum_{(x,a,x') \in \mathcal{D}_{T_k}} \varphi(x,a) V(x') \right\|_{\Lambda_{T_k}} \\ &\quad \le Q_{\max} \sqrt{2d \log\left(\frac{1 + KL_{\max}B}{\delta}\right) + 16d^2 \log\left(60\beta^2\alpha B^{9/2} Q_{\max}^2 L_{\max}^{5/2} K^{7/2}\sqrt{d}\right) + 8} \\ &\quad \le Q_{\max} \sqrt{26d^2 \log\left(\frac{60 W_{\max} R_{\max} \beta^2 \alpha B^{9/2} Q_{\max}^2 L_{\max}^{5/2} K^{7/2}\sqrt{d}}{\delta}\right)} \\ &\quad = \sqrt{26 Q_{\max}^2 d^2 \log\left(\frac{60 W_{\max} R_{\max} \beta^2 \alpha B^{9/2} Q_{\max}^2 L_{\max}^{5/2} K^{7/2}\sqrt{d}}{\delta}\right)} \end{aligned}$$

At this point, to find a value for β such that

$$\beta^2 \ge 26 Q_{\max}^2 d^2 \log\left(\frac{60 W_{\max} R_{\max} \beta^2 \alpha B^{9/2} Q_{\max}^2 L_{\max}^{5/2} K^{7/2}\sqrt{d}}{\delta}\right) \,,$$

we invoke Lemma 14 with $z = 26Q_{\max}^2 d^2$ *and* $R = \frac{60 W_{\max} R_{\max} \alpha B^{9/2} Q_{\max}^2 L_{\max}^{5/2} K^{7/2} \sqrt{d}}{\delta}$ *which gives that the desired inequality holds for all* $\beta \in \mathbb{R}$ *such that*

$$\beta^2 \geq 52 Q_{\max}^2 d^2 \log(c\alpha W_{\max} R_{\max} B^{9/2} Q_{\max}^4 L_{\max}^{5/2} K^{7/2} d^{5/2} \delta^{-1}) ,$$

where $c = 60 \cdot 26$. *Therefore, we select*

$$\beta = 8 Q_{\max} d \log(c\alpha W_{\max} R_{\max} B^{9/2} Q_{\max}^4 L_{\max}^{5/2} K^{7/2} d^{5/2} \delta^{-1}) .$$

Remark 1 *For the proof of Lemma 4, we need to compute a bound on the covering number of the function class* $\mathcal{V}$. *We find this is done in a neat and more direct way than previous analysis Zhong and Zhang [2024], Sherman et al. [2023a], Cassel and Rosenberg [2024] that needed to introduce a policy class for the iterates* $\{\pi_k\}_{k=1}^K$ *generated by Algorithm 1 as an intermediate step.*

B.6.1 Proof of Lemma 5 (covering number)

Lemma 5 *Let us consider the function class* $\mathcal{V}$ *defined in Equation (7) and an* ϵ*-covering set* $\mathcal{R}(\mathcal{V})$ *such that for any* $V \in \mathcal{V}$, *there exists* $V' \in \mathcal{R}(\mathcal{V})$ *such that* $\|V - V'\|_\infty \leq \frac{1}{K L_{\max} B}$. *The covering number of the class* $\mathcal{V}$ *can be bounded as follows*

$$\log \mathcal{N}_{\frac{1}{K}} \leq 4d^2 \log\left(4\left(W_{\max} B + Q_{\max} L_{\max} K B^2 + 3\sqrt{d} + \beta B + R_{\max} H\right)\sqrt{K^5 L_{\max}^3} \alpha\beta B^{5/2}\right) .$$

Proof 12 *We will use the following intermediate class of log sum exp state value functions*

$$\tilde{\mathcal{V}} = \left\{ V : \mathcal{X} \to \mathbb{R} \quad s.t. \quad \forall x, V(x) = \frac{1}{\eta} \log \sum_{a \in \mathcal{A}} \exp\left(\eta \sum_{\ell=1}^K Q_\ell(x,a)\right), Q_\ell \in \mathcal{Q} \right\} .$$

Consider any $V, V' \in \mathcal{V}$, *and notice that for any* $x \in \mathcal{X}$,

$$|V(x) - V'(x)| \leq \left|\bar{V}(x) - \bar{V}'(x)\right| + \left|\tilde{V}(x) - \tilde{V}'(x)\right| .$$

with $\bar{V}, \tilde{V} \in \tilde{\mathcal{V}}$ *such that* $V(x) = \bar{V}(x) - \tilde{V}(x)$ *for all* $x \in \mathcal{X}$ *and with* $\bar{V}', \tilde{V}' \in \tilde{\mathcal{V}}$ *such that* $V'(x) = \bar{V}'(x) - \tilde{V}'(x)$ *for all* $x \in \mathcal{X}$. *Therefore, the above bound guarantees that an* $\epsilon/2$ *covering set on the function class* $\tilde{\mathcal{V}}$ *implies an* ϵ *covering for the class* $\mathcal{V}$. *Hence, in the following we focus on computing a* $\epsilon/2$ *covering number for* $\tilde{\mathcal{V}}$. *By definition of* $\bar{V}, \bar{V}'$ *and Lemma 12, we have*

$$\begin{aligned}
\left|\bar{V}(x) - \bar{V}'(x)\right| &= \left|\frac{1}{\eta} \log \sum_{a \in \mathcal{A}} \exp\left(\eta \sum_{\ell=1}^K \bar{Q}_\ell(x,a)\right) - \frac{1}{\eta} \log \sum_{a \in \mathcal{A}} \exp\left(\eta \sum_{\ell=1}^K \bar{Q}'_\ell(x,a)\right)\right| \\
&\leq \max_{a \in \mathcal{A}} \left|\sum_{\ell=1}^K \bar{Q}_\ell(x,a) - \sum_{\ell=1}^K \bar{Q}'_\ell(x,a)\right| \\
&\leq K \max_{\ell \in [K]} \left\|\bar{Q}_\ell - \bar{Q}'_\ell\right\|_\infty .
\end{aligned}$$

For any $\ell \in [K]$, *we denote* Λ_ℓ, θ_ℓ *the parameters of the function* $\bar{Q}_\ell$ *and* Λ'_ℓ, θ'_ℓ *the parameters of the function* $\bar{Q}'_\ell$. *We now prove that* $\bar{Q}_\ell, \bar{Q}'_\ell$ *are Lipschitz functions. Let us denote* $Q_{\theta,\Lambda}$ *and* $Q_{\theta',\Lambda'}$ *two functions in* $\mathcal{Q}$ *for different parameters* $\theta, \Lambda, \theta', \Lambda'$. *For any state-action pair* (x,a), *the difference between the two functions can be written as*

$$\begin{aligned}
Q_{\theta,\Lambda}(x,a) - Q_{\theta',\Lambda'}(x,a) = &\left(\varphi(x,a)^\intercal \theta + \beta \|\varphi(x,a)\|_\Lambda - R_{\max} H\right) \cdot \sigma\left(-\alpha\beta \|\varphi(x,a)\|_\Lambda + \omega\right) \\
&- \left(\varphi(x,a)^\intercal \theta' + \beta \|\varphi(x,a)\|_{\Lambda'} - R_{\max} H\right) \cdot \sigma\left(-\alpha\beta \|\varphi(x,a)\|_{\Lambda'} + \omega\right) .
\end{aligned}$$

Next, our goal is to show that the function

$$f(\theta, \Lambda; x, a) := (\varphi(x,a)^\intercal \theta + \beta \|\varphi(x,a)\|_\Lambda - R_{\max} H) \cdot \sigma(-\alpha\beta \|\varphi(x,a)\|_\Lambda + \omega)$$

is Lipshitz in both parameters θ, Λ. *The Lipshitzness with respect to* β *does not need to be established since it is kept fixed throughout the learning process. We show this showing that the gradients are bounded. In particular,*

$$\|\nabla_\theta f(\theta, \Lambda; x, a)\| = \|\varphi(x,a)\| \cdot \sigma(-\alpha\beta \|\varphi(x,a)\|_\Lambda + \omega) \leq \|\varphi(x,a)\| \leq B .$$

For the Lipshitzness with respect to Λ, we have that

$$\begin{aligned}
&f(\theta,\Lambda;x,a)-f(\theta,\Lambda';x,a)\\
&=(\varphi(x,a)^\intercal\theta+\beta\left\|\varphi(x,a)\right\|_\Lambda-R_{\max}H)\cdot\sigma(-\alpha\beta\left\|\varphi(x,a)\right\|_\Lambda+\omega)\\
&\quad-(\varphi(x,a)^\intercal\theta+\beta\left\|\varphi(x,a)\right\|_{\Lambda'}-R_{\max}H)\cdot\sigma(-\alpha\beta\left\|\varphi(x,a)\right\|_{\Lambda'}+\omega)\\
&=(\varphi(x,a)^\intercal\theta+\beta\left\|\varphi(x,a)\right\|_\Lambda-R_{\max}H)\cdot\Big(\sigma(-\alpha\beta\left\|\varphi(x,a)\right\|_\Lambda+\omega)\\
&\quad-\sigma(-\alpha\beta\left\|\varphi(x,a)\right\|_{\Lambda'}+\omega)\Big)+\sigma(-\alpha\beta\left\|\varphi(x,a)\right\|_{\Lambda'}+\omega)\left(\beta\left\|\varphi(x,a)\right\|_\Lambda-\beta\left\|\varphi(x,a)\right\|_{\Lambda'}\right)
\end{aligned}$$

Then, using the fact that σ is 1-Lipshitz, we have that

$$\begin{aligned}
&|f(\theta,\Lambda;x,a)-f(\theta,\Lambda';x,a)|\\
&\quad\le\alpha\beta\left|\varphi(x,a)^\intercal\theta+\beta\left\|\varphi(x,a)\right\|_\Lambda-R_{\max}H\right|\cdot\left|\left\|\varphi(x,a)\right\|_\Lambda-\left\|\varphi(x,a)\right\|_{\Lambda'}\right|\\
&\qquad+\sigma(-\alpha\beta\left\|\varphi(x,a)\right\|_{\Lambda'}+\omega)\left|\beta\left\|\varphi(x,a)\right\|_\Lambda-\beta\left\|\varphi(x,a)\right\|_{\Lambda'}\right|\\
&\le\alpha\beta\left|\varphi(x,a)^\intercal\theta+\beta\left\|\varphi(x,a)\right\|_\Lambda-R_{\max}H\right|\cdot\left|\left\|\varphi(x,a)\right\|_\Lambda-\left\|\varphi(x,a)\right\|_{\Lambda'}\right|\\
&\quad+\beta\left|\left\|\varphi(x,a)\right\|_\Lambda-\left\|\varphi(x,a)\right\|_{\Lambda'}\right|\\
&\le\alpha\beta(\|\theta\|B+\beta B+R_{\max}H+1)\left|\left\|\varphi(x,a)\right\|_\Lambda-\left\|\varphi(x,a)\right\|_{\Lambda'}\right|.
\end{aligned}$$

where we used the fact that $\sigma(x)\le 1$, for all $x\in\mathbb{R}$ and $\alpha\ge 1$ in the last inequality. Using that $\|\varphi(x,a)\|_\Lambda=\left\|\Lambda^{1/2}\varphi(x,a)\right\|$ and the triangular inequality we have that

$$\begin{aligned}
f(\theta,\Lambda;x,a)-f(\theta,\Lambda';x,a)&\le\alpha\beta(\|\theta\|B+\beta B+R_{\max}H+1)\left|\left\|\Lambda^{1/2}\varphi(x,a)\right\|-\left\|(\Lambda')^{1/2}\varphi(x,a)\right\|\right|\\
&\le\alpha\beta(\|\theta\|B+\beta B+R_{\max}H+1)\left\|(\Lambda^{1/2}-(\Lambda')^{1/2})\varphi(x,a)\right\|\\
&\le\alpha\beta B(\|\theta\|B+\beta B+R_{\max}H+1)\left\|\Lambda^{1/2}-(\Lambda')^{1/2}\right\|
\end{aligned}$$

where the last inequality holds for $\|\varphi(x,a)\|\le B$. Finally, using the definition of the class $\mathcal{Q}$, we have that the matrices Λ and Λ' are positive definite in particular $\lambda_{\min}(\Lambda)\ge\frac{1}{2KBL_{\max}}$ and $\lambda_{\min}(\Lambda')\ge\frac{1}{2KBL_{\max}}$. Therefore by [Cassel and Rosenberg, 2024, Lemma 17], it holds that $\left\|\Lambda^{1/2}-(\Lambda')^{1/2}\right\|\le\frac{1}{2\sqrt{\lambda_{\min}}}\|\Lambda-\Lambda'\|=\sqrt{\frac{BKL_{\max}}{2}}\|\Lambda-\Lambda'\|$. Therefore, all in all we have that

$$\begin{aligned}
f(\theta,\Lambda;x,a)-f(\theta,\Lambda';x,a)&\le\sqrt{KL_{\max}}\alpha\beta B^{3/2}(\|\theta\|B+\beta B+R_{\max}H+1)\|\Lambda-\Lambda'\|\\
&\le\sqrt{KL_{\max}}\alpha\beta B^{3/2}(\|\theta\|B+\beta B+R_{\max}H+1)\|\Lambda-\Lambda'\|_F,
\end{aligned}$$

where $\|\cdot\|_F$ denote the Frobenious norm of a matrix. Hence, we have that

$$Q_{\theta,\Lambda}(x,a)-Q_{\theta',\Lambda'}(x,a)\le\sqrt{KL_{\max}}\alpha\beta B^{3/2}(\|\theta\|B+\beta B+R_{\max}H+1)\|\Lambda-\Lambda'\|_F+B\|\theta-\theta'\|.$$

At this point, if we have a ϵ_Λ-covering set for the set

$$\mathbf{\Lambda}=\left\{\Lambda\in\mathbb{R}^{d\times d}:\lambda_{\max}(\Lambda)\le 1,\quad\lambda_{\min}(\Lambda)\ge\frac{1}{2BKL_{\max}}\right\}$$

and an ϵ_θ-covering set for the set

$$\mathbf{\Theta}=\left\{\theta\in\mathbb{R}^d:\|\theta\|\le W_{\max}+Q_{\max}L_{\max}KB\right\}$$

we would have that

$$\begin{aligned}
|V(x)-V'(x)|&\le 2\sqrt{K^3L_{\max}}\alpha\beta B^{3/2}(\|\theta\|B+\beta B+R_{\max}H+1)\epsilon_F+2BK\epsilon_\theta\\
&\le 2\sqrt{K^3L_{\max}}\alpha\beta B^{3/2}(\|\theta\|B+\beta B+R_{\max}H+1)(\epsilon_F+\epsilon_\theta),
\end{aligned}$$

where in the last inequality we assumed that $\beta\ge 1$ and $B\ge 1$. Therefore, to have an ϵ-covering set for $\mathcal{V}$, we need to construct an ϵ_Λ-covering set for $\mathbf{\Lambda}$, where

$$\epsilon_\Lambda=\frac{\epsilon}{4\sqrt{K^3L_{\max}}\alpha\beta B^{3/2}(\|\theta\|B+\beta B+R_{\max}H+1)}$$

and an $\epsilon_\theta = \frac{\epsilon}{4\sqrt{K^3 L_{\max}}\alpha\beta B^{3/2}(\|\theta\|B+\beta B+R_{\max}H+1)}$*-covering set for* $\mathbf{\Theta}$*. Then, using the fact that the* ϵ*-covering number for the Euclidean ball of radius* R *in* d *dimension is given by* $(1+2R/\epsilon)^d$*, we obtain*

$$\log \mathcal{N}_{\epsilon_\theta}(\mathbf{\Theta}) \leq d \log\left(1 + 8\frac{(W_{\max}+Q_{\max}L_{\max}KB)\sqrt{K^3L_{\max}}\alpha\beta B^{3/2}(\|\theta\|\,B+\beta B+R_{\max}H+1)}{\epsilon}\right)$$

Moreover, noticing that for all matrices $\Lambda \in \mathbf{\Lambda}$ *it holds that* $\|\Lambda\|_F \leq \sqrt{d}\lambda_{\max}(\Lambda) \leq \sqrt{d}$*, we need to cover the Frobenious norm ball with radius* $\sqrt{d}$*. Recalling that the Frobenious norm of a matrix is equivalent to the euclidean norm of the vectorization of the matrix, this equivalent to cover the euclidean ball in* $\mathbb{R}^{d^2}$ *with radius* $\sqrt{d}$*.*

$$\log \mathcal{N}_{\epsilon_\Lambda}(\mathbf{\Lambda}) \leq d^2 \log\left(1 + 8\sqrt{d}\frac{\sqrt{K^3L_{\max}}\alpha\beta B^{3/2}(\|\theta\|\,B+\beta B+R_{\max}H+1)}{\epsilon}\right).$$

Therefore, using the fact that

$$\begin{aligned}
&\log \mathcal{N}_\epsilon(\mathcal{V}) = \log \mathcal{N}_{\epsilon_\Lambda}(\mathbf{\Lambda}) + \log \mathcal{N}_{\epsilon_\theta}(\mathbf{\Theta}) \\
&\leq d \log\left(1 + 8\frac{(W_{\max}+Q_{\max}L_{\max}KB)\sqrt{K^3L_{\max}}\alpha\beta B^{3/2}(\|\theta\|\,B+\beta B+R_{\max}H+1)}{\epsilon}\right) \\
&\quad + d^2 \log\left(1 + 8\sqrt{d}\frac{\sqrt{K^3L_{\max}}\alpha\beta B^{3/2}(\|\theta\|\,B+\beta B+R_{\max}H+1)}{\epsilon}\right) \\
&\leq 2d^2 \log\left(1 + 8\frac{(W_{\max}+Q_{\max}L_{\max}KB+\sqrt{d})\sqrt{K^3L_{\max}}\alpha\beta B^{3/2}(\|\theta\|\,B+\beta B+R_{\max}H+1)}{\epsilon}\right) \\
&\leq 2d^2 \log\left(1 + 8\frac{(W_{\max}B+Q_{\max}L_{\max}KB^2+\sqrt{d}+\beta B+R_{\max}H+1)^2\sqrt{K^3L_{\max}}\alpha\beta B^{3/2}}{\epsilon}\right) \\
&\leq 2d^2 \log\left(16\frac{(W_{\max}B+Q_{\max}L_{\max}KB^2+3\sqrt{d}+\beta B+R_{\max}H)^2\sqrt{K^3L_{\max}}\alpha\beta B^{3/2}}{\epsilon}\right) \\
&\leq 4d^2 \log\left(4\frac{(W_{\max}B+Q_{\max}L_{\max}KB^2+3\sqrt{d}+\beta B+R_{\max}H)\sqrt{K^3L_{\max}}\alpha\beta B^{3/2}}{\epsilon}\right) \\
&= 4d^2 \log\left(4(W_{\max}B+Q_{\max}L_{\max}KB^2+3\sqrt{d}+\beta B+R_{\max}H)\sqrt{K^5L_{\max}^3}\alpha\beta B^{5/2}\right).
\end{aligned}$$

where we used $d > 1$ *and the last step uses the fact that we are looking for a* $\epsilon = \frac{1}{KL_{\max}B}$ *covering set. Finally,*

$$\log \mathcal{N}_\epsilon \leq 4d^2 \log\left(4(W_{\max}B+Q_{\max}L_{\max}KB^2+3\sqrt{d}+\beta B+R_{\max}H)\sqrt{K^5L_{\max}^3}\alpha\beta B^{5/2}\right).$$

B.7 Putting everything together (proof of Theorem 1)

Theorem 3 *Run Algorithm 1 with parameters* $\omega = \log K$*,* $\alpha = 2\log K$*,*

$$\eta = \sqrt{\frac{5d\log\left(1+B^2T/d\right)\log|\mathcal{A}|}{8R_{\max}^2H^{5/2}K}}, \quad \text{and } \beta = CHR_{\max}d\log\left(BHW_{\max}R_{\max}dK\delta^{-1}\right),$$

for some absolute constant $C > 0$ *and* $\delta \in (0,1)$*. Then, with probability at least* $1-\delta$*, we have*

$$\begin{aligned}
\mathfrak{R}_K &= \tilde{\mathcal{O}}\left(\sqrt{d^3H^3K} + \sqrt{dH^{9/2}K\log\left(|\mathcal{A}|\right)}\right) \\
&= \tilde{\mathcal{O}}\left(\sqrt{d^3H^2T} + \sqrt{dH^{7/2}T\log\left(|\mathcal{A}|\right)}\right).
\end{aligned}$$

Proof 13 *We are now ready to prove Theorem 1. Combining Lemma 1, and the bounds in Equations (3) and Lemma 3 we first get*

$$\frac{1}{H}\mathfrak{R}_K \leq 2R_{\max}H\sum_{k=1}^{K}\left\langle \mu\left(\pi_k\right), p_k^+\right\rangle + 4Q_{\max}E\left(K\right) + \frac{E\left(K\right)\log\left|\mathcal{A}\right|}{\eta} + 2\eta Q_{\max}^2\sqrt{H}K + 2\sum_{k=1}^{K}\left\langle \mu\left(\pi_k\right), \left(1-p_k^+\right)\odot \mathrm{CB}_k\right\rangle .$$

Using the bound on the ascension functions provided in Inequality 5 and $1-p_k^+ \preceq 1$*, we further have*

$$\frac{1}{H}\mathfrak{R}_K \leq 4R_{\max}H\alpha^2\sum_{k=1}^{K}\left\langle \mu\left(\pi_k\right), \mathrm{CB}_k^2\right\rangle + 4R_{\max}e^{-\omega}HK + 4Q_{\max}E\left(K\right) + \frac{E\left(K\right)\log\left|\mathcal{A}\right|}{\eta} + 2\eta Q_{\max}^2\sqrt{H}K + 2\sum_{k=1}^{K}\left\langle \mu\left(\pi_k\right), \mathrm{CB}_k\right\rangle .$$

Lemma 5 can be used to bound the bonuses

$$\frac{1}{H}\mathfrak{R}_K \leq 32R_{\max}\alpha^2\beta^2B^2d\log\left(1+\frac{B^2T}{d}\right) + 16R_{\max}H\alpha^2\beta^2B^2\log\left(\frac{2K}{\delta}\right)^2 + 4R_{\max}e^{-\omega}HK + 4Q_{\max}E\left(K\right) + \frac{E\left(K\right)\log\left|\mathcal{A}\right|}{\eta} + 2\eta Q_{\max}^2\sqrt{H}K + \frac{8\beta B}{H}\sqrt{dT\log\left(1+\frac{B^2T}{d}\right)} + 8\beta B\log\left(\frac{2K}{\delta}\right)^2 .$$

Following Lemmas 4 and 2, we plug the values of $Q_{\max} = H\left(R_{\max}+\frac{2\omega}{\alpha}\right)$ *and* $L_{\max} = H\log\left(K/\delta\right)$*,*

$$\frac{1}{H}\mathfrak{R}_K \leq 32R_{\max}\alpha^2\beta^2B^2d\log\left(1+\frac{B^2T}{d}\right) + 16R_{\max}H\alpha^2\beta^2B^2\log\left(\frac{2K}{\delta}\right)^2 + 4R_{\max}e^{-\omega}HK + 4H\left(R_{\max}+\frac{2\omega}{\alpha}\right)E\left(K\right) + \frac{E\left(K\right)\log\left|\mathcal{A}\right|}{\eta} + 2\eta\left(R_{\max}+\frac{2\omega}{\alpha}\right)^2H^{5/2}K + \frac{8\beta B}{H}\sqrt{dT\log\left(1+\frac{B^2T}{d}\right)} + 8\beta B\log\left(\frac{2K}{\delta}\right)^2 .$$

By Lemma 9, we can bound $E\left(K\right) \leq 5d\log\left(1+\frac{B^2T}{d}\right)$

$$\frac{1}{H}\mathfrak{R}_K \leq 32R_{\max}\alpha^2\beta^2B^2d\log\left(1+\frac{B^2T}{d}\right) + 16R_{\max}\alpha^2\beta^2B^2H\log\left(\frac{2K}{\delta}\right)^2 + 4R_{\max}e^{-\omega}HK + 20Hd\left(R_{\max}+\frac{2\omega}{\alpha}\right)\log\left(1+\frac{B^2T}{d}\right) + \frac{5d}{\eta}\log\left(1+\frac{B^2T}{d}\right)\log\left|\mathcal{A}\right| + 2\eta\left(R_{\max}+\frac{2\omega}{\alpha}\right)^2H^{5/2}K + \frac{8\beta B}{H}\sqrt{dT\log\left(1+\frac{B^2T}{d}\right)} + 8\beta B\log\left(\frac{2K}{\delta}\right)^2 .$$

It remains to choose the parameters. We start by setting $\alpha = 2\omega$ and use $R_{\max} \geq 1$ to get

$$\begin{aligned}\frac{1}{H}\mathfrak{R}_K \leq\ & 128R_{\max}\omega^2\beta^2B^2d\log\left(1+\frac{B^2T}{d}\right)+64R_{\max}\omega^2\beta^2B^2H\log\left(\frac{2K}{\delta}\right)^2\\ &+4R_{\max}e^{-\omega}HK+40HdR_{\max}\log\left(1+\frac{B^2T}{d}\right)\\ &+\frac{5d}{\eta}\log\left(1+\frac{B^2T}{d}\right)\log|\mathcal{A}|+8\eta R_{\max}^2H^{5/2}K\\ &+\frac{8\beta B}{H}\sqrt{dT\log\left(1+\frac{B^2T}{d}\right)}+8\beta B\log\left(\frac{2K}{\delta}\right)^2.\end{aligned}$$

Then, we set $\omega = \log K$

$$\begin{aligned}\frac{1}{H}\mathfrak{R}_K \leq\ & 128R_{\max}\beta^2B^2d\log(K)^2\log\left(1+\frac{B^2T}{d}\right)+64R_{\max}\beta^2B^2H\log(K)^2\log\left(\frac{2K}{\delta}\right)^2\\ &+4R_{\max}H+40HdR_{\max}\log\left(1+\frac{B^2T}{d}\right)\\ &+\frac{5d}{\eta}\log\left(1+\frac{B^2T}{d}\right)\log|\mathcal{A}|+8\eta R_{\max}^2H^{5/2}K\\ &+\frac{8\beta B}{H}\sqrt{dT\log\left(1+\frac{B^2T}{d}\right)}+8\beta B\log\left(\frac{2K}{\delta}\right)^2.\end{aligned}$$

We choose the learning rate as $\eta = \sqrt{\frac{5d\log(1+B^2T/d)\log|\mathcal{A}|}{8R_{\max}^2H^{5/2}K}}$ and we obtain

$$\begin{aligned}\frac{1}{H}\mathfrak{R}_K \leq\ & 128R_{\max}\beta^2B^2d\log(K)^2\log\left(1+\frac{B^2T}{d}\right)+64R_{\max}\beta^2B^2H\log(K)^2\log\left(\frac{2K}{\delta}\right)^2\\ &+4R_{\max}H+40HdR_{\max}\log\left(1+\frac{B^2T}{d}\right)\\ &+4\sqrt{10R_{\max}^2H^{5/2}d\log\left(1+\frac{B^2T}{d}\right)\log(|\mathcal{A}|)K}\\ &+\frac{8\beta B}{H}\sqrt{dT\log\left(1+\frac{B^2T}{d}\right)}+8\beta B\log\left(\frac{2K}{\delta}\right)^2.\end{aligned}$$

Finally, following Lemma 6 we set $\beta = CHR_{\max}d\log\left(BHW_{\max}R_{\max}dK\delta^{-1}\right)$ where $C > 0$ is an absolute constant and we get

$$\begin{aligned}\frac{1}{H}\mathfrak{R}_K \leq\ & 128C^2R_{\max}^3B^2d^3H^2\log(K)^2\log\left(1+\frac{B^2T}{d}\right)\log\left(BHW_{\max}R_{\max}dK\delta^{-1}\right)^2\\ &+64C^2R_{\max}^3d^2B^2H^3\log(K)^2\log\left(\frac{2K}{\delta}\right)^2\log\left(BHW_{\max}R_{\max}dK\delta^{-1}\right)^2\\ &+4R_{\max}H+40HdR_{\max}\log\left(1+\frac{B^2T}{d}\right)\\ &+4\sqrt{10R_{\max}^2H^{5/2}d\log\left(1+\frac{B^2T}{d}\right)\log(|\mathcal{A}|)K}\\ &+8CR_{\max}Bd\sqrt{dT\log\left(1+\frac{B^2T}{d}\right)}\log\left(BHW_{\max}R_{\max}dK\delta^{-1}\right)\\ &+8CR_{\max}dBH^2\log\left(\frac{2K}{\delta}\right)^2\log\left(BHW_{\max}R_{\max}dK\delta^{-1}\right).\end{aligned}$$

After multiplying by H, we get

$$\begin{aligned}\mathfrak{R}_K &= \tilde{\mathcal{O}}\left(\sqrt{d^3H^3K} + \sqrt{dH^{9/2}K\log\left(|\mathcal{A}|\right)}\right) \\ &= \tilde{\mathcal{O}}\left(\sqrt{d^3H^2T} + \sqrt{dH^{7/2}T\log\left(|\mathcal{A}|\right)}\right) .\end{aligned}$$

C Motivation for *Learning from Features Alone* and related works in imitation learning

Related works in theoretical imitation learning. A special case of our setting is imitation learning from state-only expert trajectories, which is recovered when $\varphi_r(x, a) = \mathbf{e}_x$. This setting was first studied in Sun et al. [2019] in the finite-horizon setting with general function approximation. There are some notable differences between their work and ours, primarily that they focus on the finite-horizon setting and learn a non-stationary policy. In principle, their algorithm could be applied to the infinite-horizon setting by truncating the trajectories after $\tilde{\mathcal{O}}(1-\gamma)^{-1}$ steps. However, this would still result in a non-stationary policy, whereas our approach outputs a stationary policy. Their realizability assumption on the expert policy and expert state-value function is not required in our work which leverages, instead, the linear MDP assumption. These assumptions are not directly comparable, even when the function classes in Sun et al. [2019] are assumed to be linear. Indeed, the realizability assumption imposed in Sun et al. [2019] would imply having access to the values of the features $\sum_a \pi_E(a|x)\varphi(x, a)$ for each state $x \in \mathcal{X}$. In contrast, our approach does not require this additional knowledge about the expert.

Furthermore, the guarantees on the number of expert trajectories in [Sun et al., 2019, Theorem 3.3] adapted to the infinite-horizon setting, would scale as $\tilde{\mathcal{O}}((1-\gamma)^{-4}\epsilon^{-2})$ whereas we only require $\tilde{\mathcal{O}}((1-\gamma)^{-2}\epsilon^{-2})$ state-only samples from the expert occupancy measure.

Similarly, Arora et al. [2020] develop a framework for imitation and representation learning from observation alone based on bilevel optimization but assume the realizability of the state-value function, which is not needed in our work.

The work of Kidambi et al. [2021] investigates the idea of exploration in state-only imitation learning. Unlike our work, they focus on the finite-horizon setting and on different structural assumptions on the MDP. Specifically, Kidambi et al. [2021] consider tabular MDPs, nonlinear kernel regulators, and MDPs with Gaussian transition kernels and bounded Eluder dimension, whereas our work focuses on infinite-horizon linear MDPs and observing only the feature directions visited by the expert, which is a weaker requirement than observing the states directly. Moreover, our algorithm `FRA-IL` is computationally efficient, whereas the model fitting step in Kidambi et al. [2021] cannot be implemented efficiently for various situations, including linear MDPs [Jin et al., 2019] and KNRs [Kakade et al., 2020].

Wu et al. [2024a] operate under a different set of assumptions, namely that the learner has access to a function class for the expert's score function and that the expected state norm remains bounded during learning. Under this setting, the authors are the first to achieve first- and second-order bounds for imitation learning, which lead to a faster rate in the case of low-variance expert policies and transitions. The authors do not quantify the MDP trajectory complexity, but it would scale suboptimally with $1/\epsilon$ because they require an expensive *RL in the loop* routine that we avoid in our work.

Xu et al. [2022] develop an analysis for horizon-free bounds on τ_E for a special class of MDPs, where expert states can be visited only by visiting all preceding expert states.

The trajectory access to the MDP $\mathcal{M} \setminus r_{\text{true}}$ assumed in this work should not be confused with interactive/online imitation learning, where the expert can be queried during learning [Ross and Bagnell, 2010, Ross et al., 2011, Swamy et al., 2021, Li and Zhang, 2022, Lavington et al., 2022, Sekhari et al., 2024b, Sun et al., 2017, Sekhari et al., 2024a]. Furthermore, our trajectory access is a much weaker requirement compared to generative model access used in [Swamy et al., 2022, Kamoutsi et al., 2021].

Moreover, it is important to note that we do not require any ergodicity or self-exploration properties of the dynamics, whereas such assumptions are needed in [Viano et al., 2022, Zeng et al., 2022b]. Additionally, uniformly good evaluation error, which is essentially possible only under generative model or ergodic dynamics assumptions, is required in [Wu et al., 2023, Zeng et al., 2022a, 2023]. Also, the use of exploration bonuses in imitation learning has also been useful for the related problem of finding the reward feasible set without using a generative model [Lazzati et al., 2024, Lindner et al., 2022].

Finally, we present Table 1, which compares our bounds with existing ones. We show the number of expert trajectories and MDP interactions required for ϵ-suboptimal expected performance. The

Table 1: Comparison with related imitation learning algorithms.

Algorithm	**Setting**	**F.O.**	**Expert Traj.** (τ_E)	**MDP Traj.** (K)
Behavioural Cloning	Function Approximation, Episodic Foster et al. [2024]	✗	$\mathcal{O}\left(\frac{H^2 \log\lvert\Pi\rvert}{\epsilon^2}\right)$	-
	Tabular, Episodic Rajaraman et al. [2020]	✗	$\tilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{X}\rvert}{\epsilon}\right)$	-
	Deterministic Linear Expert, Episodic Rajaraman et al. [2021]	✗	$\tilde{\mathcal{O}}\left(\frac{H^2 d}{\epsilon}\right)$	-
Mimic-MD Rajaraman et al. [2020]	Tabular, Known P, Deterministic Expert, Episodic	✗	$\mathcal{O}\left(\frac{H^{3/2}\lvert\mathcal{X}\rvert}{\epsilon}\right)$	-
OAL Shani et al. [2021]	Episodic Tabular	✗	$\tilde{\mathcal{O}}\left(\frac{H^2\lvert\mathcal{X}\rvert}{\epsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{H^4\lvert\mathcal{X}\rvert^2\lvert\mathcal{A}\rvert}{\epsilon^2}\right)$
MB-TAIL Xu et al. [2023]	Episodic, Tabular, Deterministic Expert	✗	$\mathcal{O}\left(\frac{H^{3/2}\lvert\mathcal{X}\rvert}{\epsilon}\right)$	$\mathcal{O}\left(\frac{H^3\lvert\mathcal{X}\rvert^2\lvert\mathcal{A}\rvert}{\epsilon^2}\right)$
FAIL Sun et al. [2019]	Episodic, $\pi_E \in \Pi$ and $V^{\pi_E} \in \mathcal{F}$	✓⋆	$\tilde{\mathcal{O}}\left(\frac{H^4 \log(\lvert\Pi\rvert\lvert\mathcal{F}\rvert H)}{\epsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{H^4 \log(\lvert\Pi\rvert\lvert\mathcal{F}\rvert H)}{\epsilon^2}\right)$
Mobile Kidambi et al. [2021]	Episodic, $r_{\text{true}} \in \mathcal{R}$ and $P \in \mathcal{P}$	✓⋆	$\tilde{\mathcal{O}}\left(\frac{H^2 \log(\lvert\mathcal{R}\rvert H)}{\epsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{H^5 \log\lvert\mathcal{P}\rvert}{\epsilon^2}\right)$
OGAIL Liu et al. [2022]	Episodic Linear Mixture MDP, $W_{\max} = \sqrt{d}$	✓	$\tilde{\mathcal{O}}\left(\frac{H^3 d^2}{\epsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{H^4 d^3}{\epsilon^2}\right)$
ILARL Viano et al. [2024]	Linear MDP, $W_{\max} = 1$	✓	$\tilde{\mathcal{O}}\left(\frac{d}{(1-\gamma)^2\epsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{d^3}{(1-\gamma)^8\epsilon^4}\right)$
`FRA-IL` (This Work)	Linear MDP	✓	$\tilde{\mathcal{O}}\left(\frac{W_{\max}^2}{(1-\gamma)^2\epsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{d^3}{(1-\gamma)^{4.5}\epsilon^2}\right)$
Lower Bound (This Work)	Linear MDP	✓	$\Omega\left(\frac{W_{\max}^2}{(1-\gamma)^2\epsilon^2}\right)$	$\Omega\left(\frac{d}{(1-\gamma)^2\epsilon^2}\right)$

acronym F.O. refers to "Features Only" and indicates whether the algorithm applies to the setting we consider here. The star ✓⋆ specifies that the algorithm only applies to state-only imitation learning. "Linear expert" refers to the case where an expert policy is of the form

$$\pi(x) = \max_{a \in \mathcal{A}} \varphi(x, a)^\intercal \theta\,,$$

for some vector θ. Finally, in the work by Kidambi et al. [2021], the bound on K can be tighter than what we report in the table. We report this slightly looser version for sake of simplicity and avoiding to introduce the information gain (see Kidambi et al. [2021] for details).

D Omitted proofs for Section 5

To improve readability, we define the feature expectation vector as $\lambda(\pi) = \Phi_r^\intercal \mu(\pi)$ for any policy π, where $\mu(\pi_k)$ denotes the occupancy measure of policy π_k. This notation will be used in the following proofs.

D.1 Proof of Theorem 2 (guarantee for the output of Algorithm 2)

Theorem 2 *Algorithm 2, when run for $K = \tilde{\mathcal{O}}\left(d^3 H^{9/2} \varepsilon^{-2} \log(|\mathcal{A}|)\right)$ iterations with an expert dataset of size $\tau_E = \widetilde{\mathcal{O}}\left(W_{\max}^2 H^2 \varepsilon^{-2}\right)$, outputs an ε-suboptimal policy.*

Proof 14 *Using the decomposition presented in Section 5, we can express the regret as*

$$(1-\gamma)\,\mathfrak{R}_{\mathrm{K}}^{\mathrm{IL}} = \underbrace{\sum_{k=1}^{K} \langle r_k, \mu\left(\pi_E\right) - \mu\left(\pi_k\right)\rangle}_{(1-\gamma)\mathfrak{R}_K^{\pi}(\mu(\pi_E))} + \underbrace{\sum_{k=1}^{K} \langle \Phi_r^\intercal \mu(\pi_k) - \Phi_r^\intercal \mu(\pi_E), w_k - w_{\mathrm{true}}\rangle}_{(1-\gamma)\mathfrak{R}_K^{w}(w_{\mathrm{true}})} .$$

By bounding $\mathfrak{R}_T^w(w_{\mathrm{true}})$ and $\mathfrak{R}_T^\pi(\mu(\pi_E))$ using Theorem 4 and Theorem 1, respectively, we obtain that with probability $1-3\delta$

$$\begin{aligned}\frac{1}{K}\,\mathfrak{R}_{\mathrm{K}}^{\mathrm{IL}} \leq\; & 10H(BW_{\max} + R_{\max})\sqrt{\frac{\log \delta^{-1}}{K}} + 24HW_{\max}B\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\tau_E}} \\ & + \tilde{\mathcal{O}}(d^{3/2}(1-\gamma)^{-9/4}\log^{1/2}|\mathcal{A}|\,K^{-1/2})\,.\end{aligned}$$

Therefore, by considering B and $R_{\max}$ as constants and choosing $K = \widetilde{\mathcal{O}}\left(\frac{d^3 \log(|\mathcal{A}|\delta^{-1})}{(1-\gamma)^{4.5}\varepsilon^2}\right)$ and $\tau_E = \widetilde{\mathcal{O}}\left(\frac{W_{\max}^2 \log(1/\delta)}{(1-\gamma)^2\varepsilon^2}\right)$ we have that with probability $1-3\delta$ it holds that

$$\frac{1}{K}\,\mathfrak{R}_{\mathrm{K}}^{\mathrm{IL}} \leq 4\epsilon\,.$$

Since $\frac{1}{K}\,\mathfrak{R}_{\mathrm{K}}^{\mathrm{IL}}$ is a random variable bounded by $(1-\gamma)^{-1}$ almost surely, in expectation we have the following bound

$$\mathbb{E}_{\mathrm{Alg}}\left[\frac{1}{K}\,\mathfrak{R}_{\mathrm{K}}^{\mathrm{IL}}\right] \leq \frac{3\delta}{1-\gamma} + 4\varepsilon\,.$$

Thus, by choosing $\delta \leq \varepsilon/3(1-\gamma)$ we can conclude that

$$\mathbb{E}_{\mathrm{Alg}}\left[\frac{1}{K}\,\mathfrak{R}_{\mathrm{K}}^{\mathrm{IL}}\right] \leq 5\varepsilon\,.$$

Finally, by selecting π^{out} uniformly at random from the policies generated by Algorithm 2 we have that

$$\mathbb{E}_{\mathrm{Alg}}\left\langle \nu_0, V_{r_{\mathrm{true}}}^{\pi_E} - V_{r_{\mathrm{true}}}^{\pi^{\mathrm{out}}}\right\rangle \leq 5\varepsilon\,.$$

D.2 Proof of Theorem 4 (regret bound for the reward player)

Theorem 4 *Assume that $w_{\mathrm{true}} \in \mathcal{W}$ for some non-empty closed convex set $\mathcal{W}$ and that for any $w \in \mathcal{W}$, $\|w\| \leq W_{\max}$. Then, OGD with $\eta_r = W_{\max}/B\sqrt{K}$ ran for K iterations satisfies with probability at least $1-2\delta$ that*

$$\mathfrak{R}_K^w(w_{\mathrm{true}}) \leq 10H(BW_{\max} + R_{\max})\sqrt{K\log 1/\delta} + 24HW_{\max}KB\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\tau_E}}\,.$$

Proof 15 *Given the definition of the feature expectation vector $\lambda(\pi)$, we can rewrite the regret for the reward player as follows*

$$(1-\gamma)\,\mathfrak{R}_T^w(w_{\mathrm{true}}) = \sum_{k=1}^{K} \langle \lambda(\pi_k) - \lambda(\pi_E), w_k - w_{\mathrm{true}}\rangle\,.$$

Then, adding and subtracting the estimators for the occupancy measures, we get

$$\begin{aligned}(1-\gamma)\mathfrak{R}_T^w(w_{\text{true}}) = &\sum_{k=1}^K \left\langle \varphi_r(X_k,A_k) - \widehat{\lambda(\pi_E)}, w_k - w_{\text{true}} \right\rangle \\ &+ \sum_{k=1}^K \langle \lambda(\pi_k) - \varphi_r(X_k,A_k), w_k - w_{\text{true}} \rangle \\ &+ \sum_{k=1}^K \left\langle \widehat{\lambda(\pi_E)} - \lambda(\pi_E), w_k - w_{\text{true}} \right\rangle .\end{aligned}$$

Now, using the regret bound for OGD [Zinkevich, 2003], we can bound the first term in the decomposition above as

$$\begin{aligned}&\sum_{k=1}^K \left\langle \varphi_r(X_k,A_k) - \widehat{\lambda(\pi_E)}, w_k - w_{\text{true}} \right\rangle \\ &\quad\le \frac{\max_{w\in\mathcal{W}} \|w_{\text{true}} - w_1\|_2^2}{2\eta_r} + \frac{\eta_r}{2}\sum_{k=1}^K \left\| \widehat{\lambda(\pi_E)} - \varphi_r(X_k,A_k) \right\|_2^2 \\ &\quad\le \frac{2W_{\max}^2}{\eta_r} + 2\eta_r B^2 K ,\end{aligned}$$

Looking at the term $\sum_{k=1}^K \langle \lambda(\pi_k) - \varphi_r(X_k,A_k), w_k - w_{\text{true}} \rangle$*, we notice that*

$$\psi_k = \langle \lambda(\pi_k) - \varphi_r(X_k,A_k), w_k - w_{\text{true}} \rangle$$

is a martingale difference sequence such that

$$|\langle \lambda(\pi_k) - \varphi_r(X_k,A_k), w_k - w_{\text{true}} \rangle| \le 4R_{\max} .$$

Applying Azuma-Hoeffding's inequality, we have that with probability $1-\delta$

$$\sum_{k=1}^K \langle \lambda(\pi_k) - \varphi_r(X_k,A_k), w_k - w_{\text{true}} \rangle \le R_{\max}\sqrt{8K\log\left(\frac{1}{\delta}\right)}.$$

Then, plugging in this bound in the regret decomposition we obtain

$$\begin{aligned}(1-\gamma)\mathfrak{R}_T^w(w_{\text{true}}) \le &\frac{2W_{\max}^2}{\eta_r} + 2\eta_r B^2 K + R_{\max}\sqrt{8K\log 1/\delta} \\ &+ \sum_{k=1}^K \left\langle \widehat{\lambda(\pi_E)} - \lambda(\pi_E), w_k - w_{\text{true}} \right\rangle .\end{aligned}$$

Then, we treat the last term using Cauchy-Schwartz's inequality

$$\begin{aligned}\sum_{k=1}^K \left\langle \widehat{\lambda(\pi_E)} - \lambda(\pi_E), w_k - w_{\text{true}} \right\rangle &\le \sum_{k=1}^K \left\| w_{\text{true}} - w^k \right\|_2 \left\| \widehat{\lambda(\pi_E)} - \lambda(\pi_E) \right\|_2 \\ &\le 2W_{\max} K \left\| \widehat{\lambda(\pi_E)} - \lambda(\pi_E) \right\|_2 .\end{aligned}$$

It remains to find a high probability (dimension-free) upper bound on $\left\| \widehat{\lambda(\pi_E)} - \lambda(\pi_E) \right\|_2$*. First, notice that* $\left\| \widehat{\lambda(\pi_E)} - \lambda(\pi_E) \right\|_2 = \left\| (\tau_E)^{-1} \left(\sum_{i=1}^{\tau_E} \varphi_r(X_E^i, A_E^i) - \lambda(\pi_E) \right) \right\|_2$*. Then, we use the notation* $u_{x,a} = \varphi_r(x,a) - \lambda(\pi_E)$ *for all state action pairs* x, a *and using that for all* $x, a \in \mathcal{X}\times\mathcal{A}$*,* $\|\varphi_r(x,a)\|_2 \le B$*, we have*

$$\sum_{i=1}^{\tau_E} \mathbb{E}\left[\left\| u_{X_E^i, A_E^i} \right\|_2^2 \right] \le \sum_{i=1}^{\tau_E} \mathbb{E}\left[\left\| \varphi_r(X_E^i, A_E^i) - \lambda(\pi_E) \right\|_2^2 \right] \le 4\tau_E B^2 .$$

Moreover, for any $x, a \in \mathcal{X} \times \mathcal{A}$, $\|u_{x,a}\| \leq 2B$ and $\mathbb{E}\left[u_{X_E^i, A_E^i}\right] = 0$ because of the distribution of the dataset $\mathcal{D}_{\pi_E}$. Thus, by applying [Hsu et al., 2012, Proposition 2], it holds that for all $t > 0$

$$\mathbb{P}\left[\left\|\sum_{i=1}^{\tau_E} u_{X_E^i, A_E^i}\right\| > \sqrt{4\tau_E}B + \sqrt{32\tau_E t}B + (8/3)2Bt\right] \leq e^{-t}$$

Therefore, choosing $t = \log\frac{1}{\delta}$, we obtain that with probability $1-\delta$

$$\begin{aligned}\left\|\sum_{i=1}^{\tau_E} \varphi_r(X_E^i, A_E^i) - \lambda(\pi_E)\right\| &\leq \sqrt{4\tau_E}B + \sqrt{32\tau_E \log\left(\frac{1}{\delta}\right)}B + \frac{16B}{3}\log\left(\frac{1}{\delta}\right) \\ &\leq 6B\sqrt{\tau_E \log\left(\frac{1}{\delta}\right)} + \frac{16B}{3}\log\left(\frac{1}{\delta}\right).\end{aligned}$$

Then, dividing by τ_E we obtain that

$$\left\|\widehat{\lambda(\pi_E)} - \lambda(\pi_E)\right\|_2 \leq 6B\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\tau_E}} + \frac{16B}{3\tau_E}\log\left(\frac{1}{\delta}\right).$$

Then, for $\tau_E \geq \frac{64}{18^2}\log\frac{1}{\delta}$, we have that

$$6\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\tau_E}} \geq \frac{8}{3\tau_E}\log\left(\frac{1}{\delta}\right),$$

and hence that with probability $1-\delta$,

$$\left\|\widehat{\lambda(\pi_E)} - \lambda(\pi_E)\right\|_2 \leq 12B\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\tau_E}}.$$

Thus, by a union bound and choosing $\eta_r = {}^{W_{\max}}/_{B\sqrt{K}}$, we have that with probability $1-2\delta$,

$$\begin{aligned}(1-\gamma)\mathfrak{R}_T^w(w_{\text{true}}) &= 4BW_{\max}\sqrt{K} + R_{\max}\sqrt{8K\log\delta^{-1}} + 24W_{\max}BK\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\tau_E}} \\ &\leq 10(BW_{\max} + R_{\max})\sqrt{K\log\delta^{-1}} + 24W_{\max}KB\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\tau_E}}.\end{aligned}$$

E Lower bounds for imitation learning

In this section, we prove lower bounds for both K and τ_E for all algorithms following Protocol 1 given hereafter.

Protocol 1 Imitation learning from features alone in Linear MDPs.

1: The learner adopts a learning algorithm Alg that receives as input
(1) a features dataset $\mathcal{D}_{\pi_E} = \left\{\varphi_r\left(X_E^i, A_E^i\right)\right\}_{i=1}^{\tau_E}$ where for any $i \in [\tau_E]$, $X_E^i, A_E^i \sim \mu\left(\pi_E\right)$,
(2) read access to $\varphi_P\left(x, a\right)$ for all $x, a \in \mathcal{X} \times \mathcal{A}$,
(3) trajectory access to $\mathcal{M} \setminus r_{\text{true}}$, and
(4) the reward class $\mathcal{R}$ such that $r_{\text{true}} \in \mathcal{R}$.
2: Alg samples K trajectories from $\mathcal{M} \setminus r_{\text{true}}$ and outputs π^{out} s.t. $\mathbb{E}\left[\left\langle \nu_0, V_{r_{\text{true}}}^{\pi_E} - V_{r_{\text{true}}}^{\pi^{\text{out}}} \right\rangle\right] \leq \varepsilon$.

We prove an $\Omega\left(\varepsilon^{-2}\right)$ lower bound for both cases, demonstrating that Algorithm 2 is rate optimal. First, we state the lower bound K that holds even with perfect knowledge of the expert feature expectation vector $\lambda\left(\pi_E\right)$, a strictly easier setting compared the one under which Theorem 2 is proven.

Theorem 5 ***(Lower Bound on** K**)*** *For any algorithm* Alg*, there exists an MDP $\mathcal{M}$ and an expert policy π_E such that* Alg*, taking as input $\Phi_r^\intercal \mu_{\mathcal{M}}\left(\pi_E\right)$, requires $K = \Omega\left(\frac{d}{(1-\gamma)^2\varepsilon^2}\right)$ to guarantee $\mathbb{E}_{\text{Alg}}\left[\left\langle \nu_0, V_{\mathcal{M}}^{\pi_E} - V_{\mathcal{M}}^{\pi^{\text{out}}} \right\rangle\right] = \mathcal{O}\left(\varepsilon\right)$.*

Next, we establish a lower bound on the required number of expert demonstration τ_E. The result holds even with perfect knowledge of the transition dynamics (*i.e.* for $K = \infty$).

Theorem 6 ***(Lower Bound on** τ_E**)*** *Let $\gamma \geq \frac{1}{2}$. For any algorithm* Alg*, there exists an MDP $\mathcal{M}$ and an expert policy π_E such that* Alg *taking as input the transitions dynamics and an expert dataset of size τ_E requires $\tau_E = \Omega\left(\frac{W_{\max}^2}{(1-\gamma)^2\varepsilon^2}\right)$ to guarantee $\mathbb{E}_{\text{Alg}}\left[\left\langle \nu_0, V_{\mathcal{M}}^{\pi_E} - V_{\mathcal{M}}^{\pi^{\text{out}}} \right\rangle\right] = \mathcal{O}\left(\varepsilon\right)$.*

The proofs are provided in the following sections.

E.1 Proof of Theorem 5 (lower bound on the number of interactions)

We start with the proof of the lower bound on K. We consider a class of possibly randomized algorithms that output a policy π^{out} given a dataset of expert features $\mathcal{D}_{\pi_E}$ and K trajectories collected by the learner in the MDP $\mathcal{M}$.

Proof Idea. To construct a lower bound, we consider the case of imitation learning from states alone (*i.e.* $\Phi_r\left(x, a\right) = \mathbf{e}_x$), and $\lambda\left(\pi_E\right)$ represents the state expert occupancy measure. We consider the case of a two-state MDP, where $\mathcal{X} = \{x_0, x_1\}$, and the learner knows the *good* state x_0 that maximizes the expert's occupancy measure due to having access to $\lambda\left(\pi_E\right)$. The leaner's objective is to maximize the time spent in this good state. All actions in the *bad* state x_1 share the same transition kernel. Therefore, the agent's decisions in the state x_0 is the only factor that influences the outcome. An action labeled as $a^\star$ is available in the state x_0. The transition kernel $P\left(x_0 \mid x_0, a\right)$ is identical for all actions $a \neq a^\star$, while for $a^\star$, it is defined as $P\left(x_0 \mid x_0, a\right) + \epsilon$. We then consider a family of $|\mathcal{A}|$ MDPs, where each MDP assigns the role of $a^\star$ to a different action. We will formally demonstrate that for any algorithm in Alg, there exists at least one MDP within this family where achieving $\mathbb{E}_{\text{Alg}}\left[\left\langle \nu_0, V_{\mathcal{M}}^{\pi_E} - V_{\mathcal{M}}^{\pi^{\text{out}}} \right\rangle\right] = \mathcal{O}\left(\varepsilon\right)$ requires $K = \Omega\left(\frac{|\mathcal{A}|}{(1-\gamma)^2\varepsilon^2}\right)$. Finally, the bound for an arbitrary dimension d is obtained noticing that this MDP can be written as a linear MDP with features dimension $d = 2 + 2\left|\mathcal{A}\right|$.

Proof 16 *For any policy π, we denote $\lambda_{\mathcal{M}}\left(\pi\right) = \Phi_r^\intercal \mu_{\mathcal{M}}\left(\pi\right)$ the expected feature vector of the policy π in the MDP $\mathcal{M}$. We consider a deterministic algorithm* Alg *that maps $\lambda_{\mathcal{M}}\left(\pi_E\right)$ and K environment trajectories to a policy. The extension to randomized algorithms can be done by an*

application of Fubini's theorem (see Bubeck et al. [2012]). The hard instance we consider for the lower bound is an MDP $\mathcal{M}$ with two states, x_0 and x_1, and $|\mathcal{A}|$ actions per state. For any action a, the reward function is given by $r_{\text{true}}(x_0, a) = 1$, and $r_{\text{true}}(x_1, a) = 0$. We will refer to state x_0 as the "good" state and to state x_1 as the "bad" state. In state x_1, the transition kernel induced by any action a is the same, i.e. $P(x_1 \mid x_1, a) = 1 - \delta_1$, and $P(x_0 \mid x_1, a) = \delta_1$ for some $\delta_1 \in (0, 1)$. Let $\delta_0 \in (0, 1)$ and $\epsilon \in (0, \delta_0)$. In state x_0, there is an action $a^\star$ with a slightly different transition kernel

$$P(x_1 \mid x_0, a^\star) = \delta_0 - \epsilon, \quad P(x_0 \mid x_0, a^\star) = 1 - \delta_0 + \epsilon,$$

whereas for any action $a \neq a^\star$, we set

$$P(x_1 \mid x_0, a) = \delta_0, \quad P(x_0 \mid x_0, a) = 1 - \delta_0.$$

We set the unknown expert policy π_E such that it always select action $a^\star$ in both states, i.e. $\pi_E(a^\star \mid x_0) = \pi_E(a^\star \mid x_1) = 1$. Setting $\nu_0(x_0) = 1$, we can write the flow constraints and get

$$\begin{aligned}
\nu(\pi_E, x_0) &= 1 - \gamma + \gamma(1 - \delta_0 + \epsilon)\nu(\pi_E, x_0) + \gamma\delta_1\nu(\pi_E, x_1), \\
\nu(\pi_E, x_1) &= \gamma(1 - \delta_1)\nu(\pi_E, x_1) + \gamma(\delta_0 - \epsilon)\nu(\pi_E, x_0).
\end{aligned}$$

The second equation gives $\nu(\pi_E, x_1) = \frac{\gamma(\delta_0 - \epsilon)}{1 - \gamma(1 - \delta_1)}\nu(\pi_E, x_0)$, which we can plug back into the first equation to obtain

$$\nu(\pi_E, x_0) = 1 - \gamma + \left(\gamma(1 - \delta_0 + \epsilon) + \frac{\gamma^2\delta_1(\delta_0 - \epsilon)}{1 - \gamma(1 - \delta_1)}\right)\nu(\pi_E, x_0),$$

which we can rearrange to get

$$\nu(\pi_E, x_0) = \frac{1 - \gamma + \gamma\delta_1}{1 - \gamma + \gamma\delta_1 + \gamma\delta_0 - \gamma\epsilon}.$$

Using the normalization constraint $\nu(\pi_E, x_0) + \nu(\pi_E, x_1) = 1$, we also get

$$\nu(\pi_E, x_1) = \frac{\gamma\delta_0 - \gamma\epsilon}{1 - \gamma + \gamma\delta_1 + \gamma\delta_0 - \gamma\epsilon}.$$

Furthermore, let π_{bad} be a "bad" policy that always plays an action $a \neq a^\star$. The same calculation with $\epsilon = 0$ shows that the state occupancy measure for the policy π_{bad} is given by

$$\begin{aligned}
\nu(\pi_{\text{bad}}, x_0) &= \frac{1 - \gamma + \gamma\delta_1}{1 - \gamma + \gamma\delta_1 + \gamma\delta_0}, \\
\nu(\pi_{\text{bad}}, x_1) &= \frac{\gamma\delta_0}{1 - \gamma + \gamma\delta_1 + \gamma\delta_0}.
\end{aligned}$$

Let $\tilde{\pi}$ be any policy. Noting that for any x, $V^{\pi_E}(x) = Q^{\pi_E}(x, a^\star)$, we can use the performance difference lemma and get

$$\begin{aligned}
\langle \mu(\pi_E) - \mu(\tilde{\pi}), r_{\text{true}} \rangle &= \mathbb{E}_{(x,a)\sim\mu(\tilde{\pi})}\left[V^{\pi_E}(x) - Q^{\pi_E}(x, a)\right] \\
&= \mathbb{E}_{(x,a)\sim\mu(\tilde{\pi})}\left[Q^{\pi_E}(x, a^\star) - Q^{\pi_E}(x, a)\right].
\end{aligned}$$

All actions share the same transition kernel in x_1 thus for any action a, $Q^{\pi_E}(x_1, a^\star) = Q^{\pi_E}(x_1, a)$ and we have

$$\langle \mu(\pi_E) - \mu(\tilde{\pi}), r_{\text{true}} \rangle = \nu(\tilde{\pi}, x_0) \sum_{a \in \mathcal{A}\setminus\{a^\star\}} \tilde{\pi}(a \mid x_0)\left(Q^{\pi_E}(x_0, a^\star) - Q^{\pi_E}(x_0, a)\right).$$

Next, we need to compute the difference of Q-values. Using the Bellman equations for π_E in state x_0, we have

$$\forall a \neq a^\star, Q^{\pi_E}(x_0, a) = 1 + \gamma\delta_0 Q^{\pi_E}(x_1, a^\star) + \gamma(1 - \delta_0)Q^{\pi_E}(x_0, a^\star) \tag{8}$$

$$Q^{\pi_E}(x_0, a^\star) = 1 + \gamma(\delta_0 - \epsilon)Q^{\pi_E}(x_1, a^\star) + \gamma(1 - \delta_0 + \epsilon)Q^{\pi_E}(x_0, a^\star). \tag{9}$$

Solving the second equation for $Q^{\pi_E}(x_0, a^\star)$ gives

$$Q^{\pi_E}(x_0, a^\star) = \frac{1}{1 - \gamma(1 - \delta_0 + \epsilon)}\left(1 + \gamma(\delta_0 - \epsilon)Q^{\pi_E}(x_1, a^\star)\right). \tag{10}$$

By the Bellman equation in state x_1 and action $a^\star$, we further have

$$Q^{\pi_E}(x_1, a^\star) = 0 + \gamma\delta_1 Q^{\pi_E}(x_0, a^\star) + \gamma(1-\delta_1) Q^{\pi_E}(x_1, a^\star) ,$$

which implies that

$$Q^{\pi_E}(x_1, a^\star) = \frac{\gamma\delta_1}{1-\gamma(1-\delta_1)} Q^{\pi_E}(x_0, a^\star) . \tag{11}$$

Replacing (11) into (10), we get

$$Q^{\pi_E}(x_0, a^\star) = \frac{1}{1-\gamma(1-\delta_0+\epsilon)} + \frac{\gamma^2\delta_1(\delta_0-\epsilon)}{(1-\gamma(1-\delta_0+\epsilon))(1-\gamma(1-\delta_1))} Q^{\pi_E}(x_0, a^\star) .$$

Rearranging the terms gives

$$\begin{aligned} Q^{\pi_E}(x_0, a^\star) &= \left(1 - \frac{\gamma^2\delta_1(\delta_0-\epsilon)}{(1-\gamma(1-\delta_0+\epsilon))(1-\gamma(1-\delta_1))}\right)^{-1} \frac{1}{1-\gamma(1-\delta_0+\epsilon)} \\ &= \frac{1-\gamma(1-\delta_1)}{(1-\gamma(1-\delta_0+\epsilon))(1-\gamma(1-\delta_1)) - \gamma^2\delta_1(\delta_0-\epsilon)} . \end{aligned} \tag{12}$$

Plugging Equation (12) into Equation (11), we can deduce the value of the expert at $(x_1, a^\star)$

$$Q^{\pi_E}(x_1, a^\star) = \frac{\gamma\delta_1}{(1-\gamma(1-\delta_0+\epsilon))(1-\gamma(1-\delta_1)) - \gamma^2\delta_1(\delta_0-\epsilon)} .$$

Looking at the difference $Q^{\pi_E}(x_0, a^\star) - Q^{\pi_E}(x_0, a)$, we can take the difference of Equations (9) and (8) to get

$$\begin{aligned} Q^{\pi_E}(x_0, a^\star) - Q^{\pi_E}(x_0, a) &= \gamma\epsilon\left(Q^{\pi_E}(x_0, a^\star) - Q^{\pi_E}(x_1, a^\star)\right) \\ &= \underbrace{\frac{\gamma\epsilon(1-\gamma)}{(1-\gamma(1-\delta_0+\epsilon))(1-\gamma(1-\delta_1)) - \gamma^2\delta_1(\delta_0-\epsilon)}}_{(\diamondsuit)} . \end{aligned}$$

Next, we upper bound the denominator as follows

$$\begin{aligned} (\diamondsuit) &= 1 - \gamma(1-\delta_0+\epsilon) - \gamma(1-\delta_1) \\ &\quad + \gamma^2(1-\delta_0+\epsilon-\delta_1+\delta_0\delta_1-\epsilon\delta_1-\delta_0\delta_1+\epsilon\delta_1) \\ &= 1 - \gamma(1-\delta_0+\epsilon) - \gamma(1-\delta_1) + \gamma^2(1-\delta_0-\delta_1+\epsilon) \\ &= 1 - \gamma + \gamma\delta_0(1-\gamma) + \gamma\delta_1(1-\gamma) - \gamma(1-\gamma) - \gamma\epsilon(1-\gamma) \\ &= (1-\gamma)^2 + \gamma\delta_0(1-\gamma) + \gamma\delta_1(1-\gamma) - \gamma\epsilon(1-\gamma) \\ &\leq (1-\gamma)^2 + \gamma\delta_0(1-\gamma) + \gamma\delta_1(1-\gamma) , \end{aligned}$$

where the inequality follows from $\gamma\epsilon(1-\gamma) > 0$. Setting $\delta_1 = \delta_0 = \frac{1-\gamma}{\gamma}$, we obtain

$$(\diamondsuit) \leq 3(1-\gamma)^2 ,$$

and it holds that

$$Q^{\pi_E}(x_0, a^\star) - Q^{\pi_E}(x_0, a) \geq \frac{\gamma\epsilon}{3(1-\gamma)} .$$

Moreover, the choice of δ_0 and δ_1 implies that $\nu(\pi_{\text{bad}}, x_0) = \frac{2}{3}$. By definition of the transitions, note that always playing $a \neq a^\star$ like π_{bad} does minimizes the probability of being in state x_0. Thus, for any policy $\tilde{\pi}$, $\nu(\tilde{\pi}, x_0) \geq \nu(\pi_{\text{bad}}, x_0)$, and we have

$$\begin{aligned} \langle \mu(\pi_E) - \mu(\tilde{\pi}), r_{\text{true}} \rangle &\geq \nu(\tilde{\pi}, x_0) \sum_{a \in \mathcal{A}\setminus\{a^\star\}} \tilde{\pi}(a \mid x_0) \frac{\gamma\epsilon}{3(1-\gamma)} \\ &\geq \nu(\pi_{\text{bad}}, x_0)(1 - \tilde{\pi}(a^\star \mid x_0)) \frac{\gamma\epsilon}{3(1-\gamma)} \\ &= 2(1 - \tilde{\pi}(a^\star \mid x_0)) \frac{\gamma\epsilon}{9(1-\gamma)} \\ &\geq \frac{(1 - \tilde{\pi}(a^\star \mid x_0))\epsilon}{9(1-\gamma)} . \end{aligned}$$

where the last inequality follows from $\gamma \geq 1/2$. We now consider the policy $\tilde{\pi} = \bar{\pi}$ produced by a learning algorithm Alg *interacting with the MDP described above (with $\epsilon > 0$). We also consider $\underline{\pi}$ the output of the same learning algorithm* Alg *when interacting with the MDP $\underline{\mathcal{M}}$, a copy of $\mathcal{M}$ with $\epsilon = 0$ (note that in $\underline{\mathcal{M}}$, all actions are identical in* both *states x_0 and x_1, so there is nothing to learn). In $\mathcal{M}$, all actions are identical in state x_1, thus we can assume both policies are the same in state x_1, i.e. $\bar{\pi}(\cdot \mid x_1) = \underline{\pi}(\cdot \mid x_1) = \mathbf{e}_{a^\star}$, and focus exclusively on learning in state x_0. By Pinkser's inequality, we have that*

$$\bar{\pi}(a^\star \mid x_0) - \underline{\pi}(a^\star \mid x_0) \leq \sqrt{2\mathcal{D}_{KL}(\underline{\pi}(\cdot \mid x_0) \| \bar{\pi}(\cdot \mid x_0))},$$

and the previous inequality becomes

$$\langle \mu(\pi_E) - \mu(\bar{\pi}), r_{\text{true}} \rangle \geq \frac{\epsilon}{9(1-\gamma)} \left(1 - \underline{\pi}(a^\star \mid x_0) - \sqrt{2\mathcal{D}_{KL}(\underline{\pi}(\cdot \mid x_0) \| \bar{\pi}(\cdot \mid x_0))}\right).$$

Denote $A = |\mathcal{A}|$ and let $\mathcal{H} = \{\mathcal{M}_i\}_{i=1}^A$ be a collection of MDPs instances where for any $i = 1, \dots, A$, the MDP $\mathcal{M}_i$ is a copy of $\mathcal{M}$ where the ith action is equal to $a^\star$, i.e. $a_i = a^\star$. We denote P_i the corresponding transitions. For any $i \in [\![1, A]\!]$, we denote $\bar{\pi}^i$ the policy output by the learning algorithm Alg *after interacting with the instance $\mathcal{M}_i$, and π_E^i be the expert policy for the instance $\mathcal{M}_i$, i.e. the policy that always plays a_i. We denote $\mu_i(\pi)$ the occupancy measure of any policy π in the MDP $\mathcal{M}_i$. Then, notice that the previous derivations apply for any MDP in $\mathcal{H}$. Thus, summing over $i \in [\![1, A]\!]$ and noting that $\underline{\pi}(\cdot \mid x_0)$ is a probability distribution over $\mathcal{A}$, we get*

$$\sum_{i=1}^A \left\langle \mu_i\left(\pi_E^i\right) - \mu_i\left(\bar{\pi}^i\right), r_{\text{true}} \right\rangle \geq \frac{\epsilon}{9(1-\gamma)} \left(A - 1 - \sum_{i=1}^A \sqrt{2\mathcal{D}_{KL}(\underline{\pi}(\cdot \mid x_0) \| \bar{\pi}^i(\cdot \mid x_0))}\right). \tag{13}$$

For any $i \in [A]$ and $T \in \mathbb{N}^\star$, denote $\mathbb{P}_i^T$ the probability distribution over sets $\mathcal{D}_i^T = \left\{x_0, A_t^i, X_t^i\right\}_{t \in [T]}$ of T transitions starting from x_0 induced by the interaction between the algorithm Alg *and the MDP $\mathcal{M}_i$. Likewise, we denote $\underline{\mathbb{P}}^T$ the probability distribution corresponding to $\underline{\mathcal{M}}$. Then, by the data processing inequality for the KL divergence, for any $i \in [A]$, it holds that*

$$\mathcal{D}_{KL}\left(\underline{\pi}(\cdot \mid x_0) \| \bar{\pi}^i(\cdot \mid x_0)\right) \leq \mathcal{D}_{KL}\left(\underline{\mathbb{P}}^T \| \mathbb{P}_i^T\right).$$

Denoting $\underline{\mathbb{E}}$ the expectation with respect to $\underline{\mathbb{P}}^T$, we can use the Markov property of the environment and continue as follows

$$\begin{aligned}
\mathcal{D}_{KL}\left(\underline{\mathbb{P}}^T \| \mathbb{P}_i^T\right) &= \underline{\mathbb{E}}\left[\log\left(\frac{\prod_{t=1}^T \underline{P}(\underline{X}_t \mid x_0, \underline{A}_t)\, \underline{\mathbb{P}}^T\left(\underline{A}_t \mid \underline{X}_1, \underline{A}_1, \dots, \underline{X}_{t-1}\right)}{\prod_{t=1}^T P_i(\underline{X}_t \mid x_0, \underline{A}_t)\, \mathbb{P}_i^T\left(\underline{A}_t \mid \underline{X}_1, \underline{A}_1, \dots, \underline{X}_{t-1}\right)}\right)\right] \\
&= \underline{\mathbb{E}}\left[\log\left(\frac{\prod_{t=1}^T \underline{P}(\underline{X}_t \mid x_0, \underline{A}_t)}{\prod_{t=1}^T P_i(\underline{X}_t \mid x_0, \underline{A}_t)}\right)\right] \\
&= \underline{\mathbb{E}}\left[\sum_{t=1}^T \log\left(\frac{\underline{P}(\underline{X}_t \mid x_0, \underline{A}_t)}{P_i(\underline{X}_t \mid x_0, \underline{A}_t)}\right)\right],
\end{aligned}$$

where the probabilities on the actions are equal due to running the same algorithm Alg *with the same history up to time $t-1$. Next, we have*

$$\begin{aligned}
\mathcal{D}_{KL}\left(\underline{\mathbb{P}}^T \| \mathbb{P}_i^T\right) &= \sum_{t=1}^T \sum_{(x,a) \in \mathcal{X} \times \mathcal{A}} \underline{\mathbb{P}}^T\left[(\underline{X}_t, \underline{A}_t) = (x, a)\right] \log\left(\frac{\underline{P}(x \mid x_0, a)}{P_i(x \mid x_0, a)}\right) \\
&= \sum_{t=1}^T \sum_{x \in \mathcal{X}} \underline{\mathbb{P}}^T\left[(\underline{X}_t, \underline{A}_t) = (x, a_i)\right] \log\left(\frac{\underline{P}(x \mid x_0, a_i)}{P_i(x \mid x_0, a_i)}\right),
\end{aligned}$$

where we used that the transitions $\underline{P}$ and P_i are the same for any action $a \neq a_i$. By definition of the transitions, we further have

$$\begin{aligned}
\mathcal{D}_{KL}\left(\underline{\mathbb{P}}^T \| \mathbb{P}_i^T\right) &= \sum_{t=1}^T \underline{\mathbb{P}}^T\left[(\underline{X}_t, \underline{A}_t) = (x_0, a_i)\right] \log\left(\frac{1-\delta_0}{1-\delta_0+\epsilon}\right) \\
&\quad + \sum_{t=1}^T \underline{\mathbb{P}}^T\left[(\underline{X}_t, \underline{A}_t) = (x_1, a_i)\right] \log\left(\frac{\delta_0}{\delta_0 - \epsilon}\right).
\end{aligned}$$

Next, by definition of $\underline{\mathbb{P}}^T$, we have

$$\begin{aligned}\mathcal{D}_{KL}\left(\underline{\mathbb{P}}^T \| \mathbb{P}_i^T\right) &= \sum_{t=1}^{T} \underline{\mathbb{P}}^T\left[\underline{A}_t = a_i\right] \underline{P}\left(x_0 \mid x_0, a_i\right) \log\left(\frac{1-\delta_0}{1-\delta_0+\epsilon}\right) \\ &\quad + \sum_{t=1}^{T} \underline{\mathbb{P}}^T\left[\underline{A}_t = a_i\right] \underline{P}\left(x_1 \mid x_0, a_i\right) \log\left(\frac{\delta_0}{\delta_0-\epsilon}\right) \\ &= \underline{\mathbb{E}}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\underline{A}_t = a_i\right\}\right]\left((1-\delta_0)\log\left(\frac{1-\delta_0}{1-\delta_0+\epsilon}\right) + \delta_0 \log\left(\frac{\delta_0}{\delta_0-\epsilon}\right)\right).\end{aligned}$$

By Auer et al., 2008, Lemma 20, we can bound the KL divergence as follows

$$\begin{aligned}\mathcal{D}_{KL}\left(\underline{\mathbb{P}}^T \| \mathbb{P}_i^T\right) &\leq \frac{\epsilon^2}{\delta_0 \log(2)} \underline{\mathbb{E}}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\underline{A}_t = a_i\right\}\right] \\ &\leq \frac{\epsilon^2}{(1-\gamma)\log(2)} \underline{\mathbb{E}}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\underline{A}_t = a_i\right\}\right],\end{aligned}$$

where the last inequality is due to the choice of $\delta_0 = \frac{1-\gamma}{\gamma}$ and $\gamma < 1$. Plugging this into Equation (13) and dividing by A, we have

$$\frac{1}{A}\sum_{i=1}^{A}\left\langle \mu_i\left(\pi_E^i\right) - \mu_i\left(\bar{\pi}^i\right), r_{\text{true}}\right\rangle \geq \frac{\epsilon}{9(1-\gamma)}\left(1 - \frac{1}{A} - \frac{\epsilon}{A}\sum_{i=1}^{A}\sqrt{\frac{\underline{\mathbb{E}}\left[\sum_{t=1}^{T}\mathbb{1}\left\{\underline{A}_t = a_i\right\}\right]}{(1-\gamma)\log(2)}}\right).$$

By Jensen's inequality, we further get

$$\begin{aligned}\frac{1}{A}\sum_{i=1}^{A}\left\langle \mu_i\left(\pi_E^i\right) - \mu_i\left(\bar{\pi}^i\right), r_{\text{true}}\right\rangle &\geq \frac{\epsilon}{9(1-\gamma)}\left(1 - \frac{1}{A} - \epsilon\sqrt{\frac{\underline{\mathbb{E}}\left[\sum_{i=1}^{A}\sum_{t=1}^{T}\mathbb{1}\left\{\underline{A}_t = a_i\right\}\right]}{A(1-\gamma)\log(2)}}\right) \\ &\geq \frac{1}{9(1-\gamma)}\left(\frac{\epsilon}{2} - \epsilon^2\sqrt{\frac{T}{A(1-\gamma)\log(2)}}\right),\end{aligned}$$

where the second inequality follows from $\sum_{i=1}^{A}\mathbb{1}\left\{\underline{A}_t = a_i\right\} = 1$ almost surely for any t and $1 - \frac{1}{A} \geq \frac{1}{2}$. Note that the value of ϵ maximizing the lower bound is given by $\epsilon^\star = \frac{1}{4}\sqrt{\frac{A(1-\gamma)\log(2)}{T}}$. To satisfy the constraint $\epsilon^\star \in (0, \delta_0)$ with $\delta_0 = \frac{1-\gamma}{\gamma}$, assume we have $T \geq \frac{\gamma^2 A \log(2)}{16(1-\gamma)}$. We plug the value of $\epsilon^\star$ in the previous inequality to get

$$\begin{aligned}\frac{1}{A}\sum_{i=1}^{A}\left\langle \mu_i\left(\pi_E^i\right) - \mu_i\left(\bar{\pi}^i\right), r_{\text{true}}\right\rangle &\geq \frac{1}{16 \cdot 9(1-\gamma)}\sqrt{\frac{A(1-\gamma)\log(2)}{T}} \\ &= \frac{1}{144}\sqrt{\frac{A\log(2)}{(1-\gamma)T}},\end{aligned}$$

The average can be upper bounded by the maximum, thus

$$\begin{aligned}\max_{i=1,\ldots,A}\left\langle \nu_0, V_{\mathcal{M}_i}^{\pi_E^i} - V_{\mathcal{M}_i}^{\bar{\pi}^i}\right\rangle &= \frac{1}{1-\gamma}\max_{i=1,\ldots,A}\left\langle \mu_i\left(\pi_E^i\right) - \mu_i\left(\bar{\pi}^i\right), r_{\text{true}}\right\rangle \\ &\geq \frac{1}{144}\sqrt{\frac{A\log(2)}{(1-\gamma)^3 T}}.\end{aligned}$$

What remains is to set the number of samples T to make the lower bound small enough to make $\max_{i=1,\ldots,A}\left\langle \nu_0, V_{\mathcal{M}_i}^{\pi_E^i} - V_{\mathcal{M}_i}^{\bar{\pi}^i}\right\rangle = \mathcal{O}(\varepsilon)$ possible, i.e. we need to have $T = \Omega\left(\frac{A}{(1-\gamma)^3\varepsilon^2}\right)$ samples.

Therefore, we need $T = \Omega\left(\frac{A}{(1-\gamma)^3\varepsilon^2}\right)$ *samples to learn a* $\mathcal{O}(\varepsilon)$*-suboptimal policy in the MDP that achieves the maximum. In order to derive a lower bound on the episodes number* K *we can divide the sample complexity lower bound for* T *by the the expected number of transitions per episode which is* $(1-\gamma)^{-1}$*. This gives* $K = \Omega\left(\frac{A}{(1-\gamma)^2\varepsilon^2}\right)$*. We can conclude by noting that our construction used in the lower bound is a linear MDP with dimensionality* $d = 2 + 2|\mathcal{A}|$*, thus we have* $K = \Omega\left(\frac{d}{(1-\gamma)^2\varepsilon^2}\right)$.

E.2 Proof of Theorem 6 (lower bound on the number of expert transitions)

Proof Idea: The construction of the lower bound consists in relating the problem to that of distinguishing two Bernoullis distributions with close means. For that, we consider two MDPs $\mathcal{M}_0$ and $\mathcal{M}_1$ that only differ in their reward function. They have two states $\mathcal{X} = \{x_0, x_1\}$ and $|\mathcal{A}|$ actions available at each state. The initial distribution ν_0 is chosen to be the uniform distribution over $\mathcal{X}$. In state x_1, any action a induces the same transition kernel: $P(x_0 \mid x_1, a) = \delta$. In state x_0, any action a except some action $a^\star$ is such that $P(x_1|x_0, a) = \delta$. However, the special action $a^\star$ allows to stay in the state x_0 with a slightly higher probability, *i.e.* $P(x_1 \mid x_0, a^\star) = \delta - \epsilon$. Then, the reward function in $\mathcal{M}_0$ is defined as $r^0_{\text{true}}(x_0, \cdot) = W_{\max}$, and $r^0_{\text{true}}(x_0, \cdot) = 0$, while in $\mathcal{M}_1$, it is defined as $r^1_{\text{true}}(x_0, \cdot) = 0$, $r^1_{\text{true}}(x_1, \cdot) = W_{\max}$. Finally, we define an expert π^0_E for $\mathcal{M}_0$ as the policy that always play the action $a^\star$, and an expert π^1_E for $\mathcal{M}_1$ that always play some action $a \neq a^\star$. We then show that the expert occupancy measures satisfy $\nu\left(\pi^0_E, x_0\right) = 1/2 + \Delta$, for some small $\Delta > 0$, while $\nu\left(\pi^1_E, x_0\right) = \frac{1}{2}$. The remaining step is to reduce this problem to a lower bound on the regret of a two-arm Bernoulli bandits instance with means $(1/2, 1/2)$ and $(1/2 + \Delta, 1/2 - \Delta)$. The proof is formally presented hereafter.

Theorem 6 ***(Lower Bound on*** τ_E***)*** *Let* $\gamma \geq \frac{1}{2}$*. For any algorithm* Alg*, there exists an MDP* $\mathcal{M}$ *and an expert policy* π_E *such that* Alg *taking as input the transitions dynamics and an expert dataset of size* τ_E *requires* $\tau_E = \Omega\left(\frac{W^2_{\max}}{(1-\gamma)^2\varepsilon^2}\right)$ *to guarantee* $\mathbb{E}_{\text{Alg}}\left[\left\langle \nu_0, V^{\pi_E}_{\mathcal{M}} - V^{\pi^{\text{out}}}_{\mathcal{M}} \right\rangle\right] = \mathcal{O}(\varepsilon)$.

Proof 17 *As mentioned earlier, it is sufficient to consider deterministic algorithms that map histories to policies. The lower bound for randomized algorithms follows by an application of Fubini's theorem (see Bubeck et al., 2012). We consider two MDPs* $\mathcal{H} = \{\mathcal{M}_0, \mathcal{M}_1\}$ *with the same state space* $\mathcal{X} = \{x_0, x_1\}$ *and* $|\mathcal{A}|$ *actions available in each state. The initial distribution* ν_0 *is chosen to be the uniform distribution over* $\mathcal{X}$*, i.e.* $\nu_0(x_0) = \nu_0(x_1) = \frac{1}{2}$*. The transitions are the same in both MDPs: in state* x_1*, each action* $a \in \mathcal{A}$ *induces the following transition kernel*

$$P(x_0 \mid x_1, a) = \delta, \quad P(x_1 \mid x_1, a) = 1 - \delta$$

while in state x_0*, there is an action* $a^\star$ *giving a slightly higher probability on staying in state* x_0*, i.e.*

$$\begin{aligned} P(x_0 \mid x_0, a^\star) &= 1 - \delta + \epsilon, \quad P(x_1 \mid x_0, a^\star) = \delta - \epsilon \\ \forall a \neq a^\star, P(x_0 \mid x_0, a) &= 1 - \delta, \quad P(x_1 \mid x_0, a) = \delta\,. \end{aligned}$$

The reward functions, r^0_{true} *and* r^1_{true}*, are different. In* $\mathcal{M}_0$*, the "good" state is* x_0*, i.e. for any action* $a \in \mathcal{A}$*, we set* $r^0_{\text{true}}(x_0, a) = W_{\max}$*,* $r^0_{\text{true}}(x_1, a) = 0$*, and in* $\mathcal{M}_1$*, the "good" state is* x_1*, i.e.* $r^1_{\text{true}}(x_0, a) = 0$*, and* $r^1_{\text{true}}(x_1, a) = W_{\max}$*. Note that* $W_{\max} = R_{\max}$ *due to using the features* $\varphi_r(x, a) = \mathbf{e}_x$ *for any state-action pair* x, a*.*

Then, we define one expert policy for each MDP. In $\mathcal{M}_0$*, the expert* π^0_E *is the policy that always plays* $a^\star$ *and in* $\mathcal{M}_1$*, the expert* π^1_E *is the policy that always plays an action* $a \neq a^\star$*. Therefore, the state occupancy measure of expert* π^0_E *in MDP* $\mathcal{M}_0$ *has the highest mass in state* x_0*, while* π^1_E *put equal mass on both states. Indeed, writing the flow constraints for both experts, we have*

$$\begin{pmatrix} 1 - \gamma + \gamma\delta - \gamma\epsilon & -\gamma\delta \\ -\gamma(\delta - \epsilon) & 1 - \gamma + \gamma\delta \end{pmatrix} \nu\left(\pi^0_E\right) = \nu_0\,,$$

$$\begin{pmatrix} 1 - \gamma + \gamma\delta & -\gamma\delta \\ -\gamma\delta & 1 - \gamma + \gamma\delta \end{pmatrix} \nu\left(\pi^1_E\right) = \nu_0\,.$$

Solving these linear systems using, e.g., Cramer's rule, we obtain

$$\nu\left(\pi_E^0, x_0\right) = \frac{1-\gamma+2\gamma\delta}{2\left(1-\gamma-\gamma\epsilon+2\gamma\delta\right)}, \qquad \nu\left(\pi_E^0, x_1\right) = \frac{1-\gamma-2\gamma\epsilon+2\gamma\delta}{2\left(1-\gamma-\gamma\epsilon+2\gamma\delta\right)},$$
$$\nu\left(\pi_E^1, x_0\right) = \frac{1}{2}, \qquad \nu\left(\pi_E^1, x_1\right) = \frac{1}{2}.$$

For $i \in \{0,1\}$, let $\bar{\pi}^i$ be the policy output by Alg *when given a dataset $\mathcal{D}_{\pi_E^i}$ as input and let $V_i^{\bar{\pi}^i}$ be the value function of policy $\bar{\pi}^i$ corresponding to the reward function r^i_{true} from the MDP $\mathcal{M}_i$. By definition of r^i_{true}, we can write*

$$\begin{aligned}\frac{1}{2}\sum_{i\in\{0,1\}}\left\langle \nu_0, V_i^{\pi_E^i} - V_i^{\bar{\pi}^i}\right\rangle &= \frac{1}{2\left(1-\gamma\right)}\sum_{i\in\{1,2\}}\left\langle \mu\left(\pi_E^i\right) - \mu\left(\bar{\pi}^i\right), r^i_{\text{true}}\right\rangle \\ &= \frac{W_{\max}}{2\left(1-\gamma\right)}\left(\nu\left(\pi_E^0, x_0\right) - \nu\left(\bar{\pi}^0, x_0\right) + \nu\left(\pi_E^1, x_1\right) - \nu\left(\bar{\pi}^1, x_1\right)\right).\end{aligned} \tag{14}$$

Thus, we need to compute the difference between state occupancy measures. Let $\tilde{\pi}$ be an arbitrary policy and denote $\alpha \in [0,1]$ the probability of playing action $a^\star$ in state x_0, i.e. $\tilde{\pi}\left(a^\star \mid x_0\right) = \alpha$. Writing down the flow constraints again, we can show that

$$\nu\left(\tilde{\pi}, x_0\right) = \frac{1-\gamma+2\gamma\delta}{2\left(1-\gamma-\gamma\alpha\epsilon+2\gamma\delta\right)}, \quad \nu\left(\tilde{\pi}, x_1\right) = \frac{1-\gamma-2\gamma\alpha\epsilon+2\gamma\delta}{2\left(1-\gamma-\gamma\alpha\epsilon+2\gamma\delta\right)}.$$

Looking at the difference with π_E^0 in state x_0, we have

$$\begin{aligned}\nu\left(\pi_E^0, x_0\right) - \nu\left(\tilde{\pi}, x_0\right) &= \frac{1-\gamma+2\gamma\delta}{2\left(1-\gamma-\gamma\epsilon+2\gamma\delta\right)} - \frac{1-\gamma+2\gamma\delta}{2\left(1-\gamma-\gamma\alpha\epsilon+2\gamma\delta\right)} \\ &= \frac{\left(1-\gamma+2\gamma\delta\right)\left(\left(-\gamma\alpha\epsilon\right) - \left(-\gamma\epsilon\right)\right)}{2\left(1-\gamma-\gamma\epsilon+2\gamma\delta\right)\left(1-\gamma-\gamma\alpha\epsilon+2\gamma\delta\right)} \\ &= \frac{\left(1-\gamma+2\gamma\delta\right)\gamma\epsilon\left(1-\alpha\right)}{2\left(1-\gamma-\gamma\epsilon+2\gamma\delta\right)\left(1-\gamma-\gamma\alpha\epsilon+2\gamma\delta\right)}.\end{aligned}$$

Setting $\delta = \frac{1-\gamma}{\gamma}$ and noting $\epsilon \geq 0$, $\gamma \geq \frac{1}{2}$, we can lower bound the difference as follows

$$\begin{aligned}\nu\left(\pi_E^0, x_0\right) - \nu\left(\tilde{\pi}, x_0\right) &= \frac{3\left(1-\gamma\right)\gamma\epsilon\left(1-\alpha\right)}{2\left(3\left(1-\gamma\right)-\gamma\epsilon\right)\left(3\left(1-\gamma\right)-\gamma\alpha\epsilon\right)} \\ &\geq \frac{\epsilon\left(1-\alpha\right)}{12\left(1-\gamma\right)}.\end{aligned} \tag{15}$$

Likewise, the difference between $\nu\left(\pi_E^1\right)$ and $\nu\left(\tilde{\pi}\right)$ in state x_1 is given by

$$\begin{aligned}\nu\left(\pi_E^1, x_1\right) - \nu\left(\tilde{\pi}, x_1\right) &= \frac{1}{2} - \frac{1-\gamma-2\gamma\alpha\epsilon+2\gamma\delta}{2\left(1-\gamma-\gamma\alpha\epsilon+2\gamma\delta\right)} \\ &= \frac{\gamma\alpha\epsilon}{2\left(1-\gamma-\gamma\alpha\epsilon+2\gamma\delta\right)}.\end{aligned}$$

Using the definition of δ, and again $\epsilon \geq 0$, $\gamma \geq \frac{1}{2}$, we get

$$\begin{aligned}\nu\left(\pi_E^1, x_1\right) - \nu\left(\tilde{\pi}, x_1\right) &= \frac{\gamma\alpha\epsilon}{2\left(3\left(1-\gamma\right)-\gamma\alpha\epsilon\right)} \\ &\geq \frac{\epsilon\alpha}{12\left(1-\gamma\right)}.\end{aligned} \tag{16}$$

Plugging Inequalities (15) and (16) into Equation (14) with $\alpha = \bar{\pi}^0\left(a^\star \mid x_0\right)$ and $\alpha = \bar{\pi}^1\left(a^\star \mid x_0\right)$ respectively, we get

$$\begin{aligned}\frac{1}{2}\sum_{i\in\{0,1\}}\left\langle \nu_0, V_i^{\pi_E^i} - V_i^{\bar{\pi}^i}\right\rangle &\geq \frac{\epsilon W_{\max}}{24\left(1-\gamma\right)^2}\left(1 - \bar{\pi}^0\left(a^\star \mid x_0\right) + \bar{\pi}^1\left(a^\star \mid x_0\right)\right) \\ &= \frac{\epsilon W_{\max}}{24\left(1-\gamma\right)^2}\left(\sum_{a\neq a^\star}\bar{\pi}^0\left(a \mid x_0\right) + \bar{\pi}^1\left(a^\star \mid x_0\right)\right).\end{aligned}$$

Next, we can lower bound the right hand side using the Bretagnolle-Huber inequality (see Bretagnolle and Huber, 1979, and Lattimore and Szepesvári, 2020, Theorem 14.2), which gives

$$\frac{1}{2}\sum_{i\in\{0,1\}}\left\langle \nu_0, V_i^{\pi_E^i} - V_i^{\bar{\pi}^i}\right\rangle \geq \frac{\epsilon W_{\max}}{24\left(1-\gamma\right)^2}\exp\left(-\mathcal{D}_{KL}\left(\bar{\pi}^0\left(\cdot \mid x_0\right)\|\bar{\pi}^1\left(\cdot \mid x_0\right)\right)\right). \tag{17}$$

Then, using the data processing inequality and using the fact that the learning algorithm produces $\bar{\pi}^i$ as a deterministic function of the dataset $\mathcal{D}_{\pi_E^i}$ for $i=0,1$, we have that

$$\mathcal{D}_{KL}\left(\bar{\pi}^0\left(\cdot \mid x_0\right)\|\bar{\pi}^1\left(\cdot \mid x_0\right)\right) \leq \mathcal{D}_{KL}\left(\mathbb{P}_0^{\tau_E}\|\mathbb{P}_1^{\tau_E}\right),$$

where, for $i\in\{0,1\}$, we denoted $\mathbb{P}_i^{\tau_E}$ the probability distribution over datasets of size τ_E induced by the interaction between the expert π_E^i and the environment (analog to what is done in the proof of Theorem 5). Next, we denote $\mathrm{kl}\left(p,q\right)$ and $\chi^2\left(p,q\right)$ the KL and chi-squared divergences between bernoulli distributions of means p and p', i.e.

$$\mathrm{kl}\left(p,q\right) = p\log\left(\frac{p}{q}\right) + \left(1-p\right)\log\left(\frac{1-p}{1-q}\right)$$
$$\chi^2\left(p,q\right) = \frac{\left(p-q\right)^2}{q\left(1-q\right)}.$$

By definition of the KL, we have

$$\begin{aligned}
\mathcal{D}_{KL}\left(\mathbb{P}_0^{\tau_E}\|\mathbb{P}_1^{\tau_E}\right) &= \tau_E\cdot\mathrm{kl}\left(\frac{3\left(1-\gamma\right)}{2\left(3\left(1-\gamma\right)-\gamma\epsilon\right)},\frac{1}{2}\right)\\
&\leq \tau_E\cdot\chi^2\left(\frac{3\left(1-\gamma\right)}{2\left(3\left(1-\gamma\right)-\gamma\epsilon\right)},\frac{1}{2}\right)\\
&= \tau_E\cdot\chi^2\left(\frac{1}{2}+\frac{\gamma\epsilon}{3\left(1-\gamma\right)-\gamma\epsilon},\frac{1}{2}\right)\\
&= \frac{4\tau_E\gamma^2\epsilon^2}{\left(3\left(1-\gamma\right)-\gamma\epsilon\right)^2}\\
&\leq \frac{\tau_E\gamma^2\epsilon^2}{\left(1-\gamma\right)^2},
\end{aligned}$$

where the first inequality follows from the concavity of the logarithm function, and the second inequality uses the fact that $\epsilon\leq\delta=\frac{1-\gamma}{\gamma}$. Thus, plugging in this last inequality into Equation (17), we obtain

$$\begin{aligned}
\frac{1}{2}\sum_{i\in\{0,1\}}\left\langle \nu_0, V_i^{\pi_E^i} - V_i^{\bar{\pi}^i}\right\rangle &\geq \frac{\epsilon W_{\max}}{24\left(1-\gamma\right)^2}\exp\left(-\frac{\tau_E\gamma^2\epsilon^2}{\left(1-\gamma\right)^2}\right)\\
&\geq \frac{\epsilon W_{\max}}{24\left(1-\gamma\right)^2}\exp\left(-\frac{\tau_E\epsilon^2}{\left(1-\gamma\right)^2}\right),
\end{aligned}$$

where we used $\gamma<1$ in the second inequality. Introducing $\epsilon'=\epsilon\left(1-\gamma\right)^{-1}$, we can rewrite the previous inequality as

$$\frac{1}{2}\sum_{i\in\{0,1\}}\left\langle \nu_0, V_i^{\pi_E^i} - V_i^{\bar{\pi}^i}\right\rangle \geq \frac{W_{\max}\epsilon'}{24\left(1-\gamma\right)}\exp\left(-\tau_E\left(\epsilon'\right)^2\right).$$

It remains to make the lower bound small enough. To bound the average suboptimality gap by $\frac{W_{\max}\epsilon'}{24e(1-\gamma)}$ and have $\frac{1}{2}\sum_{i\in\{0,1\}}\left\langle \nu_0, V_i^{\pi_E^i} - V_i^{\bar{\pi}^i}\right\rangle \leq \frac{W_{\max}\epsilon'}{24e(1-\gamma)}$, we need at least $\tau_E\geq\frac{1}{(\epsilon')^2}$ expert transitions. Therefore, to achieve

$$\frac{1}{2}\sum_{i\in\{0,1\}}\left\langle \nu_0, V_i^{\pi_E^i} - V_i^{\bar{\pi}^i}\right\rangle \leq \varepsilon,$$

for some $\varepsilon>0$, we need to choose $\epsilon'=24e(1-\gamma)\varepsilon/W_{\max}$, which means that every algorithm needs at least $\tau_E\geq\frac{W_{\max}^2}{24^2e^2(1-\gamma)^2\varepsilon^2}$ to guarantee a suboptimality gap of order ε.

F Technical tools

F.1 Reinforcement learning

Proposition 1 *The occupancy measure $\mu(\pi)$ of any policy π satisfies the following system of equations:*

$$E^{\intercal}\mu\left(\pi\right)=\gamma P^{\intercal}\mu\left(\pi\right)+\left(1-\gamma\right)\nu_{0}\,. \tag{18}$$

Proof 18 *Define the transition kernel induced by policy π as P_{π}, with $P_{\pi}(\cdot|x) = \mathbb{E}_{A\sim\pi(\cdot|x)}\left[P(\cdot|x,A)\right]$. The proof follows from the following standard calculation:*

$$\begin{aligned}
E^{\intercal}\mu\left(\pi\right) &= \left(1-\gamma\right)\sum_{\tau=0}^{\infty}\left(\gamma P_{\pi}^{\intercal}\right)^{\tau}\nu_{0}\\
&= \left(1-\gamma\right)\sum_{\tau=1}^{\infty}\left(\gamma P_{\pi}^{\intercal}\right)^{\tau}\nu_{0}+\left(1-\gamma\right)\nu_{0}\\
&= \gamma P_{\pi}\left(1-\gamma\right)\sum_{\tau=0}^{\infty}\left(\gamma P_{\pi}^{\intercal}\right)^{\tau}\nu_{0}+\left(1-\gamma\right)\nu_{0}\\
&= \gamma P_{\pi}E^{\intercal}\mu\left(\pi\right)+\left(1-\gamma\right)\nu_{0}\\
&= \gamma P\mu\left(\pi\right)+\left(1-\gamma\right)\nu_{0},
\end{aligned}$$

where the last step follows from the easily-checked fact that $P\mu\left(\pi\right)=P_{\pi}E^{\intercal}\mu\left(\pi\right)$.

Lemma 6 *Let π be any policy, $Q\in\mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ be any function defined on $\mathcal{X}\times\mathcal{A}$, and $V\in\mathbb{R}^{\mathcal{X}}$ be such that for any x, $V\left(x\right)=\mathbb{E}_{A\sim\pi(\cdot|x)}\left[Q\left(x,A\right)\right]$. Then*

$$\left\langle \mu\left(\pi\right),EV\right\rangle =\left\langle \mu\left(\pi\right),Q\right\rangle \,.$$

Proof 19 *We have*

$$\begin{aligned}
\left\langle \mu\left(\pi\right),EV\right\rangle &= \sum_{x\in\mathcal{X}}\nu\left(\pi,x\right)V\left(x\right)\\
&= \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\nu\left(\pi,x\right)\pi\left(a\mid x\right)Q\left(x,a\right)\\
&= \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\mu\left(\pi,x,a\right)Q\left(x,a\right)\\
&= \left\langle \mu\left(\pi\right),Q\right\rangle \,,
\end{aligned}$$

where the second equality follows from the definition of the function V and the first equality from the definition of the state-action occupancy measure.

Lemma 7 *Let π and π' be two policies. Then,*

$$\mathcal{D}_{KL}\left(\mu\left(\pi\right)\|\mu\left(\pi'\right)\right)\leq\frac{1}{1-\gamma}\left\langle \nu\left(\pi\right),\mathcal{D}_{KL}\left(\pi\|\pi'\right)\right\rangle \,.$$

Proof 20 *Using the chain rule of the relative entropy, we write*

$$\mathcal{D}_{KL}\left(\mu\left(\pi\right)\|\mu\left(\pi'\right)\right)=\mathcal{D}_{KL}\left(\nu\left(\pi\right)\|\nu\left(\pi'\right)\right)+\left\langle \nu\left(\pi\right),\mathcal{D}_{KL}\left(\pi\|\pi'\right)\right\rangle \,.$$

By the flow constraints and the joint convexity of the relative entropy, we bound the first term as

$$\begin{aligned}
\mathcal{D}_{KL}\left(\nu\left(\pi\right)\|\nu\left(\pi'\right)\right) &= \mathcal{D}_{KL}\left(\gamma P^{\intercal}\mu\left(\pi\right)+\left(1-\gamma\right)\nu_{0}\|\gamma P^{\intercal}\mu\left(\pi'\right)+\left(1-\gamma\right)\nu_{0}\right)\\
&\leq \left(1-\gamma\right)\mathcal{D}_{KL}\left(\nu_{0}\|\nu_{0}\right)+\gamma\mathcal{D}_{KL}\left(P^{\intercal}\mu\left(\pi\right)\|P^{\intercal}\mu\left(\pi'\right)\right)\\
&= \gamma\mathcal{D}_{KL}\left(P^{\intercal}\mu\left(\pi\right)\|P^{\intercal}\mu\left(\pi'\right)\right)\\
&\leq \gamma\mathcal{D}_{KL}\left(\mu\left(\pi\right)\|\mu\left(\pi'\right)\right)\,,
\end{aligned}$$

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

Lemma 8 *For any MDP $\mathcal{M}$, any ascension function p^+, and any policy π, we have for any state-action pair $(x,a) \in \mathcal{X} \times \mathcal{A}$,*

$$\mu^+(\pi, x, a) \leq \mu(\pi, x, a) ,$$

where $\mu^+(\pi)$ denotes the state-action occupancy of π in $\mathcal{M}^+$, the optimistically augmented MDP induced by p^+.

Proof 21 *Let us consider a process $(X_\tau, A_\tau)_{\tau \in \mathbb{N}}$ generated by the policy π in the MDP $\mathcal{M}$, that is, such that $X_0 \sim \nu_0$, and for any $\tau \in \mathbb{N}$, $A_\tau \sim \pi(\cdot \mid X_\tau)$, and $X_{\tau+1} \sim P(\cdot \mid X_\tau, A_\tau)$. Additionally, we define a process $(X_\tau^+, A_\tau^+)_{\tau \in \mathbb{N}}$ coupled to the process defined above as follows. At the first stage, we set $X_\tau^+ = X_0$. Then for any $\tau \geq 1$, the coupled process evolves as*

$$X_{\tau+1}^+, A_{\tau+1}^+ = \begin{cases} X_{\tau+1}, A_{\tau+1} & \text{w.p.} \quad 1 - p^+(X_\tau, A_\tau) \quad \text{if} \quad X_\tau^+, A_\tau^+ = X_\tau, A_\tau \\ x^+, a & \text{w.p.} \quad p^+(X_\tau, A_\tau) \quad \text{if} \quad X_\tau^+, A_\tau^+ = X_\tau, A_\tau \\ x^+, a & \text{if} \quad X_\tau^+, A_\tau^+ \neq X_\tau, A_\tau \end{cases} .$$

It is straightforward to check that this process follows the dynamics of the optimistically augmented MDP $\mathcal{M}^+(r, p^+)$ (since its transitions obey the kernel P^+). By definition, for any state-action pair $(x,a) \in \mathcal{X} \times \mathcal{A}$, we have

$$\begin{aligned}
\mu^+(\pi, x, a) &= (1-\gamma) \sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{P}[X_\tau^+ = x, A_\tau^+ = a] \\
&= (1-\gamma) \sum_{\tau=0}^{\infty} \gamma^\tau \left(\mathbb{P}[X_\tau^+ = x, A_\tau^+ = a, X_\tau^+ \neq x^+] + \mathbb{P}[X_\tau^+ = x, A_\tau^+ = a, X_\tau^+ = x^+] \right) \\
&= (1-\gamma) \sum_{\tau=0}^{\infty} \gamma^\tau \left(\mathbb{P}[X_\tau = x, A_\tau = a, X_\tau^+ \neq x^+] + 0 \right) \\
&\leq (1-\gamma) \sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{P}[X_\tau = x, A_\tau = a] \\
&= \mu(\pi, x, a) .
\end{aligned}$$

In the third equality, the second term within the sum is equal to zero because $x \neq x^+$, and in the other term we replaced (X_τ^+, A_τ^+) by (X_τ, A_τ) because the two coincide long as $X_\tau^+ \neq x^+$. This concludes the proof.

F.2 Linear algebra and analysis

Lemma 9 *Under the event $\mathcal{E}_L$, the number of epochs $E(K)$ in Algorithm 1 is bounded as*

$$E(K) \leq 5d \log\left(1 + \frac{B^2 T}{d}\right) .$$

where $T = L_{\max} K = \frac{\log\left(\frac{K}{\delta}\right) K}{1-\gamma}$.

Proof 22 *In the following, we denote $\varphi_t = \varphi(x_t, a_t)$ for any t. The bound on the number of epochs is derived observing that since the determinant of the matrix Λ_k can grow at most linearly then the condition is triggered at most a logarithmic number of times. In particular notice that*

$$\det\left(\Lambda_{t_{E(K)}}\right) \geq 2 \det\left(\Lambda_{t_{E(K)-1}}\right) \geq 2^2 \det\left(\Lambda_{t_{E(K)-2}}\right) \geq 2^{E(K)-1} \det(I) = 2^{E(K)-1} .$$

Hence, it holds that $E(K) - 1 \leq \frac{1}{\log 2} \log\left(\det \Lambda_{t_{E(K)}}\right)$. Then, denoting $T_{K+1} = T_K + L_K$ where L_K is the length of episode K, we have that

$$\begin{aligned}
E(K) &\leq 1 + \frac{1}{\log 2} \log\left(\det\left(\Lambda_{T_{K+1}}\right)\right) && \left(\Lambda_{t_{E(K)}} \preceq \Lambda_{T_{K+1}}\right) \\
&\leq 1 + \frac{d}{\log 2} \log\left(\frac{\operatorname{trace}\left(\Lambda_{T_{K+1}}\right)}{d}\right) && \textit{(trace-determinant inequality)} .
\end{aligned}$$

By definition of the covariance matrix,

$$\begin{aligned} E(K) &\leq 1 + \frac{d}{\log 2} \log\left(\frac{\operatorname{trace}\left(\sum_{t\in[T_{K+1}]} \varphi_t \varphi_t^\intercal\right) + d}{d} \right) \\ &= 1 + \frac{d}{\log 2} \log\left(1 + \frac{\sum_{t\in[T_{K+1}]} \|\varphi_t\|_2^2}{d} \right) \\ &\leq 1 + \frac{d}{\log 2} \log\left(1 + \frac{B^2 T}{d} \right) \\ &\leq 5d \log\left(1 + \frac{B^2 T}{d} \right), \end{aligned}$$

where the first equality follows from properties of the trace and the second inequality follows from $\|\varphi_t\|_2 \leq B$ *and* $T_{K+1} \leq T$ *which holds under* $\mathcal{E}_L$*.*

The following lemma is a generalization of Lemma 19 of Cassel and Rosenberg [2024] for an arbitrary threshold $\omega \geq 0$.

Lemma 10 *For all* $z \geq 0$*,* $\omega \geq 0$ *it holds that* $\sigma(z-\omega) \leq 2\left(z^2 + \exp(-\omega)\right)$*.*

Proof 23 *Let us consider the function* $g : z \mapsto \sigma(z-\omega) - \left(z + \frac{1}{e^{\omega/2}}\right)^2$*. Note that for any* z*, we have* $\sigma'(z) = \sigma(z)\sigma(-z)$*. Thus, the first two derivatives of* g *are given by*

$$\begin{aligned} g'(z) &= \sigma(z-\omega)\sigma(\omega-z) - 2\left(z + \frac{1}{e^{\omega/2}}\right), \\ g''(z) &= \sigma(z-\omega)\sigma(\omega-z)^2 - \sigma(z-\omega)^2\sigma(\omega-z) - 2. \end{aligned}$$

Since $\sigma(z) \in (0,1)$ *for any* z*, the second derivative of* g *is nonpositive,* $g''(z) \leq 0$*, and* g *is concave. By the first order condition, for any* $z \geq 0$*,*

$$g(z) \leq g(0) + g'(0)\, z.$$

Furthermore, note that

$$g(0) = \sigma(-\omega) - \frac{1}{e^\omega} = \frac{1}{1+e^\omega} - \frac{1}{e^\omega} \leq 0,$$

and

$$\begin{aligned} g'(0) &= \frac{1}{1+e^\omega}\frac{1}{1+e^{-\omega}} - \frac{2}{e^{\omega/2}} \\ &\leq \frac{1}{1+e^\omega} - \frac{2}{e^{\omega/2}} \\ &\leq -\frac{1}{e^{\omega/2}} \\ &\leq 0, \end{aligned}$$

where we first used that $e^{-\omega} \geq 0$ *for any* $\omega \geq 0$ *and then that* $x + 1 \geq \sqrt{x}$ *for any* $x \geq 0$*. Thus, it holds that* $g(z) \leq 0$ *for all* $z \geq 0$*, i.e.* $\sigma(z-\omega) \leq \left(z + \frac{1}{e^{\omega/2}}\right)^2$*. Using* $(a+b)^2 \leq 2\left(a^2+b^2\right)$*, it holds that*

$$\sigma(z-\omega) \leq 2\left(z^2 + e^{-\omega}\right).$$

We present a variant of Lemma 18 of Cassel and Rosenberg [2024] which is valid for $\omega \geq 2$ instead of $\omega \geq 0$, but is sharper by a factor of 2.

Lemma 11 *For all* $\omega \geq 2$*, it holds that*

$$\max_{z \geq 0} z \cdot \sigma(\omega - \alpha z) \leq \frac{\omega}{\alpha}.$$

Proof 24 *Let $\alpha > 0$, $\omega \geq 2$, and $g : z \geq 0 \mapsto z \cdot \sigma(\omega - \alpha z)$. We recall that the derivative of the sigmoid function is given for any z by $\sigma'(z) = \sigma(z)\sigma(-z)$, and that $\sigma(-z) = 1 - \sigma(z)$. g is twice differentiable. Its first derivative is given by*

$$\begin{aligned} g'(z) &= \sigma(\omega - \alpha z) - \alpha z \sigma(\omega - \alpha z)\left[1 - \sigma(\omega - \alpha z)\right] \\ &= \sigma(\omega - \alpha z)\left[1 - \alpha z\left(1 - \sigma(\omega - \alpha z)\right)\right] . \end{aligned}$$

We set the derivative to zero and solve the equation to find the critical points. We have

$$\begin{aligned} g'(z) = 0 \quad \textit{iff} \quad & \alpha z = \frac{1}{1 - \sigma(\omega - \alpha z)} \\ \textit{iff} \quad & \alpha z = 1 + e^{\omega - \alpha z} \\ \textit{iff} \quad & (\alpha z - 1) e^{\alpha z - 1} = e^{\omega - 1} . \end{aligned} \tag{19}$$

For $x > 0$, the equation $we^w = x$ has exactly one positive solution $w = W(x)$ which increases with x and where W denotes the Lambert function. Thus, $g'(z) = 0$ if and only if $\alpha z - 1 = W\left(e^{\omega - 1}\right)$, i.e. $z^\star = \frac{W\left(e^{\omega-1}\right)+1}{\alpha}$. We check that $z^\star$ is a local maximum. The second derivative of g is given by

$$\begin{aligned} g''(z) = & -2\alpha\sigma(\omega - \alpha z)\left[1 - \sigma(\omega - \alpha z)\right] \\ & + \alpha^2 z \sigma(\omega - \alpha z)\left[1 - \sigma(\omega - \alpha z)\right]^2 \\ & - \alpha^2 z \sigma(\omega - \alpha z)^2\left[1 - \sigma(\omega - \alpha z)\right] \\ = & -2\alpha\sigma(\omega - \alpha z)\left[1 - \sigma(\omega - \alpha z)\right] \\ & + \alpha^2 z \sigma(\omega - \alpha z)\left[1 - \sigma(\omega - \alpha z)\right]\left[1 - 2\sigma(\omega - \alpha z)\right] . \end{aligned}$$

We evaluate it at the critical point $z^\star$ and simplify the expression using Equation 19

$$\begin{aligned} g''(z^\star) &= -2\alpha\sigma(\omega - \alpha z^\star)\left[1 - \sigma(\omega - \alpha z^\star)\right] \\ &\quad + \alpha\sigma(\omega - \alpha z^\star)\left[1 - 2\sigma(\omega - \alpha z^\star)\right] \\ &= -\alpha\sigma(\omega - \alpha z^\star) \\ &< 0 , \end{aligned}$$

thus $z^\star > 0$ is a local maximum. Since $g(0) = 0$, $\lim_{z \to +\infty} g(z) = 0$, $g(z^\star)$ and $z^\star$ is the only positive critical point, this means $z^\star$ is a global maximum. We evaluate g to get the maximum

$$\begin{aligned} g(z^\star) &= \frac{W\left(e^{\omega-1}\right)+1}{\alpha} \frac{1}{1 + \exp\left(W\left(e^{\omega-1}\right)\right) e^{1-\omega}} \\ &= \frac{W\left(e^{\omega-1}\right)+1}{\alpha} \frac{W\left(e^{\omega-1}\right)}{W\left(e^{\omega-1}\right) + W\left(e^{\omega-1}\right) \exp\left(W\left(e^{\omega-1}\right)\right) e^{1-\omega}} \\ &= \frac{W\left(e^{\omega-1}\right)}{\alpha} , \end{aligned}$$

where we used $W\left(e^{\omega-1}\right) \exp\left(W\left(e^{\omega-1}\right)\right) = e^{\omega-1}$ in the third equality. We now upper bound the Lambert function. Taking the log of the equation that defines it, we have $W(x) = \log x - \log W(x)$ for any $x > 0$. Note that $W(e) = 1$ and that W is increasing, so for any $x > e$, we have $W(x) > 1$ and thus $W(x) < \log x$. Using it on $g(z^\star)$, we further have

$$g(z^\star) \leq \frac{\omega - 1}{\alpha} \leq \frac{\omega}{\alpha} ,$$

where we used $\omega \geq 2$. This concludes the proof.

Lemma 12 *Let $n \in \mathbb{R}^n$, and define $\mathrm{LSE} : \mathbb{R}^n \to \mathbb{R}$ the function defined for any $x \in \mathbb{R}^n$ as*

$$\mathrm{LSE}(x) = \log \sum_{i=1}^{n} e^{x_i} .$$

Then LSE is 1-Lipschitz with respect to the norm $\|\cdot\|_\infty$, i.e. for any $x, y \in \mathbb{R}^n$,

$$|\mathrm{LSE}(x) - \mathrm{LSE}(y)| \leq \|x - y\|_\infty .$$

Proof 25 *For any $i \in [n]$ and any $x \in \mathbb{R}^n$, the gradient of* LSE *is given by*

$$\nabla \operatorname{LSE}(x) = \frac{e^x}{\langle e^x, \mathbf{1} \rangle} .$$

Let $y \in \mathbb{R}^n$. By the intermediate mean value theorem, there exists a z on the segment $[x, y]$ such that

$$\begin{aligned} |\operatorname{LSE}(x) - \operatorname{LSE}(y)| &= |\langle \nabla \operatorname{LSE}(z), x - y \rangle| \\ &\leq \|\nabla \operatorname{LSE}(z)\|_1 \|x - y\|_\infty \\ &= \|x - y\|_\infty , \end{aligned}$$

where the inequality follows from Hölder's inequality.

Lemma 13 (Cohen et al., 2019, Lemma 27) *If $0 \prec M \preceq N$ then for any vector v,*

$$\|v\|_N^2 \leq \frac{\det N}{\det M} \|v\|_M^2 .$$

Lemma 14 (Sherman et al., 2023b, Lemma 15) *Let $R, z \geq 1$, then $\beta \geq 2z \log(Rz)$ ensures $\beta \geq z \log(R\beta)$.*

Lemma 15 (Rosenberg et al., 2020, Lemma D.4) *Let $\{X_k\}_{k \in [K]}$ be a sequence of random variables adapted to the filtration $\{\mathcal{F}_k\}_{k \in [K]}$ and suppose that $0 \leq X_k \leq X_{\max}$ almost surely. Then, with probability at least $1 - \delta$, the following holds for all $k \geq 1$ simultaneously*

$$\sum_{k=1}^{K} \mathbb{E}[X_k | \mathcal{F}_{k-1}] \leq 2 \sum_{k=1}^{K} X_k + 4 X_{\max} \log \frac{2K}{\delta} .$$

Lemma 16 (Jin et al., 2019, Lemma D.2) *Let $\{\varphi_t\}_{t \geq 0}$ be a bounded sequence in $\mathbb{R}^d$ satisfying $\sup_{t \geq 0} \|\varphi_t\| \leq 1$. Let $\Lambda_0 \in \mathbb{R}^{d \times d}$ be a positive definite matrix. For any $t \geq 0$, we define $\Lambda_t = \Lambda_0 + \sum_{j=1}^{t} \varphi_j \varphi_j^\intercal$. Then, if the smallest eigenvalue of Λ_0 satisfies $\lambda_{\min}(\Lambda_0) \geq 1$, we have*

$$\sum_{j=1}^{t} \varphi_j \Lambda_{j-1}^{-1} \varphi_j \leq 2 \log \left(\frac{\det \Lambda_t}{\det \Lambda_0} \right) .$$