On the Theory of Reinforcement Learning with Once-per-Episode Feedback

Niladri S. Chatterji[*]
Stanford University
niladri@cs.stanford.edu

Aldo Pacchiano[*]
Microsoft Research
apacchiano@microsoft.com

Peter L. Bartlett
UC Berkeley
peter@berkeley.edu

Michael I. Jordan
UC Berkeley
jordan@cs.berkeley.edu

Abstract

We study a theory of reinforcement learning (RL) in which the learner receives binary feedback only once at the end of an episode. While this is an extreme test case for theory, it is also arguably more representative of real-world applications than the traditional requirement in RL practice that the learner receive feedback at every time step. Indeed, in many real-world applications of reinforcement learning, such as self-driving cars and robotics, it is easier to evaluate whether a learner's complete trajectory was either "good" or "bad," but harder to provide a reward signal at each step. To show that learning is possible in this more challenging setting, we study the case where trajectory labels are generated by an unknown parametric model, and provide a statistically and computationally efficient algorithm that achieves sublinear regret.

1 Introduction

The Reinforcement Learning (RL) paradigm involves a learning agent interacting with an unknown dynamical environment over multiple time steps. The learner receives a reward signal after each step which it uses to improve its performance over time. This formulation of RL has had significant empirical success in the recent past [24, 23, 33, 32].

While this empirical success is encouraging, as RL starts to tackle a more wide-ranging class of consequential real-world problems, such as self-driving cars, supply chains, and medical care, a new set of challenges arise. Foremost among them is the lack of a well-specified reward signal associated with every state-action pair in many real-world settings. For example, consider a robot manipulation task where the robot must fold a pile of clothes. It is not clear how to design a useful reward signal that aids the robot to learn to complete this task. However, it is fairly easy to check whether the task was successfully completed (that is, whether the clothes were properly folded) and provide feedback at the end of the episode.

This is a classical challenge but it is one that is often neglected in theoretical treatments of RL. To address this challenge we introduce a framework for RL that eschews the need for a Markovian reward signal at every step and provides the learner only with binary feedback based on its complete trajectory in an episode. In our framework, the learner interacts with the environment for a fixed number of time steps (H) in each episode to produce a trajectory (τ) which is the collection of all

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

[*]Equal contribution.

states visited and actions taken in these rounds. At the end of the episode a binary reward $y_\tau \in \{0, 1\}$ is drawn from an unknown distribution $\mathbb{Q}(\cdot|\tau)$ and handed to the learner. This protocol continues for N episodes and the learner's goal is to maximize the number of expected binary "successes."

One approach to deal with the lack of a reward function in the literature is Inverse Reinforcement Learning [25], which uses demonstrations of good trajectories to learn a reward function. However, this approach is difficult to use when good demonstrations are either prohibitively expensive or difficult to obtain. Another closely related line of work studies reinforcement learning with preference feedback [2, 15, 3, 5, 37, 26, 38]. Our framework provides the learner with an even weaker form of feedback than that studied in this line of work. Instead of providing preferences between trajectories, we only inform the learner whether the task was completed successfully or not at the end.

To study whether it is possible to learn under such drastically limited feedback we study the case where the conditional rewards (y_τ) are drawn from an unknown logistic model (see Assumption 2.1). Under this assumption we show that learning is possible—we provide an optimism-based algorithm that achieves sublinear regret (see Theorem 3.2). Technically our theory leverages recent results of Russac et al. [31] for the online estimation of the parameters of the underlying logistic model, and combining them with the UCBVI algorithm [4] to obtain regret bounds. Under an explorability assumption we also show that our algorithm is computationally efficient and we provide a dynamic programming algorithm to solve for the optimistic policy at every episode.

We note that Efroni et al. [11] study a similar problem to ours, such that a reward is revealed only at the end of the episode, but they assume that there exists an underlying linear model that determines the reward associated with each state-action pair, and reward revealed to the learner is the sum of rewards over the state-action pairs with added stochastic noise. This assumption ensures that the reward function is Markovian, and allows them to use an online linear bandit algorithm [1] to directly estimate the underlying reward function. This is not possible in our setting since we do not assume the existence of an underlying Markovian reward function. Cohen et al. [6] provided an algorithm that learns in this setting even when the noise is adversarially chosen. An open problem posed by Efroni et al. [11] was to find an algorithm that learns in this setting of reinforcement learning, with once per episode feedback, when the rewards are drawn from an unknown generalized linear model (GLM). In this paper we consider a specific GLM—the logistic model.

The remainder of the paper is organized as follows. In Section 2 we introduce notation and describe our setting. In Section 3 we present our algorithm and main results. Under an explorability assumption we prove that our algorithm is computationally efficient (in Appendix E). Section 4 points to other related work and we conclude with a discussion in Section 5. Other technical details, proofs and experiments are deferred to the appendix.

2 Preliminaries

This section presents notational conventions and a description of the setting.

2.1 Notation

For any $k \in \mathbb{N}$ we denote the set $\{1, \dots, k\}$ by $[k]$. Given any set $\mathcal{T}$, let $\Delta_{\mathcal{T}}$ denote the simplex over this set. Given a vector $\mathbf{v}$, for any $p \in \mathbb{N}$, let $\|\mathbf{v}\|_p$ denote the ℓ_p norm of the vector. Given a vector $\mathbf{v}$ and positive semi-definite matrix $\mathbf{M}$, define $\|\mathbf{v}\|_{\mathbf{M}} := \sqrt{\mathbf{v}^\top \mathbf{M} \mathbf{v}}$. Given a matrix $\mathbf{M}$ let $\|\mathbf{M}\|_{op}$ denote its operator norm. For any positive semi-definite matrix $\mathbf{M}$ we use $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ to denote its maximum and minimum eigenvalues respectively. We will use $C_1, C_2, \dots$ to denote absolute constants whose values are fixed throughout the paper, and $c, c', \dots$ to denote "local" constants, which may take different values in different contexts. We use the standard "big Oh notation" [see, e.g., 7].

2.2 The Setting

We study a Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, H)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathbb{P}(\cdot|s, a)$ is the law that governs the transition dynamics given a state and action pair

(s, a), and $H \in \mathbb{N}$ is the length of an episode. Both the state space $\mathcal{S}$ and action space $\mathcal{A}$ are finite in our paper. The learner's trajectory τ is the concatenation of all states and actions visited during an episode; that is, $\tau := (s_1, a_1, \cdots, s_H, a_H)$. Given any $h \in [H]$ and trajectory τ, a sub-trajectory $\tau_h := (s_1, a_1, \ldots, s_h, a_h)$ is all the states and actions taken up to step h. Also set $\tau_0 := \emptyset$. Let $\tau_{h:H} := (s_h, a_h, \ldots, s_H, a_H)$ denote the states and action from step h until the end of the episode. Let Γ be the set of all possible trajectories τ. Analogously, for any $h \in [H]$ let Γ_h be the set of all sub-trajectories up to step h. At the start of each episode the initial state s_1 is drawn from a fixed distribution ρ that is known to the learner.

At the end of an episode the trajectory τ gets mapped to a feature map $\phi(\tau) \in \mathbb{R}^d$. We also assume that the learner has access to this feature map ϕ. Here are two examples of feature maps:

1. **Direct parametrization:** Without loss of generality assume that $\mathcal{S} = \{1, \ldots, |\mathcal{S}|\}$ and $\mathcal{A} = \{1, \ldots, |\mathcal{A}|\}$. The feature map $\phi(\tau) = \sum_{h=1}^{H} \phi_h(s_h, a_h)$, where the per-step maps $\phi_h(s, a) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|H}$ are defined as follows:

$$(\phi_h(s,a))_j = \begin{cases} 1 & \text{if } j = (h-1)|\mathcal{S}||\mathcal{A}| + (s-1)|\mathcal{A}| + a, \\ 0 & \text{otherwise.} \end{cases}$$

   The complete feature map $\phi(\tau) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|H}$ is therefore an encoding of the trajectory τ.

2. **Reduced parametrization:** Any trajectory τ is associated with a feature $\phi(\tau) \in \mathbb{R}^d$, where $d < |\mathcal{S}||\mathcal{A}|H$.

After the completion of an episode the learner is given a random binary reward $y_\tau \in \{0, 1\}$. Let $\mathbf{w}_\star \in \mathbb{R}^d$ be a vector that is unknown to the learner. We study the case where the rewards are drawn from a binary logistic model as described below.

Assumption 2.1 (Logistic model)**.** *Given any trajectory $\tau \in \Gamma$, the rewards are said to be drawn from a logistic model if the law of $y_\tau|\tau$ is*

$$y_\tau|\tau = \begin{cases} 1 & \text{w.p.} \quad \mu\left(\mathbf{w}_\star^\top \phi(\tau)\right) \\ 0 & \text{w.p.} \quad 1 - \mu\left(\mathbf{w}_\star^\top \phi(\tau)\right), \end{cases} \tag{1}$$

where for any $z \in \mathbb{R}$, $\mu(z) = \frac{1}{1+\exp(-z)}$ is the logistic function. We shall refer to $\mathbf{w}_\star$ as the "reward parameters."

We make the following boundedness assumptions on the features and reward parameters.

Assumption 2.2 (Bounded features and parameters)**.** *We assume that*

- $\|\mathbf{w}_\star\|_2 \leq B$ *for some known value* $B > 0$ *and*
- *for all* $\tau \in \Gamma$, $\|\phi(\tau)\|_2 \leq 1$.

We note that such boundedness assumptions are standard in the logistics bandits literature [13, 31, 14].

A policy π is a collection of per-step policies $(\pi_1, \ldots, \pi_H)$ such that

$$\pi_h : \Gamma_{h-1} \times \mathcal{S} \to \Delta_{\mathcal{A}}.$$

If the agent is using the policy π then at round h of the episode the learner plays according to the policy π_h. We let Π_h denote the set of all valid policies at step h and let Π denote the set of valid policies over the trajectory. Let $\mathbb{P}^\pi(\cdot|s_1)$ denote the joint probability distribution over the learner's trajectory τ and the reward y_τ when the learner plays according to the policy π and the initial state is s_1. Often when the initial state is clear from the context we will refer to $\mathbb{P}^\pi(\cdot|s_1)$ by simply writing $\mathbb{P}^\pi$. Also with some abuse of notation we will sometimes let $\mathbb{P}^\pi$ denote the distribution of the trajectory and the reward where the initial state is drawn from the distribution ρ.

Given an initial state $s \in \mathcal{S}$ the value function corresponding to a policy π is

$$V^\pi(s) := \mathbb{E}_{y_\tau, \tau \sim \mathbb{P}^\pi}\left[y_\tau \mid s_1 = s\right] = \mathbb{E}_{\tau \sim \mathbb{P}^\pi}\left[\mu\left(\mathbf{w}_\star^\top \phi(\tau)\right) \;\middle|\; s_1 = s\right],$$

where the second equality follows as the mean of y_τ conditioned on τ is $\mu(\mathbf{w}_\star^\top \phi(\tau))$. With some abuse of notation we denote the average value function as $V^\pi := \mathbb{E}_{s_1 \sim \rho}\left[V^\pi(s_1)\right]$.

Define the optimal policy as $\pi_\star \in \arg\max_{\pi \in \Pi} V^\pi$. It is worth noting that in our setting the optimal policy may be *non-Markovian*. The learner plays for a total of N episodes. The policy played in episode $t \in [N]$ is $\pi^{(t)}$ and its value function is $V^{(t)} := V^{\pi^{(t)}}$. Also define the value function for the optimal policy to be $V_\star := V^{\pi_\star}$. Our goal shall be to control the regret of the learner, which is defined as

$$\mathcal{R}(N) := \sum_{t=1}^{N} V_\star - V^{(t)}. \tag{2}$$

The trajectories in these N episodes are denoted by $\{\tau^{(t)}\}_{t=1}^N$ and rewards received are denoted by $\{y^{(t)}\}_{t=1}^N$.

3 Optimistic Algorithms that Use Trajectory Labels

We now present an algorithm to learn from labeled trajectories. Throughout this section we assume that both Assumptions 2.1 and 2.2 are in force.

The derivative of the logistic function is $\mu'(z) = \frac{\exp(-z)}{(1+\exp(-z))^2}$, and therefore, μ is $1/4$-Lipschitz. The following quantity will play an important role in our bounds

$$\kappa := \max_{\tau \in \Gamma} \sup_{\mathbf{w}:\|\mathbf{w}\| \le B} \frac{1}{\mu'(\mathbf{w}^\top \phi(\tau))}.$$

A consequence of Assumption 2.2 is that $\kappa \le \exp(B)$. We briefly note that κ is a measure of curvature of the logistic model. It also plays an important role in the analysis of logistic bandit algorithms [13, 31].

Since the true reward parameter $\mathbf{w}_\star$ is unknown we will estimate it using samples. At any episode $t \in [N]$, a natural way of computing an estimator of $\mathbf{w}_\star$, given past trajectories $\{\tau^{(q)}\}_{q \in [t-1]}$ and labels $\{y^{(q)}\}_{q \in [t-1]}$, is by minimizing the ℓ_2-regularized cross-entropy loss:

$$\mathcal{L}_t(\mathbf{w}) := -\sum_{q=1}^{t-1} y^{(q)} \log\left(\mu\left(\mathbf{w}^\top \phi(\tau^{(q)})\right)\right) - (1-y^{(q)}) \log\left(1-\mu\left(\mathbf{w}^\top \phi(\tau^{(q)})\right)\right) + \frac{\|\mathbf{w}\|_2^2}{2}.$$

This function is strictly convex and its minimizer is defined to be

$$\widehat{\mathbf{w}}_t := \arg\min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}_t(\mathbf{w}). \tag{3}$$

Define a design matrix at every episode

$$\mathbf{\Sigma}_1 := \kappa \mathbf{I}, \quad \text{and} \quad \mathbf{\Sigma}_t := \kappa \mathbf{I} + \sum_{q=1}^{t-1} \phi(\tau^{(q)}) \phi(\tau^{(q)})^\top, \quad \text{for all } t \ge 1.$$

Further, define the confidence radius $\beta_t(\delta)$ as follows

$$\beta_t(\delta) := \left(1 + B + \rho_t(\delta)\left(\sqrt{1+B} + \rho_t(\delta)\right)\right)^{3/2} \tag{4}$$

$$\text{where,} \quad \rho_t(\delta) := d \log\left(4 + \frac{4t}{d}\right) + 2\log\left(\frac{N}{\delta}\right) + \frac{1}{2}.$$

We adapt a result due to Russac et al. [31, Proposition 7] who studied the online logistic bandits problem to establish that at every episode and every trajectory the difference between $\mu(\mathbf{w}_\star^\top \phi(\tau))$ and $\mu(\widehat{\mathbf{w}}_t^\top \phi(\tau))$ is small.

Lemma 3.1. *For any $\delta \in (0,1]$, define the event*

$$\mathcal{E}_\delta := \left\{ \textit{for all } t \in [N], \tau \in \Gamma : \left|\mu(\mathbf{w}_\star^\top \phi(\tau)) - \mu(\widehat{\mathbf{w}}_t^\top \phi(\tau))\right| \le \sqrt{\kappa}\beta_t(\delta) \|\phi(\tau)\|_{\mathbf{\Sigma}_t^{-1}} \right\}. \tag{5}$$

Then $\mathbb{P}(\mathcal{E}_\delta) \ge 1 - \delta$.

We provide a proof in Appendix B.2. The proof follows by simply translating [31, Proposition 7] into our setting. We note that we specifically adapt these recent results by Russac et al. [31] since they directly apply to $\widehat{\mathbf{w}}_t$, the minimizer of the ℓ_2-regularized cross-entropy loss. In contrast, previous work on the logistic bandits problem [see, e.g., 14, 13] established confidence sets for an estimator that was obtained by performing a non-convex (and potentially computationally intractable) projection of $\widehat{\mathbf{w}}_t$ onto the ball of Euclidean radius B.

Our algorithm shall construct an estimate of the transition dynamics $\widehat{\mathbb{P}}_t$. Let $N_t(s,a)$ be the number of times that the state-action pair (s,a) is encountered before the start of episode t, and let $N_t(s';s,a)$ be the number of times the learner encountered the state s' after taking action a at state s before the start of episode t. Define the estimator of the transition dynamics as follows:

$$\widehat{\mathbb{P}}_t(s'|a,s) := \frac{N_t(s';s,a)}{N_t(s,a)}. \tag{6}$$

Also define the state-action bonus at episode t

$$\xi_{s,a}^{(t)} := \min\left\{2, 4\sqrt{\frac{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8NH^2)^{|\mathcal{S}|}\log(N_t(s,a))}{\delta}\right)}{N_t(s,a)}}\right\}. \tag{7}$$

In this definition whenever $N_t(s,a)=0$, that is, when a state-action pair hasn't been visited yet, we define $\xi_{s,a}^{(t)}$ to be equal to 2. Finally, we define the optimistic reward functions

$$\bar{\mu}_t(\mathbf{w},\tau) := \min\left\{\mu\left(\mathbf{w}^\top\phi(\tau)\right) + \sqrt{\kappa}\beta_t(\delta)\|\phi(\tau)\|_{\mathbf{\Sigma}_t^{-1}}, 1\right\} \quad \text{and} \tag{8a}$$

$$\widetilde{\mu}_t(\mathbf{w},\tau) := \bar{\mu}_t(\mathbf{w},\tau) + \sum_{h=1}^{H-1}\xi_{s_h,a_h}^{(t)}. \tag{8b}$$

The first reward function $\bar{\mu}_t$ is defined as above to account for the uncertainty in the predicted value of $\mathbf{w}_\star$ in light of Lemma 3.1, and the second reward function $\widetilde{\mu}_t$ is designed to account for the error in the estimation of the transition dynamics $\mathbb{P}$. With these additional definitions in place we are ready to present our algorithms and main results.

3.1 UCBVI with Trajectory Labels

Our first algorithm is an adaptation of the UCBVI algorithm [4] to our setting with labeled trajectories.

Algorithm 1: UCBVI with trajectory labels.

Input: State and action spaces $\mathcal{S}, \mathcal{A}$.

Initialize $\widehat{\mathbb{P}}_1 = \mathbf{0}$, visitation set $\mathcal{K} = \emptyset$.

for $t = 1, \cdots$ **do**

1. Calculate the $\widehat{\mathbf{w}}_t$ by solving equation (3).
2. If $t > 1$, compute $\pi^{(t)}$

$$\pi^{(t)} \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}_t^\pi(\cdot|s_1)}\left[\widetilde{\mu}_t(\widehat{\mathbf{w}}_t,\tau)\right]. \tag{9}$$

   Else for all $h, s, \tau_{h-1} \in [H]\times\mathcal{S}\times\Gamma_{h-1}$, set $\pi_h^{(1)}(\cdot|s,\tau_{h-1})$ to be the uniform distribution over the action set.
3. Observe the trajectory $\tau^{(t)} \sim \mathbb{P}^{\pi^{(t)}}$ and update the design matrix

$$\mathbf{\Sigma}_{t+1} = \kappa\mathbf{I} + \sum_{q=1}^{t}\phi(\tau^{(q)})\phi(\tau^{(q)})^\top. \tag{10}$$

4. Update the visitation set $\mathcal{K} = \{(s,a)\in\mathcal{S}\times\mathcal{A} : N_t(s,a) > 0\}$.
5. For all $(s,a)\in\mathcal{K}$, update $\widehat{\mathbb{P}}_{t+1}(\cdot|s,a)$ according to equation (6).
6. For all $(s,a)\notin\mathcal{K}$, set $\widehat{\mathbb{P}}_{t+1}(\cdot|s,a)$ to be the uniform distribution over states.

Theorem 3.2. *For any $\bar{\delta} \in (0,1]$, set $\delta = \bar{\delta}/(6N)$ then under Assumptions 2.1 and 2.2 the regret of Algorithm 1 is upper bounded as follows:*

$$\mathcal{R}(N) \leq \widetilde{O}\left(\left[H\sqrt{(H+|\mathcal{S}|)|\mathcal{S}||\mathcal{A}|} + H^2 + \sqrt{\kappa d}(d^3 + B^{3/2})\right]\sqrt{N} + (H+|\mathcal{S}|)H|\mathcal{S}||\mathcal{A}|\right),$$

with probability at least $1-\bar{\delta}$.

The regret of our algorithm scales with $\sqrt{N}$ and polynomially with the horizon, number of states, number of actions, κ, dimension of the feature maps and length of the reward parameters (B). The minimax regret in the standard episodic reinforcement learning is $O(\sqrt{H|\mathcal{S}||\mathcal{A}|N})$ [27, 4]. Here we pay for additional factors in H, $|\mathcal{S}|$ and κ since our rewards are non-Markovian and are revealed to the learner only at the end of the episode. We provide a proof of this theorem in Appendix B. For a more detailed bound on the regret with the logarithmic factors and constants specified we point the interested reader to inequality (41) in the appendix.

Proof sketch. First we show that with high probability at each episode the value function of the optimal policy $V_\star$ is upper bounded by $\widetilde{V}^{(t)} := \mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}_t^{\pi^{(t)}}(\cdot|s_1)}\left[\widetilde{\mu}_t(\widehat{\mathbf{w}}_t,\tau)\right]$ (the value function of the policy $\pi^{(t)}$ when the rewards are dictated by $\widetilde{\mu}_t$ and the transition dynamics are given by $\widehat{\mathbb{P}}_t$). Then we provide a high probability bound on the difference between the optimistic value function $\widetilde{V}^{(t)}$ and the true value function $V^{(t)}$ to obtain our upper bound on the regret. In both of these steps we need to relate expectations with respect to the true transition dynamics $\mathbb{P}$ to expectations with respect to the empirical estimate of the transition dynamics $\widehat{\mathbb{P}}_t$. We do this by using our concentration results: Lemmas B.1 and B.2 proved in the appendix. While analogs of these concentration lemmas do exist in previous theoretical studies of episodic reinforcement learning, here we had to prove these lemmas in our setting with non-Markovian trajectory-level feedback (which explains why we pay extra factors in H and $|\mathcal{S}|$).

3.2 UCBVI with Added Exploration

Although the regret of Algorithm 1 is sublinear it is not guaranteed to be computationally efficient since finding the optimistic policy $\pi^{(t)}$ (in equation (9)) at every episode might prove to be difficult. In this section, we will show that when the features are sum-decomposable and the MDP satisfies an explorability assumption then it will be possible to find a computationally efficient algorithm with sublinear regret (albeit with a slightly worse scaling with the number of episodes N).

Assumption 3.3 (Sum-decomposable features)**.** *We assume that the feature maps $\phi \in \mathbb{R}^d$ are sum-decomposable over the different steps of the trajectory, that is, $\phi(\tau) = \sum_{h=1}^H \phi_h(s_h, a_h)$.*

Under this assumption, given any $\mathbf{w} \in \mathbb{R}^d$ and any trajectory $\tau \in \Gamma$, $\mathbf{w}^\top\phi(\tau) = \sum_{h=1}^H \mathbf{w}^\top\phi_h(s_h,a_h)$. We stress that even under this sum-decomposablity assumption, the optimal policy is potentially non-Markovian due to the presence of the logistic map that governs the reward.

We also make the following explorability assumption.

Assumption 3.4 (Explorability)**.** *For any $s, s' \in \mathcal{S}$, $a, a' \in \mathcal{A}$, and $h \neq h' \in [H]$, suppose that*

$$\phi_h(s,a)^\top\phi_{h'}(s',a') = 0.$$

Further assume that there exists $\omega \in (0,1)$ such that for any unit vector $\mathbf{v} \in \mathbb{R}^d$ we have that

$$\sup_{\pi\in\Pi} \mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^\pi}\left[\sum_{h\in[H]} \mathbf{v}^\top\phi_h(s_h,a_h)\right] \geq \omega.$$

In a setting with Markovian rewards a similar assumption has been made previously by Zanette et al. [40]. This assumption allows us to efficiently "explore" the feature space, and construct a sum-decomposable bonus $\sqrt{\kappa}\beta_t(\delta)\sum_{h=1}^H \|\phi_h(s_h,a_h)\|_{\mathbf{\Sigma}_t^{-1}}$ that we will use instead of

$\sqrt{\kappa}\beta_t(\delta)\|\phi(\tau)\|_{\mathbf{\Sigma}_t^{-1}}$ in the definition of $\bar{\mu}_t$ (see equation (8a)). Define the reward functions

$$\bar{\mu}_t^{\mathsf{sd}}(\mathbf{w}, \tau) := \min\left\{\mu\left(\mathbf{w}^\top \phi(\tau)\right) + \sqrt{\kappa}\beta_t(\delta)\sum_{h=1}^{H}\|\phi_h(s_h, a_h)\|_{\mathbf{\Sigma}_t^{-1}}, 1\right\} \quad \text{and} \tag{11a}$$

$$\widetilde{\mu}_t^{\mathsf{sd}}(\mathbf{w}, \tau) := \bar{\mu}_t^{\mathsf{sd}}(\mathbf{w}, \tau) + \sum_{h=1}^{H-1}\xi_{s_h, a_h}^{(t)}. \tag{11b}$$

To prove a regret bound for an algorithm that uses these rewards our first step shall be to prove that the sum-decomposable bonus also leads to an optimistic reward function (that is, the value function defined by these rewards sufficiently over-estimates the true value function). To this end, we will first use Algorithm 2 to find an exploration mixture policy $\bar{U}$ and play according to it at episode t with probability $1/t^{1/3}$. This policy $\bar{U}$ will be such that the minimum eigenvalue of

$$\mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^{\bar{U}}(\cdot|s_1)}\left[\phi(\tau)\phi(\tau)^\top\right] \tag{12}$$

is lower bounded by a function of d, ω and N (see Lemma 3.5). This property shall allow us to upper bound the condition number of the design matrix $\mathbf{\Sigma}_t$ and subsequently ensure that the rewards $\bar{\mu}_t^{\mathsf{sd}}$ and $\widetilde{\mu}_t^{\mathsf{sd}}$ are optimistic. Given a unit vector $\mathbf{v}$ define a reward function at step h as follows:

$$r_h^{\mathbf{v}}(s, a) := \mathbf{v}^\top \phi_h(s, a). \tag{13}$$

Let $r^{\mathbf{v}} := (r_1^{\mathbf{v}}, \ldots, r_H^{\mathbf{v}})$ be a reward function over the entire episode. As a subroutine Algorithm 2 uses the EULER algorithm [39]. (We briefly note that other reinforcement learning algorithms with PAC or regret guarantees [e.g., 4, 19] could also be used here in place of EULER.)

Algorithm 2: Find exploration mixture.

Input: Initial unit vector $\mathbf{v}_1$, Exploration lower bound ω, number of EULER episodes N_{EUL}, number of evaluation episodes N_{EVAL}.
Initialize: $\mathbf{A}_0 = \frac{\omega^2}{16}\mathbf{I}$, $n = 0$ and $\lambda_{\min} = \inf_{\mathbf{z}\in\mathbb{R}^d}\mathbf{z}^\top\mathbf{A}_0\mathbf{z}$.
while $\lambda_{\min} < \frac{\omega^2}{8}$ **do**
  Update the counter $n \leftarrow n + 1$.
  Set $U_n \leftarrow \mathsf{EULER}(\{r^{\mathbf{v}_n}, N_{\mathsf{EUL}}\})$ //`run EULER for` N_{EUL} `episodes`.
  for *t=1,...,*N_{EVAL} *episodes* **do**
    Sample a trajectory $\tau_n^{(t)} \sim \rho \times \mathbb{P}^{U_n}$.
  Calculate the average feature $\widehat{\mathbf{a}}_n = \sum_{t=1}^{N_{\mathsf{EVAL}}}\phi(\tau_n^{(t)})/N_{\mathsf{EVAL}}$.
  Update the matrix $\mathbf{A}_n \leftarrow \mathbf{A}_{n-1} + \widehat{\mathbf{a}}_n\widehat{\mathbf{a}}_n^\top$.
  Update the minimum eigenvalue: $\lambda_{\min} \leftarrow \inf_{\mathbf{z}\in\mathbb{R}^d}\mathbf{z}^\top\mathbf{A}_n\mathbf{z}$.
  Set $\mathbf{v}_n$ to be the minimum eigenvector of $\mathbf{A}_n$.
Set $n_{\mathsf{loop}} = n$.
Return: (i) $\bar{U} = \mathsf{Unif}(U_1, \cdots, U_{n_{\mathsf{loop}}})$ //`the uniform mixture over the policies;`
(ii) $N_{\mathsf{exp}} = n_{\mathsf{loop}} \times (N_{\mathsf{EUL}} + N_{\mathsf{EVAL}})$ //`total number of episodes.`

Lemma 3.5. *There exist positive absolute constants C_1 and C_2 such that, under Assumptions 2.2, 3.3 and 3.4, if Algorithm 2 is run with $N_{\mathsf{EUL}} = \frac{C_1|\mathcal{S}|^2|\mathcal{A}|H^2\log\left(\frac{|\mathcal{S}||\mathcal{A}|N^2d}{\delta\omega^2}\right)}{\omega^2}$ and $N_{\mathsf{EVAL}} = \frac{C_2d^3\log^3\left(\frac{Nd^2}{\delta\omega^2}\right)}{\omega^4}$, and $N > \frac{d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}(N_{\mathsf{EUL}} + N_{\mathsf{EVAL}}) =: \bar{N}_{\mathsf{exp}}$ then, with probability at least $1 - 2\delta$, we have $N_{\mathsf{exp}} \le \bar{N}_{\mathsf{exp}}$ and furthermore:*

$$\mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^{\bar{U}}(\cdot|s_1)}\left[\phi(\tau)\phi(\tau)^\top\right] \succeq \frac{\omega^2\log(3/2)}{32d\log\left(d\log\left(1+\frac{16N}{d\omega^2}\right)\right)}\mathbf{I}.$$

This lemma is proved in Appendix C. With this lemma in place we now present our modified algorithm under the explorability assumption. In the first few episodes this algorithm finds the exploration mixture policy $\bar{U}$. In a subsequent episode t this algorithm acts according to the policy $\pi^{(t)}$ which maximizes the value function associated with the rewards $\widetilde{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t, \tau)$ with probability $1 - \frac{1}{t^{1/3}}$. Otherwise it uses the exploration mixture policy $\bar{U}$.

Algorithm 3: UCBVI with trajectory labels and added exploration.

Input: State and action spaces $\mathcal{S}, \mathcal{A}$, Initial unit vector $\mathbf{v}_1$, Exploration lower bound ω, number of EULER episodes N_{EUL}, number of evaluation episodes N_{EVAL}.

Initialize $\widehat{\mathbb{P}}_1 = \mathbf{0}$, visitation set $\mathcal{K} = \emptyset$.

Find exploration mixture policy $\bar{U}$ in $N_{\exp}$ episodes by running Algorithm 2.

for $t = N_{\mathsf{exp}} + 1, \cdots, N$ **do**

1. Calculate $\widehat{\mathbf{w}}_t$ by solving equation (3).
2. If $t > N_{\mathsf{exp}} + 1$, compute $\pi^{(t)}$

$$\pi^{(t)} \in \arg\max_{\pi} \mathbb{E}_{s_1 \sim \rho,\ \tau \sim \widehat{\mathbb{P}}_t^{\pi}(\cdot|s_1)} \left[\widetilde{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t, \tau)\right]. \tag{14}$$

   Else for all $h, s, \tau_{h-1} \in [H] \times \mathcal{S} \times \Gamma_{h-1}$, set $\pi_h^{(1)}(\cdot|s, \tau_{h-1})$ to be the uniform distribution over the action set.
3. Sample $b_t = \begin{cases} 0 & \text{w.p. } 1 - \frac{1}{t^{1/3}}, \\ 1 & \text{w.p. } \frac{1}{t^{1/3}}. \end{cases}$
4. If $b_t = 1$ then set $\pi^{(t)} \leftarrow \bar{U}$.
5. Observe the trajectory $\tau^{(t)} \sim \mathbb{P}^{\pi^{(t)}}$ and update the design matrix

$$\mathbf{\Sigma}_{t+1} = \kappa \mathbf{I} + \sum_{q=N_{\mathsf{exp}}+1}^{t} \phi(\tau^{(q)}) \phi(\tau^{(q)})^{\top}. \tag{15}$$

6. Update the visitation set $\mathcal{K} = \{(s, a) \in \mathcal{S} \times \mathcal{A} : N_t(s, a) > 0\}$.
7. For all $(s, a) \in \mathcal{K}$, update $\widehat{\mathbb{P}}_{t+1}(\cdot|s, a)$ according to equation (6).
8. For all $(s, a) \notin \mathcal{K}$, set $\widehat{\mathbb{P}}_{t+1}(\cdot|s, a)$ to be the uniform distribution over states.

The following is our regret bound for Algorithm 3.

Theorem 3.6. *For any $\bar{\delta} \in (0, 1]$, set $\delta = \bar{\delta}/(12N)$. Under Assumptions 2.1, 2.2, 3.3 and 3.4, and for all $N > \bar{N}_{\mathsf{exp}}$ (see its definition in Lemma 3.5) if Algorithm 3 is run with the parameters N_{EUL} and N_{EVAL} set as specified in Lemma 3.5 then its regret is upper bounded as follows:*

$$\mathcal{R}(N) \leq \widetilde{O}\left(\frac{\sqrt{\kappa H} d}{\omega}(d^3 + B^{3/2})N^{2/3} + \left[H\sqrt{(H + |\mathcal{S}|)|\mathcal{S}||\mathcal{A}|} + H^2\right]\sqrt{N} \right.$$
$$\left. + (H + |\mathcal{S}|)H|\mathcal{S}||\mathcal{A}| + \frac{d^2}{\omega^2}\left(\frac{d^2}{\omega^2} + |\mathcal{S}|^2|\mathcal{A}|H^2\right)\right),$$

with probability at least $1 - \bar{\delta}$.

The proof of Theorem 3.6 is in Appendix D. For a more detailed bound on the regret with the logarithmic factors and constants specified we point the interested reader to inequality (58) in the appendix. The bound on the regret of this algorithm scales with $N^{2/3}$ up to poly-logarithmic factors. This is larger than the $\sqrt{N}$ regret bound (again up to poly-logarithmic factors) that we proved above for Algorithm 1 since here the learner plays according to the exploration policy $\bar{U}$ with probability $1/t^{1/3}$ throughout the run of the algorithm. However, the next proposition shows that by using the sum-decomposable reward function $\widetilde{\mu}_t^{\mathsf{sd}}$ the policy $\pi^{(t)}$ defined in equation (14) can be efficiently approximated.

Proposition 3.7. *For any $t \in [N]$ define $\widetilde{V}_t^{\mathsf{sd}}(\pi) := \mathbb{E}_{s_1 \sim \rho,\ \tau \sim \widehat{\mathbb{P}}_t^{\pi}(\cdot|s_1)} \left[\widetilde{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t, \tau)\right]$. Given any $\varepsilon > 0$, under Assumptions 2.2, 3.3 and 3.4 it is possible to find a policy $\widehat{\pi}^{(t)}$ that satisfies*

$$\widetilde{V}_t^{\mathsf{sd}}(\pi^{(t)}) - \widetilde{V}_t^{\mathsf{sd}}(\widehat{\pi}^{(t)}) \leq \varepsilon,$$

using at most $\mathsf{poly}\left(|\mathcal{S}|, |\mathcal{A}|, H, d, B, \|\widehat{\mathbf{w}}_t\|_2, \frac{1}{\varepsilon}, \log\left(\frac{N}{\delta}\right)\right)$ *time and memory.*

We describe the approximate dynamic programming algorithm that can be used to find this policy $\widehat{\pi}^{(t)}$ and present a proof of this proposition in Appendix E. We also note that if we use an

ε-approximate policy $\widehat{\pi}^{(t)}$ instead of $\pi^{(t)}$ in Algorithm 3 then its regret increases by an additive factor of at most εN. (It is possible to easily check this by inspecting the proof of Theorem 3.6.) Thus, for example a choice of $\varepsilon = 1/N^{1/3}$ ensures that the regret of Algorithm 3 is bounded by $O(N^{2/3})$ with high probability if the approximate policy $\widehat{\pi}^{(t)}$ (which can be found efficiently) is used instead.

4 Additional Related Work

There have been many theoretical results that analyze regret minimization in standard episodic reinforcement [18, 29, 16, 28, 4, 19, 8, 39, 34, 10, 30]. Recently Efroni et al. [12] introduced a framework of "sequential budgeted learning" which includes as a special case the setting of episodic reinforcement learning with the constraint that the learner is allowed to query the reward function only a limited number of times per episode. They show learning is possible in this setting by using a modified UCBVI algorithm.

As stated above to estimate the reward parameter we rely on the recent results by Russac et al. [31] who in term built on earlier work [13, 14] that analyzed the GLM-UCB algorithm. Dong et al. [9] provided and analyzed a Thompson sampling approach for the logistic bandits problem.

5 Discussion

We have shown that efficient learning is possible when the rewards are non-Markovian and delivered to the learner only once per episode. It would be interesting to see if one can establish guarantees under more general reward models than the logistic model that we study here. Another interesting question is if faster rates of learning are possible when the learner obtains ranked trajectories (that is, moving beyond binary labels).

Acknowledgments

The authors would like to thank Louis Faury, Tor Lattimore, Yoan Russac and Csaba Szepesvári for helpful conversations regarding the literature on logistic bandits. We thank Yonathan Efroni, Nadav Merlis and Shie Mannor for pointing us to prior related work.

Funding

We gratefully acknowledge the support of the NSF through the grant DMS-2023505 in support of the FODSI Institute.

References

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

[2] R. Akrour, M. Schoenauer, and M. Sebag. Preference-based policy learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 12–27, 2011.

[3] R. Akrour, M. Schoenauer, M. Sebag, and J.-C. Souplet. Programming by feedback. In *International Conference on Machine Learning*, pages 1503–1511, 2014.

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

[5] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, page 4299–4307, 2017.

[6] A. Cohen, H. Kaplan, T. Koren, and Y. Mansour. Online Markov decision processes with aggregate bandit feedback. *arXiv preprint arXiv:2102.00490*, 2021.

[7] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT Press, 2009.

[8] C. Dann, T. Lattimore, and E. Brunskill. Unifying PAC and regret: uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5717–5727, 2017.

[9] S. Dong, T. Ma, and B. Van Roy. On the performance of Thompson sampling on logistic bandits. In *Conference on Learning Theory*, pages 1158–1160, 2019.

[10] Y. Efroni, N. Merlis, M. Ghavamzadeh, and S. Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pages 12224–12234, 2019.

[11] Y. Efroni, N. Merlis, and S. Mannor. Reinforcement learning with trajectory feedback. In *AAAI Conference on Artificial Intelligence*, pages 7288–7295, 2021.

[12] Y. Efroni, N. Merlis, A. Saha, and S. Mannor. Confidence-budget matching for sequential budgeted learning. *arXiv preprint arXiv:2102.03400*, 2021.

[13] L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060, 2020.

[14] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

[15] J. Fürnkranz, E. Hüllermeier, W. Cheng, and S.-H. Park. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine learning*, 89(1-2): 123–156, 2012.

[16] A. Gopalan and S. Mannor. Thompson sampling for learning parameterized Markov decision processes. In *Conference on Learning Theory*, pages 861–898, 2015.

[17] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[18] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[19] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4868–4878, 2018.

[20] C. Jin, A. Krishnamurthy, M. Simchowitz, and T. Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879, 2020.

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

[23] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[25] A. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, pages 663–670, 2000.

[26] E. Novoseller, Y. Wei, Y. Sui, Y. Yue, and J. Burdick. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 1029–1038, 2020.

[27] I. Osband and B. Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

[28] I. Osband and B. Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*, pages 2701–2710, 2017.

[29] I. Osband, D. Russo, and B. Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

[30] A. Pacchiano, P. Ball, J. Parker-Holder, K. Choromanski, and S. Roberts. On optimism in model-based reinforcement learning. *arXiv preprint arXiv:2006.11911*, 2020.

[31] Y. Russac, L. Faury, O. Cappé, and A. Garivier. Self-concordant analysis of generalized linear bandits with forgetting. In *International Conference on Artificial Intelligence and Statistics*, pages 658–666, 2021.

[32] A. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[33] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550 (7676):354–359, 2017.

[34] M. Simchowitz and K. G. Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, pages 1153–1162, 2019.

[35] J. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

[36] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[37] C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

[38] Y. Xu, R. Wang, L. Yang, A. Singh, and A. Dubrawski. Preference-based reinforcement learning with finite-time guarantees. In *Advances in Neural Information Processing Systems*, pages 18784–18794, 2020.

[39] A. Zanette and E. Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

[40] A. Zanette, A. Lazaric, M. J. Kochenderfer, and E. Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. *arXiv preprint arXiv:2008.07737*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All assumptions are stated clearly with the accompanying theoretical results.
   (b) Did you include complete proofs of all theoretical results? [Yes] The complete proofs of all theoretical results are carefully proved in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section F.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section F.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 1.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Contents

A Technical Lemmas

In this section we collect some useful technical results used in the proofs that follow. First we present a time-uniform martingale concentration inequality.

Lemma A.1. *Let $\{x_t\}_{t=1}^{\infty}$ be a martingale difference sequence with $|x_t| \leq \zeta$ and let $\delta \in (0,1]$. Then with probability $1-\delta$ for all $T \in \mathbb{N}$*

$$\sum_{t=1}^{T} x_t \leq 2\zeta\sqrt{T\log\left(\frac{6\log T}{\delta}\right)}.$$

Proof Observe that $\frac{|x_t|}{\zeta} \leq 1$. By invoking a time-uniform Hoeffding-style concentration inequality [17, Equation (11)] we find that

$$\mathbb{P}\left[\forall\, t \in \mathbb{N} \;:\; \sum_{t=1}^{T} \frac{x_t}{\zeta} \leq 1.7\sqrt{T\left(\log\log(T) + 0.72\log\left(\frac{5.2}{\delta}\right)\right)}\right] \geq 1-\delta.$$

Rounding up the constants for the sake of simplicity we get

$$\mathbb{P}\left[\forall\, t \in \mathbb{N} \;:\; \sum_{t=1}^{T} x_t \leq 2\zeta\sqrt{T\left(\log\left(\frac{6\log(T)}{\delta}\right)\right)}\right] \geq 1-\delta,$$

which establishes our claim. ■

Next we state a matrix concentration theorem [35, Theorem 1.1].

Theorem A.2 (Matrix Freedman inequality)**.** *Consider a matrix martingale $\{\mathbf{Y}_k\}_{k=1}^{\infty}$ whose values are self adjoint matrices with dimension d and let $\{\mathbf{X}_k\}_{k=1}^{\infty}$ be its difference sequence. Assume the difference sequence is uniformly bounded in the sense that:*

$$\lambda_{\max}(\mathbf{X}_k) \leq R \quad \text{almost surely for all } k = 1, 2\ldots.$$

Define the predictable quadratic variation process of the martingale

$$\mathbf{W}_k := \sum_{j=1}^{k} \mathbb{E}\left[\mathbf{X}_j^2 \mid \mathbf{X}_1, \ldots, \mathbf{X}_{j-1}\right] \quad \text{for} \quad k = 1, 2, \ldots.$$

Then for all $x \geq 0$ and $V \geq 0$,

$$\mathbb{P}\left(\exists\, k : \lambda_{\max}(\mathbf{Y}_k) \geq x \text{ and } \|\mathbf{W}_k\|_{op} \leq V\right) \leq d \cdot \exp\left(\frac{-x^2/2}{V + Rx/3}\right).$$

The following result that bounds the norm of sequence of vectors in terms of the norm induced by its inverse Gram matrix.

Lemma A.3 (Determinant Lemma)**.** *For any sequence of vectors $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)} \in \mathbb{R}^d$ such that $\|\boldsymbol{x}^{(q)}\|_2 \leq L$ for all $q \in [T]$. Given a $\lambda \geq 0$ define $\bar{\boldsymbol{\Sigma}}_1 := \lambda\mathbf{I}$ and for $t \in \{2, \ldots, T\}$ define $\bar{\boldsymbol{\Sigma}}_t := \lambda\mathbf{I} + \sum_{q=1}^{t-1} \boldsymbol{x}^{(q)}\boldsymbol{x}^{(q)\top}$. Then for all $T \in \mathbb{N}$*

$$\sum_{t=1}^{N} \|\phi(\tau^{(t)})\|^2_{\boldsymbol{\Sigma}_t^{-1}} \leq 2d\max\left\{1, \frac{1}{\lambda}\right\}\log\left(1 + \frac{TL^2}{\lambda d}\right)$$

and

$$\log\left(\frac{\det(\bar{\boldsymbol{\Sigma}}_t)}{\det(\lambda\mathbf{I})}\right) \leq d\log\left(1 + \frac{tL^2}{\lambda d}\right).$$

Proof The first part follows by combining the results of Lemmas 15 and 16 from [13]. The second part is a restatement of [22, Lemma 19.4]. ■

B Regret Analysis of Algorithm 1

In this appendix we analyze the regret of Algorithm 1 and prove Theorem 3.2. We begin by establishing some useful concentration lemmas.

B.1 Concentration Lemmas Required to Bound the Regret

We will now prove a lemma that relates the expectation of rewards between the true model $\mathbb{P}$ and an empirical model $\widehat{\mathbb{P}}_t$ when using any fixed policy π. Given any $\eta > 0$ define

$$\bar{\xi}^{(t)}_{s,a}(\eta) = \min\left\{2\eta, 4\eta\sqrt{\frac{H\log\left(|\mathcal{S}||\mathcal{A}|\right) + \log\left(\frac{6\log(N_t(s,a))}{\delta}\right)}{N_t(s,a)}}\right\}. \tag{16}$$

Lemma B.1. *Given any fixed policy $\pi \in \Pi$, and any scalar function $\check{\mu}_\tau$ that depends on the trajectory and satisfies $|\check{\mu}_\tau| \leq \eta$, with probability at least $1-\delta$ for all $t \in \mathbb{N}$*

$$\mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^\pi(\cdot|s_1)}[\check{\mu}_\tau] - \mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}^\pi_t(\cdot|s_1)}[\check{\mu}_\tau] \leq \mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}^\pi_t(\cdot|s_1)}\left[\sum_{h=1}^{H-1}\bar{\xi}^{(t)}_{s_h,a_h}(\eta)\right]. \tag{17}$$

Proof Define $\mathbb{P}^\pi_{(h)}$ to be a trajectory distribution where the initial state is $s_1 \sim \rho$, the state-action pairs up to the end of step h are drawn from $\widehat{\mathbb{P}}^\pi_t$, and the state-action pairs from step $h+1$ up until the last step H are drawn from $\mathbb{P}^\pi$. Notice that $\mathbb{P}^\pi_{(0)}(s_1,\cdot) = \rho(s_1)\mathbb{P}^\pi(\cdot|s_1)$ and $\mathbb{P}^\pi_{(H)}(s_1,\cdot) = \rho(s_1)\widehat{\mathbb{P}}^\pi_t(\cdot|s_1)$. Thus,

$$\mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^\pi(\cdot|s_1)}[\check{\mu}_\tau] - \mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}^\pi_t(\cdot|s_1)}[\check{\mu}_\tau] = \sum_{h=1}^{H}\mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(h-1)}}[\check{\mu}_\tau] - \mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(h)}}[\check{\mu}_\tau]. \tag{18}$$

Consider the term where $h=1$. The trajectory distributions $\mathbb{P}^\pi_{(0)}$ and $\mathbb{P}^\pi_{(1)}$ differ only their distributions of state-action pairs in step 1, thus,

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(0)}}[\check{\mu}_\tau] - \mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(1)}}[\check{\mu}_\tau] \\
&\quad= \mathbb{E}_{s_1\sim\rho}\left[\mathbb{E}_{a_1\sim\pi(\cdot|s_1)}\mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(0)}}[\check{\mu}_\tau|(s_1,a_1)]\right] - \mathbb{E}_{s_1\sim\rho}\left[\mathbb{E}_{a_1\sim\pi(\cdot|s_1)}\mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(0)}}[\check{\mu}_\tau|(s_1,a_1)]\right] = 0.
\end{aligned} \tag{19}$$

Consider any other term in this sum. Again the trajectory distributions $\mathbb{P}^\pi_{(h-1)}$ and $\mathbb{P}^\pi_{(h)}$ differ only their distributions of state-action pairs in step h and hence

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(h-1)}}[\check{\mu}_\tau] - \mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(h)}}[\check{\mu}_\tau] \\
&\quad= \mathbb{E}_{s_1\sim\rho,\ \tau_{h-1}\sim\widehat{\mathbb{P}}^\pi_t(\cdot|s_1)}\left(\mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(h-1)}}[\check{\mu}_\tau|\tau_{h-1}] - \mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(h)}}[\check{\mu}_\tau|\tau_{h-1}]\right) \\
&\quad= \mathbb{E}_{s_1\sim\rho,\ \tau_{h-1}\sim\widehat{\mathbb{P}}^\pi_t(\cdot|s_1)}\left(\mathbb{E}_{s_h\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}\left[\mathbb{E}_{a_h\sim\pi_h(\cdot|s_h,\tau_{h-1})}\mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(h-1)}}[\check{\mu}_\tau|(s_h,a_h,\tau_{h-1})]\right]\right. \\
&\qquad\qquad\left. -\mathbb{E}_{s_h\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}\left[\mathbb{E}_{a_h\sim\pi_h(\cdot|s_h,\tau_{h-1})}\mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(h-1)}}[\check{\mu}_\tau|(s_h,a_h,\tau_{h-1})]\right]\right).
\end{aligned} \tag{20}$$

Define the random variable

$$z_h(s,\tau'_{h-1}) := \mathbb{E}_{a\sim\pi_h(\cdot|s,\tau'_{h-1})}\mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(h-1)}}[\check{\mu}_\tau|\tau_h = \{s,a,\tau'_{h-1}\}].$$

Observe that $|z_h(s,\tau'_{h-1})| \leq \eta$, since $|\check{\mu}_\tau| \leq \eta$ by assumption. Furthermore, the distribution of $z_h(s,\tau'_{h-1})$ only depends on the true transition dynamics $\mathbb{P}$ and on the policy π but it does not depend on the empirical estimate of the transition dynamics $\widehat{\mathbb{P}}_t$. With this definition in hand continuing from equation (20) we have

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(h-1)}}[\check{\mu}_\tau] - \mathbb{E}_{\tau\sim\mathbb{P}^\pi_{(h)}}[\check{\mu}_\tau] \\
&\quad= \mathbb{E}_{s_1\sim\rho,\ \tau_{h-1}\sim\widehat{\mathbb{P}}^\pi_t(\cdot|s_1)}\left[\underbrace{\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})] - \mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]}\right].
\end{aligned} \tag{21}$$

We will upper bound the term in the under-brace above with high probability uniformly over all sub-trajectories τ_{h-1}.

Recall that $N_t(s,a)$ is the number of times the state action pair (s,a) has been visited before episode t, and $N_t(s';s,a)$ is the number of times the state s' is visited starting from state-action pair (s,a) before episode t. When $N_t(s,a)>0$, by definition $\widehat{\mathbb{P}}_t(s'|s,a)=\frac{N_t(s';s,a)}{N_t(s,a)}$. Thus for any fixed sub-trajectory $\tau_{h-1}\in\Gamma_{h-1}$ such that $N_t(s_{h-1,a_{h-1}})>0$ we have

$$\begin{aligned}
&\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]-\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]\\
&=\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]-\sum_{s'\in\mathcal{S}}\frac{N_t(s';s_{h-1},a_{h-1})}{N_t(s_{h-1},a_{h-1})}z_h(s',\tau_{h-1})\\
&=\frac{1}{N_t(s_{h-1},a_{h-1})}\Bigg[N_t(s_{h-1},a_{h-1})\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]\\
&\qquad\qquad-\sum_{s'\in\mathcal{S}}N_t(s';s_{h-1},a_{h-1})z_h(s',\tau_{h-1})\Bigg]\\
&\overset{(i)}{=}\frac{1}{N_t(s_{h-1},a_{h-1})}\left[N_t(s_{h-1},a_{h-1})\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]-\sum_{\ell=1}^{N_t(s_{h-1},a_{h-1})}z_h(s_\ell,\tau_{h-1})\right]\\
&=\frac{1}{N_t(s_{h-1},a_{h-1})}\sum_{\ell=1}^{N_t(s_{h-1},a_{h-1})}\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]-z_h(s_\ell,\tau_{h-1}),
\end{aligned}$$

where in (i) s_ℓ is the state that was visited immediately after ℓth visit to the state-action pair (s_{h-1},a_{h-1}). Note that $|\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]-z_h(s_\ell,\tau_{h-1})|\le 2\eta$, thus by invoking Lemma A.1 we have: given any fixed sub-trajectory $\tau_{h-1}\in\Gamma_{h-1}$, for all $t\in\mathbb{N}$ such that $N_t(s_{h-1},a_{h-1})>0$

$$\begin{aligned}
&\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]-\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]\\
&\qquad\le 4\eta\sqrt{\frac{\log\left(\frac{6\log(N_t(s_{h-1},a_{h-1}))}{\delta'}\right)}{N_t(s_{h-1},a_{h-1})}}
\end{aligned}\tag{22}$$

with probability at least $1-\delta'$. In the case where $N_t(s_{h-1},a_{h-1})=0$ we have the uniform upper bound

$$\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]-\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]\le 2\eta.\tag{23}$$

Therefore combining inequalities (22) and (23), we can conclude that for any fixed sub-trajectory $\tau_{h-1}\in\Gamma_{h-1}$

$$\begin{aligned}
&\forall\, t\in\mathbb{N}\ :\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]-\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]\\
&\qquad\le\min\left\{2\eta,4\eta\sqrt{\frac{\log\left(\frac{6\log(N_t(s_{h-1},a_{h-1}))}{\delta'}\right)}{N_t(s_{h-1},a_{h-1})}}\right\}
\end{aligned}$$

with probability at least $1-\delta'$. By a union bound over all sub-trajectories we find that for all $t\in\mathbb{N}$ and all $\tau_{h-1}\in\Gamma_{h-1}$

$$\begin{aligned}
&\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]-\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]\\
&\qquad\le\min\left\{2\eta,4\eta\sqrt{\frac{\log\left(\frac{6\log(N_t(s_{h-1},a_{h-1}))}{\delta'}\right)}{N_t(s_{h-1},a_{h-1})}}\right\}
\end{aligned}$$

with probability at least $1-\delta'|\mathcal{S}|^{H-1}|\mathcal{A}|^{H-1}$. Finally a union bound over all $h\in[H]$ lets us conclude that for all $t\in\mathbb{N}$, all $h\in[H]$, and all $\tau_{h-1}\in\Gamma_{h-1}$

$$\begin{aligned}
&\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]-\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h(s',\tau_{h-1})]\\
&\qquad\le\min\left\{2\eta,4\eta\sqrt{\frac{\log\left(\frac{6\log(N_t(s_{h-1},a_{h-1}))}{\delta'}\right)}{N_t(s_{h-1},a_{h-1})}}\right\}
\end{aligned}$$

with probability at least $1-\delta'|\mathcal{S}|^{H-1}|\mathcal{A}|^{H-1}H$. Setting $\delta' = \frac{\delta}{|\mathcal{S}|^{H-1}|\mathcal{A}|^{H-1}H}$ and using equation (21) from above we get that for all $t \in \mathbb{N}$, all $h \in \{2,\dots,H\}$,

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\mathbb{P}^{\pi}_{(h-1)}}[\check{\mu}_\tau] - \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}_{(h)}}[\check{\mu}_\tau] \\
&\quad\le \mathbb{E}_{s_1\sim\rho,\ \tau_{h-1}\sim\widehat{\mathbb{P}}^{\pi}_t(\cdot|s_1)}\left[\min\left\{2\eta, 4\eta\sqrt{\frac{(H-1)\log(|\mathcal{S}||\mathcal{A}|H) + \log\left(\frac{6\log(N_t(s_{h-1},a_{h-1}))}{\delta}\right)}{N_t(s_{h-1},a_{h-1})}}\right\}\right] \\
&\quad\le \mathbb{E}_{s_1\sim\rho,\ \tau_{h-1}\sim\widehat{\mathbb{P}}^{\pi}_t(\cdot|s_1)}\left[\bar{\xi}^{(t)}_{s_{h-1},a_{h-1}}\right]
\end{aligned}$$

with probability at least $1-\delta$. Summing over all $h \in [H]$ and using equations (18) and (19) we conclude that

$$\begin{aligned}
\mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^{\pi}(\cdot|s_1)}[\check{\mu}_\tau] - \mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}^{\pi}_t(\cdot|s_1)}[\check{\mu}_\tau] &\le \mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}^{\pi}_t(\cdot|s_1)}\left[\sum_{h=2}^{H}\bar{\xi}^{(t)}_{s_{h-1},a_{h-1}}\right] \\
&= \mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}^{\pi}_t(\cdot|s_1)}\left[\sum_{h=1}^{H-1}\bar{\xi}^{(t)}_{s_h,a_h}\right]
\end{aligned}$$

with the same probability. This establishes our claim. ■

Next, we shall prove a stronger version of Lemma B.1 that holds uniformly over all policies. Given any bounded scalar function $\check{\mu}$ that maps trajectories to $\mathbb{R}$ and satisfies $|\check{\mu}_\tau| \le \eta$, any transition dynamics $\bar{\mathbb{P}}$ and any policy π define

$$z_h^{\check{\mu},\bar{\mathbb{P}}^{\pi}}(s,\tau'_{h-1}) := \mathbb{E}_{a\sim\pi_h(\cdot|s,\tau'_{h-1})}\left[\mathbb{E}_{\tau\sim\bar{\mathbb{P}}^{\pi}}[\check{\mu}_\tau \mid \tau_h = \{s,a,\tau'_{h-1}\}]\right]. \tag{24}$$

This function is different from the z_h that was defined and used locally in the proof of the preceding lemma. The absolute value of the functions $z_h^{\check{\mu},\bar{\mathbb{P}}^{\pi}}$ are also bounded by η.

Suppose that $\Psi(\varepsilon) := \{f_j\}_{j=1}^{\mathcal{N}_{\mathsf{cover}}(\varepsilon)}$ is a set of bounded functions from $\mathcal{S} \mapsto [-\eta,\eta]$, such that for any $h \in [H]$ and for any sub-trajectory $\tau_{h-1} \in \Gamma_{h-1}$, there exists a $f \in \Psi(\varepsilon)$ such that

$$\max_{s\in\mathcal{S}}\left|z_h^{\check{\mu},\bar{\mathbb{P}}^{\pi}}(s,\tau_{h-1}) - f(s)\right| \le \frac{\varepsilon}{2H}. \tag{25}$$

We will construct such a net of functions of size $\mathcal{N}_{\mathsf{cover}}(\varepsilon) \le \left(\left\lceil\frac{\eta-(-\eta)}{\varepsilon/(2H)}\right\rceil\right)^{|\mathcal{S}|} = \left(\left\lceil\frac{4\eta H}{\varepsilon}\right\rceil\right)^{|\mathcal{S}|}$. Such a set of functions can be built as follows. For each $s \in \mathcal{S}$ we pick an element of the set $\{-\eta, -\eta+\frac{\varepsilon}{2H},\dots,\eta\}$. There are at most $\left\lceil\frac{4\eta H}{\varepsilon}\right\rceil$ choices for each state, and therefore there are at most $\left(\left\lceil\frac{4\eta H}{\varepsilon}\right\rceil\right)^{|\mathcal{S}|}$ unique functions that can be defined that map from the state space $\mathcal{S}$ to the set $\{-\eta, -\eta+\frac{\varepsilon}{2H},\dots,\eta\}$. Let $\Psi(\varepsilon)$ be these functions. It is easy to check that this set of functions $\Psi(\varepsilon)$ satisfies the condition specified in inequality (25). Also define the function

$$\check{\xi}^{(t)}_{s,a}(\varepsilon;\eta) := \min\left\{2\eta, 4\eta\sqrt{\frac{H\log(|\mathcal{S}||\mathcal{A}|H) + |\mathcal{S}|\log\left(\left\lceil\frac{4\eta H}{\varepsilon}\right\rceil\right) + \log\left(\frac{6\log(N_t(s_{h-1},a_{h-1}))}{\delta}\right)}{N_t(s_{h-1},a_{h-1})}}\right\}.$$

Lemma B.2. *Suppose that $\varepsilon > 0$. Then with probability at least $1-\delta$, for all $t \in \mathbb{N}$, all policies $\pi \in \Pi$ and all $\check{\mu}_\tau$ such that $|\check{\mu}_\tau| \le \eta$,*

$$\mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}^{\pi}_t(\cdot|s_1)}[\check{\mu}_\tau] - \mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^{\pi}(\cdot|s_1)}[\check{\mu}_\tau] \le \mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^{\pi}(\cdot|s_1)}\left[\sum_{h=1}^{H-1}\check{\xi}^{(t)}_{s_h,a_h}(\varepsilon;\eta)\right] + \varepsilon.$$

Proof Define $\mathbb{P}^{\pi}_{(h)}$ to be a trajectory distribution where the initial state is $s_1 \sim \rho$, the state-action pairs up to the end of step h are drawn from $\mathbb{P}^{\pi}$, and the state-action pairs from step $h+1$ up until the last

step H is drawn from $\widehat{\mathbb{P}}_t^\pi$. Notice that $\mathbb{P}_{(0)}^\pi(s_1,\cdot) = \rho(s_1)\widehat{\mathbb{P}}_t^\pi(\cdot|s_1)$ and $\mathbb{P}_{(H)}^\pi(s_1) = \rho(s_1)\mathbb{P}^\pi(\cdot|s_1)$.

$$\mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}_t^\pi(\cdot|s_1)}[\check{\mu}_\tau] - \mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^\pi(\cdot|s_1)}[\check{\mu}_\tau] = \sum_{h=1}^{H} \mathbb{E}_{\tau\sim\mathbb{P}_{(h-1)}^\pi}[\check{\mu}_\tau] - \mathbb{E}_{\tau\sim\mathbb{P}_{(h)}^\pi}[\check{\mu}_\tau]. \tag{26}$$

Consider the term where $h = 1$. The trajectory distributions $\mathbb{P}_{(0)}^\pi$ and $\mathbb{P}_{(1)}^\pi$ differ only their distributions of state-action pairs in step 1, thus,

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\mathbb{P}_{(0)}^\pi}[\check{\mu}_\tau] - \mathbb{E}_{\tau\sim\mathbb{P}_{(1)}^\pi}[\check{\mu}_\tau] \\
&\quad= \mathbb{E}_{s_1\sim\rho}\left[\mathbb{E}_{a_1\sim\pi(\cdot|s_1)}\mathbb{E}_{\tau\sim\mathbb{P}_{(0)}^\pi}[\check{\mu}_\tau|(s_1,a_1)]\right] - \mathbb{E}_{s_1\sim\rho}\left[\mathbb{E}_{a_1\sim\pi(\cdot|s_1)}\mathbb{E}_{\tau\sim\mathbb{P}_{(0)}^\pi}[\check{\mu}_\tau|(s_1,a_1)]\right] = 0.
\end{aligned} \tag{27}$$

Consider any other term in this sum. Again the trajectory distributions $\mathbb{P}_{(h-1)}^\pi$ and $\mathbb{P}_{(h)}^\pi$ differ only their distributions of state-action pairs in step h and hence

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\mathbb{P}_{(h-1)}^\pi}[\check{\mu}_\tau] - \mathbb{E}_{\tau\sim\mathbb{P}_{(h)}^\pi}[\check{\mu}_\tau] \\
&= \mathbb{E}_{s_1\sim\rho,\ \tau_{h-1}\sim\mathbb{P}^\pi(\cdot|s_1)}\left(\mathbb{E}_{\tau\sim\mathbb{P}_{(h-1)}^\pi}[\check{\mu}_\tau|\tau_{h-1}] - \mathbb{E}_{\tau\sim\mathbb{P}_{(h)}^\pi}[\check{\mu}_\tau|\tau_{h-1}]\right) \\
&= \mathbb{E}_{s_1\sim\rho,\ \tau_{h-1}\sim\mathbb{P}^\pi(\cdot|s_1)}\Bigg[\mathbb{E}_{s_h\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}\left[\mathbb{E}_{a_h\sim\pi_h(\cdot|s_h,\tau_{h-1})}\mathbb{E}_{\tau\sim\widehat{\mathbb{P}}_t^\pi}[\check{\mu}_\tau|(s_h,a_h,\tau_{h-1})]\right] \\
&\qquad\qquad - \mathbb{E}_{s_h\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}\left[\mathbb{E}_{a_h\sim\pi_h(\cdot|s_h,\tau_{h-1})}\mathbb{E}_{\tau\sim\widehat{\mathbb{P}}_t^\pi}[\check{\mu}_\tau|(s_h,a_h,\tau_{h-1})]\right]\Bigg] \\
&= \mathbb{E}_{s_1\sim\rho,\ \tau_{h-1}\sim\mathbb{P}^\pi(\cdot|s_1)}\Bigg[\underbrace{\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^\pi}(s',\tau_{h-1})] - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^\pi}(s',\tau_{h-1})]}\Bigg]
\end{aligned} \tag{28}$$

where $z_h^{\check{\mu},\widehat{\mathbb{P}}_t^\pi}$ is defined in equation (24) above. We shall now upper bound the term in the under-brace above with high probability uniformly over all sub-trajectories τ_{h-1}.

Recall that $N_t(s,a)$ is the number of times the state action pair (s,a) has been visited before episode t, and $N_t(s';s,a)$ is the number of times the state s' is visited starting from state-action pair (s,a) before episode t. When $N_t(s,a) > 0$ by its definition $\widehat{\mathbb{P}}_t(s'|s,a) = \frac{N_t(s';s,a)}{N_t(s,a)}$. Thus for any fixed sub-trajectory $\tau_{h-1} \in \Gamma_{h-1}$ and episode $t \in \mathbb{N}$ where $N_t(s_{h-1},a_{h-1}) > 0$ we have

$$\begin{aligned}
&\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t}(s',\tau_{h-1})] - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t}(s',\tau_{h-1})] \\
&= \sum_{s'\in\mathcal{S}} \frac{N_t(s';s_{h-1},a_{h-1})}{N_t(s_{h-1},a_{h-1})} z_h^{\check{\mu},\widehat{\mathbb{P}}_t^\pi}(s',\tau_{h-1}) - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^\pi}(s',\tau_{h-1})] \\
&= \frac{1}{N_t(s_{h-1},a_{h-1})}\Bigg[\sum_{s'\in\mathcal{S}} N_t(s';s_{h-1},a_{h-1}) z_h^{\check{\mu},\widehat{\mathbb{P}}_t^\pi}(s',\tau_{h-1}) \\
&\qquad\qquad - N_t(s_{h-1},a_{h-1})\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^\pi}(s',\tau_{h-1})]\Bigg] \\
&\overset{(i)}{=} \frac{1}{N_t(s_{h-1},a_{h-1})}\left[\sum_{\ell=1}^{N_t(s_{h-1},a_{h-1})} z_h^{\check{\mu},\widehat{\mathbb{P}}_t^\pi}(s_\ell,\tau_{h-1}) - N_t(s_{h-1},a_{h-1})\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^\pi}(s',\tau_{h-1})]\right] \\
&= \frac{1}{N_t(s_{h-1},a_{h-1})}\sum_{\ell=1}^{N_t(s_{h-1},a_{h-1})} z_h^{\check{\mu},\widehat{\mathbb{P}}_t^\pi}(s_\ell,\tau_{h-1}) - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^\pi}(s',\tau_{h-1})],
\end{aligned} \tag{29}$$

where in (i) s_ℓ is the state that was visited immediately after the ℓth visit to the state-action pair (s_{h-1},a_{h-1}). Let $\widehat{f} \in \Psi(\varepsilon)$ be a function such that

$$\max_{s\in\mathcal{S}}\left|z_h^{\check{\mu},\widehat{\mathbb{P}}_t^\pi}(s,\tau_{h-1}) - \widehat{f}(s)\right| \le \frac{\varepsilon}{2H}.$$

Such a function exists by the definition of the set $\Psi(\varepsilon)$. Therefore,

$$\max_{s\in\mathcal{S}}\left|z_h^{\check{\mu},\widehat{\mathbb{P}}_t^{\pi}}(s,\tau_{h-1})-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}\left[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^{\pi}}(s',\tau_{h-1})\right]-\widehat{f}(s)+\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}\left[\widehat{f}(s')\right]\right|\leq\frac{\varepsilon}{H}.$$

Continuing from equation (29) we have

$$\begin{aligned}&\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^{\pi}}(s',\tau_{h-1})]-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t}(s',\tau_{h-1})]\\&\qquad\leq\frac{1}{N_t(s_{h-1},a_{h-1})}\sum_{\ell=1}^{N_t(s_{h-1},a_{h-1})}\left(\widehat{f}(s_\ell)-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}\left[\widehat{f}(s_\ell)\right]\right)+\frac{\varepsilon}{H}.\end{aligned}\tag{30}$$

Observe that for all ℓ, $\left|\widehat{f}(s_\ell)-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[\widehat{f}(s')]\right|\leq 2\eta$. Thus by invoking Lemma A.1 and by a union bound over the elements of $\Psi(\varepsilon)$, we have that, given any fixed sub-trajectory $\tau_{h-1}\in\Gamma_{h-1}$, for all $f\in\Psi(\varepsilon)$ and all $t\in\mathbb{N}$ such that $N_t(s_{h-1},a_{h-1})>0$:

$$\frac{1}{N_t(s_{h-1},a_{h-1})}\sum_{\ell=1}^{N_t(s_{h-1},a_{h-1})}f(s_\ell)-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}\left[f(s_\ell)\right]\leq 4\eta\sqrt{\frac{\log\left(\frac{6\log(N_t(s_{h-1},a_{h-1}))}{\delta'}\right)}{N_t(s_{h-1},a_{h-1})}}$$

with probability at least $1-|\mathcal{N}_{\mathsf{cover}}(\varepsilon)|\delta'$. Combined with inequality (30) we have that given any fixed sub-trajectory $\tau_{h-1}\in\Gamma_{h-1}$, for all policies $\pi\in\Pi$, for all $\check{\mu}$ bounded by η and all $t\in\mathbb{N}$ such that $N_t(s_{h-1},a_{h-1})>0$:

$$\begin{aligned}&\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^{\pi}}(s',\tau_{h-1})]-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t}(s',\tau_{h-1})]\\&\qquad\leq 4\eta\sqrt{\frac{\log\left(\frac{6\log(N_t(s_{h-1},a_{h-1}))}{\delta'}\right)}{N_t(s_{h-1},a_{h-1})}}+\frac{\varepsilon}{H}\end{aligned}\tag{31}$$

with probability at least $1-|\mathcal{N}_{\mathsf{cover}}(\varepsilon)|\delta'$. We also have a simple upper bound,

$$\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^{\pi}}(s',\tau_{h-1})]-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t}(s',\tau_{h-1})]\leq 2\eta.\tag{32}$$

Combining inequalities (31) and (32) we get that for any fixed sub-trajectory $\tau_{h-1}\in\Gamma_{h-1}$, for all $t\in\mathbb{N}$, for all $\pi\in\Pi$, for all $\check{\mu}$ bounded by η,

$$\begin{aligned}&\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^{\pi}}(s',\tau_{h-1})]-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t}(s',\tau_{h-1})]\\&\qquad\leq\min\left\{2\eta,4\eta\sqrt{\frac{\log\left(\frac{6\log(N_t(s_{h-1},a_{h-1}))}{\delta'}\right)}{N_t(s_{h-1},a_{h-1})}}+\frac{\varepsilon}{H}\right\}.\end{aligned}$$

By a union bound over all sub-trajectories we find that for all $t\in\mathbb{N}$, all policies $\pi\in\Pi$, all $\check{\mu}$ bounded by η and all $\tau_{h-1}\in\Gamma_{h-1}$

$$\begin{aligned}&\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^{\pi}}(s',\tau_{h-1})]-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t}(s',\tau_{h-1})]\\&\qquad\leq\min\left\{2\eta,4\eta\sqrt{\frac{\log\left(\frac{6\log(N_t(s_{h-1},a_{h-1}))}{\delta'}\right)}{N_t(s_{h-1},a_{h-1})}}+\frac{\varepsilon}{H}\right\}\end{aligned}$$

with probability at least $1-(|\mathcal{S}||\mathcal{A}|)^{H-1}|\mathcal{N}_{\mathsf{cover}}(\varepsilon)|\delta'$. Finally a union bound over the steps of the episode $h\in[H]$ lets us conclude that for all $t\in\mathbb{N}$, all policies $\pi\in\Pi$, all $\check{\mu}$ bounded by η, all $h\in[H]$ and all $\tau_{h-1}\in\Gamma_{h-1}$

$$\begin{aligned}&\mathbb{E}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t^{\pi}}(s',\tau_{h-1})]-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_{h-1},a_{h-1})}[z_h^{\check{\mu},\widehat{\mathbb{P}}_t}(s',\tau_{h-1})]\\&\qquad\leq\min\left\{2\eta,4\eta\sqrt{\frac{\log\left(\frac{6\log(N_t(s_{h-1},a_{h-1}))}{\delta'}\right)}{N_t(s_{h-1},a_{h-1})}}+\frac{\varepsilon}{H}\right\}\\&\qquad\leq\min\left\{2\eta,4\eta\sqrt{\frac{\log\left(\frac{6\log(N_t(s_{h-1},a_{h-1}))}{\delta'}\right)}{N_t(s_{h-1},a_{h-1})}}\right\}+\frac{\varepsilon}{H}\end{aligned}$$

with probability at least $1 - H(|\mathcal{S}||\mathcal{A}|)^{H-1}|\mathcal{N}_{\mathsf{cover}}(\varepsilon)|\delta'$. Setting

$$\delta' = \frac{\delta}{|\mathcal{S}|^{H-1}|\mathcal{A}|^{H-1}H\left(\left\lceil\frac{4\eta H}{\varepsilon}\right\rceil\right)^{|\mathcal{S}|}} \leq \frac{\delta}{|\mathcal{S}|^{H-1}|\mathcal{A}|^{H-1}H|\mathcal{N}_{\mathsf{cover}}(\varepsilon)|}$$

and using equation (28) from above we get that for all $t \in \mathbb{N}$, all $h \in \{2, \ldots, H\}$, all $\pi \in \Pi$, and all $\check{\mu}$ bounded by η we have

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\mathbb{P}^{\pi}_{(h-1)}}[\check{\mu}_\tau] - \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}_{(h)}}[\check{\mu}_\tau] \\
&\leq \mathbb{E}_{s_1\sim\rho,\ \tau_{h-1}\sim\mathbb{P}^{\pi}(\cdot|s_1)}\left[\min\left\{2\eta, 4\eta\sqrt{\frac{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^{H-1}\left\lceil\frac{4\eta H}{\varepsilon}\right\rceil^{|\mathcal{S}|}\log(N_t(s_{h-1},a_{h-1}))}{\delta}\right)}{N_t(s_{h-1},a_{h-1})}}\right\}\right] + \frac{\varepsilon}{H} \\
&\leq \mathbb{E}_{s_1\sim\rho,\ \tau_{h-1}\sim\mathbb{P}^{\pi}(\cdot|s_1)}\left[\check{\xi}^{(t)}_{s_{h-1},a_{h-1}}(\varepsilon;\eta)\right] + \frac{\varepsilon}{H}
\end{aligned}$$

with probability at least $1 - \delta$. Summing over all $h \in [H]$ and using equations (26) and (27) we conclude that for $t \in \mathbb{N}$, all $\pi \in \Pi$ and all $\check{\mu}$ bounded by η,

$$\begin{aligned}
\mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}^{\pi}_t(\cdot|s_1)}[\check{\mu}_\tau] - \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}(\cdot|s_1)}[\check{\mu}_\tau] &\leq \mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^{\pi}(\cdot|s_1)}\left[\sum_{h=2}^{H}\check{\xi}^{(t)}_{s_{h-1},a_{h-1}}(\varepsilon;\eta)\right] + H\times\frac{\varepsilon}{H} \\
&= \mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^{\pi}(\cdot|s_1)}\left[\sum_{h=1}^{H-1}\check{\xi}^{(t)}_{s_h,a_h}(\varepsilon;\eta)\right] + \varepsilon
\end{aligned}$$

again with probability at least $1 - \delta$. This completes the proof of this lemma. ■

B.2 Proof of Lemma 3.1

Recall the statement of the lemma from above.

Lemma 3.1. *For any $\delta \in (0, 1]$, define the event*

$$\mathcal{E}_\delta := \left\{\text{for all } t \in [N], \tau \in \Gamma : \left|\mu(\mathbf{w}_\star^\top\phi(\tau)) - \mu(\widehat{\mathbf{w}}_t^\top\phi(\tau))\right| \leq \sqrt{\kappa}\beta_t(\delta)\|\phi(\tau)\|_{\mathbf{\Sigma}_t^{-1}}\right\}. \quad (5)$$

Then $\mathbb{P}(\mathcal{E}_\delta) \geq 1 - \delta$.

Proof We invoke [31, Proposition 7] by noting that in our paper: $c_\mu = 1/\kappa$, $\kappa_\mu = 1$ (Lipschitz constant of μ), $m = 1$ (scale of the rewards), $\lambda = 1$ (the ℓ_2 regularization parameter), $\tau = N$ (length of the sliding window) and $\mathcal{T}(\tau) = [N]$ (in their paper $\mathcal{T}(\tau)$ corresponds to the set of episodes where the underlying parameter $\mathbf{w}_\star$ remains unchanged. In our setting $\mathbf{w}_\star$ is constant for all episodes). ■

B.3 Definition and Properties of a "Good Event" $\mathcal{E}_{\mathsf{good}}$

The proof of Theorem 3.2 proceeds by showing that a favorable event $\mathcal{E}_{\mathsf{good}}$ that occurs with high probability. We shall then upper bound the regret of Algorithm 1 when this event occurs. Before defining this event we need some additional notation.

Definition B.3. *For all $t \in [N]$, given any policy π define*

$$\bar{V}_t^\pi := \mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^{\pi}(\cdot|s_1)}\left[\bar{\mu}_t(\widehat{\mathbf{w}}_t, \tau)\right],$$

where recall from equation (8a) that $\bar{\mu}_t(\widehat{\mathbf{w}}_t, \tau) = \min\left\{\mu\left(\widehat{\mathbf{w}}_t^\top\phi(\tau)\right) + \sqrt{\kappa}\beta_t(\delta)\|\phi(\tau)\|_{\mathbf{\Sigma}_t^{-1}}, 1\right\}$. Further, for all episodes $t \in [N]$ also define $\bar{V}^{(t)} := \bar{V}_t^{\pi^{(t)}}$ and $\bar{V}_\star^{(t)} := \bar{V}_t^{\pi_\star}$.

Also define the value function when the average rewards are $\widetilde{\mu}_t(\widehat{\mathbf{w}}_t, \tau)$ and the transition dynamics are governed by $\widehat{\mathbb{P}}_t$.

Definition B.4. *For any episode $t \in [N]$, given any policy $\pi \in \Pi$ define*

$$\widetilde{V}_t^{\pi} := \mathbb{E}_{s_1 \sim \rho,\ \tau \sim \widehat{\mathbb{P}}_t^{\pi}(\cdot|s_1)} \left[\widetilde{\mu}_t(\widehat{\mathbf{w}}_t, \tau) \right] \tag{33}$$

where $\widetilde{\mu}_t$ is defined above in equation (8b). To simplify notation we additionally define $\widetilde{V}^{(t)} := \widetilde{V}_t^{\pi^{(t)}}$ and $\widetilde{V}_\star^{(t)} := \widetilde{V}_t^{\pi_\star}$.

Consider the following events:

$$\mathcal{E}_1 := \left\{ \sum_{t=1}^{N} V_\star \leq \sum_{t=1}^{N} \widetilde{V}_\star^{(t)} \right\}; \tag{34a}$$

$$\mathcal{E}_2 := \left\{ \sum_{t=1}^{N} \bar{V}^{(t)} - V^{(t)} \leq \beta_N(\delta) \sqrt{8Nd \max\{\kappa, 1\} \log\left(1 + \frac{N}{\kappa d}\right)} + 4\sqrt{N \log\left(\frac{6\log(N)}{\delta}\right)} \right\}; \tag{34b}$$

$$\mathcal{E}_3 := \left\{ \sum_{t=1}^{N} \widetilde{V}^{(t)} - \bar{V}^{(t)} \leq (2H+1) \sum_{t=1}^{N} \sum_{h=1}^{H-1} \xi^{(t)}_{s_h^{(t)}, a_h^{(t)}} + 4H^2 \sqrt{N \log\left(\frac{6\log(N)}{\delta}\right)} + 1 \right\}, \tag{34c}$$

where $(s_h^{(t)}, a_h^{(t)})$ is the state-action pair visited at step h during episode t.

Lemma B.5. *Define the event $\mathcal{E}_{\mathsf{good}} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. Then $\mathbb{P}\left[\mathcal{E}_{\mathsf{good}}\right] \geq 1 - 6N\delta$.*

The good event occurs when the value function $\widetilde{V}_\star^{(t)}$ is optimistic, that is, it over estimates the true value function of the optimal policy $V_\star$ and when the sums of $\bar{V}^{(t)} - V^{(t)}$ and $\widetilde{V}^{(t)} - \bar{V}^{(t)}$ over the episodes can be bounded.

Proof We will show that each of the three events $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ occurs with a high probability and take union bound to prove our claim.

Event $\mathcal{E}_1$: By invoking Lemma B.1 N times, once per episode, with the choice $\eta = 1$ we get

$$\begin{aligned}
\sum_{t=1}^{N} V_\star &= \sum_{t=1}^{N} \mathbb{E}_{s_1 \sim \rho,\ \tau \sim \mathbb{P}^{\pi_\star}(\cdot|s_1)} \left[\mu(\mathbf{w}_\star^\top \phi(\tau)) \right] \\
&\leq \sum_{t=1}^{N} \mathbb{E}_{s_1 \sim \rho,\ \tau \sim \widehat{\mathbb{P}}_t^{\pi_\star}(\cdot|s_1)} \left[\mu(\mathbf{w}_\star^\top \phi(\tau)) + \sum_{h=1}^{H-1} \bar{\xi}^{(t)}_{s_h, a_h}(1) \right] \\
&\leq \sum_{t=1}^{N} \mathbb{E}_{s_1 \sim \rho,\ \tau \sim \widehat{\mathbb{P}}_t^{\pi_\star}(\cdot|s_1)} \left[\mu(\mathbf{w}_\star^\top \phi(\tau)) + \sum_{h=1}^{H-1} \xi^{(t)}_{s_h, a_h} \right] \\
&\qquad \text{(by the definition of } \xi^{(t)}_{s_h, a_h} \text{ in equation (7))}
\end{aligned} \tag{35}$$

with probability at least $1 - N\delta$. Recall the definition of the event $\mathcal{E}_\delta$ from equation (5) and observe that it occurs with probability at least $1 - \delta$ by Lemma 3.1. Under event $\mathcal{E}_\delta$ for any $t \in [N]$ and any $\tau \in \Gamma$

$$\begin{aligned}
\mu(\mathbf{w}_\star^\top \phi(\tau)) = \min\left\{\mu(\mathbf{w}_\star^\top \phi(\tau)), 1\right\} &\leq \min\left\{\mu(\widehat{\mathbf{w}}_t^\top \phi(\tau)) + \sqrt{\kappa}\beta_t(\delta)\|\phi(\tau)\|_{\mathbf{\Sigma}_t^{-1}}, 1\right\} \\
&= \bar{\mu}_t(\widehat{\mathbf{w}}_t, \tau).
\end{aligned}$$

Therefore by a union bound over $\mathcal{E}_\delta$ and the event where inequality (35) holds we infer that

$$\sum_{t=1}^{N} V_\star \leq \sum_{t=1}^{N} \mathbb{E}_{s_1 \sim \rho,\ \tau \sim \widehat{\mathbb{P}}_t^{\pi_\star}(\cdot|s_1)} \left[\bar{\mu}_t(\widehat{\mathbf{w}}_t, \tau) + \sum_{h=1}^{H-1} \xi^{(t)}_{s_h, a_h} \right] = \sum_{t=1}^{N} \widetilde{V}_\star^{(t)},$$

with probability at least $1 - (N+1)\delta$.

Event $\mathcal{E}_2$: Assume that the event $\mathcal{E}_\delta$ occurs. Lemma 3.1 guarantees that this happens with probability at least $1-\delta$. Consider the following martingale difference sequence

$$D_t := \bar{V}^{(t)} - V^{(t)} - \left[\bar{\mu}_t\left(\widehat{\mathbf{w}}_t, \tau^{(t)}\right) - \mu\left(\mathbf{w}_\star^\top \phi(\tau^{(t)})\right)\right].$$

Note that $|D_t| \leq 2$ since both $\bar{\mu}_t$ and μ take values between 0 and 1. Therefore, by applying Lemma A.1 we have that

$$\sum_{t=1}^N \bar{V}^{(t)} - V^{(t)} \leq \sum_{t=1}^N \bar{\mu}_t\left(\widehat{\mathbf{w}}_t, \tau^{(t)}\right) - \mu\left(\mathbf{w}_\star^\top \phi(\tau^{(t)})\right) + 4\sqrt{N \log\left(\frac{6\log(N)}{\delta}\right)} \tag{36}$$

with probability at least $1-\delta$. Let us now upper bound the sum in the RHS above

$$\begin{aligned}
&\sum_{t=1}^N \bar{\mu}_t\left(\widehat{\mathbf{w}}_t, \tau^{(t)}\right) - \mu\left(\mathbf{w}_\star^\top \phi(\tau^{(t)})\right) \\
&\quad \overset{(i)}{=} \sum_{t=1}^N \min\left\{\mu\left(\widehat{\mathbf{w}}_t^\top \phi(\tau^{(t)}\right) + \sqrt{\kappa}\beta_t(\delta)\|\phi(\tau^{(t)})\|_{\mathbf{\Sigma}_t^{-1}}, 1\right\} - \min\left\{\mu\left(\mathbf{w}_\star^\top \phi(\tau^{(t)})\right), 1\right\} \\
&\quad \overset{(ii)}{\leq} \sum_{t=1}^N \left|\mu\left(\widehat{\mathbf{w}}_t^\top \phi(\tau^{(t)}\right) + \sqrt{\kappa}\beta_t(\delta)\|\phi(\tau^{(t)})\|_{\mathbf{\Sigma}_t^{-1}} - \mu\left(\mathbf{w}_\star^\top \phi(\tau^{(t)})\right)\right| \\
&\quad \overset{(iii)}{\leq} 2\sqrt{\kappa}\sum_{t=1}^N \beta_t(\delta)\|\phi(\tau^{(t)})\|_{\mathbf{\Sigma}_t^{-1}} \\
&\quad \overset{(iv)}{\leq} 2\sqrt{\kappa}\beta_N(\delta)\sum_{t=1}^N \|\phi(\tau^{(t)})\|_{\mathbf{\Sigma}_t^{-1}},
\end{aligned}$$

where (i) follows by the definition of $\bar{\mu}_t$ and since μ is bounded between 0 and 1, (ii) follows since for the function $z \mapsto \min\{z, 1\}$ is 1-Lipschitz, (iii) follows since we have assumed that the event $\mathcal{E}_\delta$ occurs which provides the bound $|\mu\left(\widehat{\mathbf{w}}_t^\top \phi(\tau^{(t)}\right) - \mu\left(\mathbf{w}_\star^\top \phi(\tau^{(t)})\right)| \leq \sqrt{\kappa}\beta_t(\delta)\|\phi(\tau^{(t)})\|_{\mathbf{\Sigma}_t^{-1}}$, and (iv) follows since $\beta_t(\delta)$ is an increasing function of t.

Continuing, since for any vector $\mathbf{z} \in \mathbb{R}^N$ $\|\mathbf{z}\|_1 \leq \sqrt{N}\|\mathbf{z}\|_2$, thus

$$\begin{aligned}
\sum_{t=1}^N \bar{\mu}_t\left(\widehat{\mathbf{w}}_t, \tau^{(t)}\right) - \mu\left(\mathbf{w}_\star^\top \phi(\tau^{(t)})\right) &\leq 2\sqrt{\kappa}\beta_N(\delta)\sqrt{N}\sqrt{\sum_{t=1}^N \|\phi(\tau^{(t)})\|^2_{\mathbf{\Sigma}_t^{-1}}} \\
&\leq \beta_N(\delta)\sqrt{8Nd\max\{\kappa, 1\}\log\left(1 + \frac{N}{\kappa d}\right)}
\end{aligned}$$

where the final inequality follows by invoking the determinant lemma (Lemma A.3) from above. A union bound over the event $\mathcal{E}_\delta$ and the event where inequality (36) holds proves that this bound holds with probability at least $1-2\delta$.

Event $\mathcal{E}_3$: We wish to establish a bound on $\sum_{t=1}^N \widetilde{V}^{(t)} - \bar{V}^{(t)}$. By definition

$$\sum_{t=1}^N \widetilde{V}^{(t)} = \sum_{t=1}^N \mathbb{E}_{s_1 \sim \rho,\ \tau \sim \widehat{\mathbb{P}}_t^{\pi^{(t)}}(\cdot|s_1)}\left[\bar{\mu}_t(\widehat{\mathbf{w}}_t, \tau) + \sum_{h=1}^{H-1} \xi_{s_h, a_h}^{(t)}\right].$$

For each $t \in [N]$ define the trajectory score function $\check{\mu}_\tau^{(t)} = \bar{\mu}_t(\widehat{\mathbf{w}}_t, \tau) + \sum_{h=1}^{H-1} \xi_{s_h,a_h}^{(t)}$. Notice that since $|\xi_{s,a}^{(t)}| \leq 2$ we have that $|\check{\mu}_\tau^{(t)}| < 2H$. Thus, by invoking Lemma B.2 N times, once per episode, with the choices $\eta = 2H$ and $\varepsilon = \frac{1}{N}$ we infer that

$$\begin{aligned}
\sum_{t=1}^N \widetilde{V}^{(t)} &\leq \sum_{t=1}^N \left(\mathbb{E}_{s_1 \sim \rho,\ \tau \sim \mathbb{P}^{\pi^{(t)}}(\cdot|s_1)}\left[\bar{\mu}_t(\widehat{\mathbf{w}}_t, \tau) + \sum_{h=1}^{H-1} \xi_{s_h,a_h}^{(t)} + 2H\sum_{h=1}^{H-1} \check{\xi}_{s_h,a_h}^{(t)}\left(\frac{1}{N}, 2H\right)\right] + \frac{1}{N}\right) \\
&= \sum_{t=1}^N \mathbb{E}_{s_1 \sim \rho,\ \tau \sim \mathbb{P}^{\pi^{(t)}}(\cdot|s_1)}\left[\bar{\mu}_t(\widehat{\mathbf{w}}_t, \tau) + (2H+1)\sum_{h=1}^{H-1} \xi_{s_h,a_h}^{(t)}\right] + 1
\end{aligned} \tag{37}$$

with probability $1 - N\delta$. Assume that the event where inequality (37) holds occurs going forward. Under this event the difference

$$\sum_{t=1}^{N} \widetilde{V}^{(t)} - \bar{V}^{(t)} = \sum_{t=1}^{N} \widetilde{V}^{(t)} - \sum_{t=1}^{N} \mathbb{E}_{s_1 \sim \rho,\, \tau \sim \mathbb{P}^{\pi^{(t)}}(\cdot|s_1)} \left[\bar{\mu}_t(\widehat{\mathbf{w}}_t, \tau)\right]$$
$$\leq (2H+1) \sum_{t=1}^{N} \mathbb{E}_{s_1 \sim \rho,\, \tau \sim \mathbb{P}^{\pi^{(t)}}(\cdot|s_1)} \left[\sum_{h=1}^{H-1} \xi^{(t)}_{s_h, a_h}\right] + 1. \tag{38}$$

Finally, define the martingale-difference sequence

$$D_t := (2H+1) \sum_{t=1}^{N} \mathbb{E}_{s_1 \sim \rho,\, \tau \sim \mathbb{P}^{\pi^{(t)}}(\cdot|s_1)} \left[\sum_{h=1}^{H-1} \xi^{(t)}_{s_h, a_h}\right] - (2H+1) \sum_{h=1}^{H-1} \xi^{(t)}_{s_h^{(t)}, a_h^{(t)}}.$$

Notice that $|D_t| \leq (2H+1)(H-1) \leq 2H^2$. Applying Lemma A.1 with $\zeta = 2H^2$ we find that

$$(2H+1) \sum_{t=1}^{N} \mathbb{E}_{s_1 \sim \rho,\, \tau \sim \mathbb{P}^{\pi^{(t)}}(\cdot|s_1)} \left[\sum_{h=1}^{H-1} \xi^{(t)}_{s_h, a_h}\right]$$
$$\leq (2H+1) \sum_{t=1}^{N} \sum_{h=1}^{H-1} \xi^{(t)}_{s_h^{(t)}, a_h^{(t)}} + 4H^2 \sqrt{N \log\left(\frac{6\log(N)}{\delta}\right)}$$

with probability at least $1 - \delta$. Combining this with inequality (38) we conclude that

$$\sum_{t=1}^{N} \widetilde{V}^{(t)} - \bar{V}^{(t)} \leq (2H+1) \sum_{t=1}^{N} \sum_{h=1}^{H-1} \xi^{(t)}_{s_h^{(t)}, a_h^{(t)}} + 4H^2 \sqrt{N \log\left(\frac{6\log(N)}{\delta}\right)} + 1$$

with probability at least $1 - (N+1)\delta$. This proves that $\mathbb{P}\left[\mathcal{E}_3\right] \geq 1 - (N+1)\delta$.

Union bound over the three events: A union bound over the three events shows that $\mathbb{P}\left[\mathcal{E}_{\mathsf{good}}\right] \geq 1 - \mathbb{P}[\mathcal{E}_1^c] - \mathbb{P}[\mathcal{E}_2^c] - \mathbb{P}[\mathcal{E}_3^c] \geq 1 - (2N+4)\delta \geq 1 - 6N\delta$, which completes the proof. ■

B.4 Proof of Theorem 3.2

Recall the statement of the theorem.

Theorem 3.2. *For any $\bar{\delta} \in (0, 1]$, set $\delta = \bar{\delta}/(6N)$ then under Assumptions 2.1 and 2.2 the regret of Algorithm 1 is upper bounded as follows:*

$$\mathcal{R}(N) \leq \widetilde{O}\left(\left[H\sqrt{(H + |\mathcal{S}|)|\mathcal{S}||\mathcal{A}|} + H^2 + \sqrt{\kappa d}(d^3 + B^{3/2})\right]\sqrt{N} + (H + |\mathcal{S}|)H|\mathcal{S}||\mathcal{A}|\right),$$

with probability at least $1 - \bar{\delta}$.

Proof Let us assume that the event $\mathcal{E}_{\mathsf{good}}$ defined in Lemma B.5 occurs. By Lemma B.5 we know that $\mathbb{P}\left[\mathcal{E}_{\mathsf{good}}\right] \geq 1 - 6N\delta$. By the definition of the event $\mathcal{E}_1$ we know that the regret (which is defined in equation (2) above) is upper bounded as follows:

$$\mathcal{R}(N) = \sum_{t=1}^{N} V_\star - V^{(t)} \leq \sum_{t=1}^{N} \widetilde{V}_\star^{(t)} - V^{(t)}.$$

By the definition of the policy $\pi^{(t)}$ (see equation (9)) we have that

$$\widetilde{V}_\star^{(t)} = \mathbb{E}_{s_1 \sim \rho,\, \tau \sim \widehat{\mathbb{P}}_t^{\pi_\star}(\cdot|s_1)} \left[\widetilde{\mu}_t(\widehat{\mathbf{w}}_t, \tau)\right] \leq \mathbb{E}_{s_1 \sim \rho,\, \tau \sim \widehat{\mathbb{P}}_t^{\pi^{(t)}}(\cdot|s_1)} \left[\widetilde{\mu}_t(\widehat{\mathbf{w}}_t, \tau)\right] = \widetilde{V}^{(t)}.$$

Thus,

$$\mathcal{R}(N) \leq \sum_{t=1}^{N} \widetilde{V}^{(t)} - V^{(t)}.$$

Under event $\mathcal{E}_2$ we know that

$$\sum_{t=1}^{N} \bar{V}^{(t)} - V^{(t)} \leq \beta_N(\delta)\sqrt{8Nd\max\{\kappa,1\}\log\left(1+\frac{N}{\kappa d}\right)} + 4\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)}.$$

By combining the previous two inequalities we find that

$$\begin{aligned}\mathcal{R}(N) &\leq \sum_{t=1}^{N} \widetilde{V}^{(t)} - \bar{V}^{(t)} \\ &\quad + \beta_N(\delta)\sqrt{8Nd\max\{\kappa,1\}\log\left(1+\frac{N}{\kappa d}\right)} + 4\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)}.\end{aligned}$$

Finally under event $\mathcal{E}_3$ we have a bound on the first term on the right hand side above, this leads to the bound:

$$\begin{aligned}\mathcal{R}(N) &\leq (2H+1)\sum_{t=1}^{N}\sum_{h=1}^{H-1} \xi^{(t)}_{s_h^{(t)},a_h^{(t)}} + 4H^2\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)} \\ &\quad + \beta_N(\delta)\sqrt{8Nd\max\{\kappa,1\}\log\left(1+\frac{N}{\kappa d}\right)} + 4\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)} + 1. \qquad (39)\end{aligned}$$

It remains to bound the term $\sum_{t=1}^{N}\sum_{h=1}^{H-1} \xi^{(t)}_{s_h^{(t)},a_h^{(t)}}$. First, note that

$$\begin{aligned}\sum_{t=1}^{N}\sum_{h=1}^{H-1} \xi^{(t)}_{s_h^{(t)},a_h^{(t)}} &= \sum_{t=1}^{N}\sum_{h=1}^{H-1} \min\left\{2, 4\sqrt{\frac{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8NH^2)^{|\mathcal{S}|}\log(N_t(s_{h-1}^{(t)},a_{h-1}^{(t)}))}{\delta}\right)}{N_t(s_{h-1}^{(t)},a_{h-1}^{(t)})}}\right\} \\ &\leq \sum_{t=1}^{N}\sum_{h=1}^{H-1} \min\left\{2, 4\sqrt{\frac{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8NH^2)^{|\mathcal{S}|}\log(N)}{\delta}\right)}{N_t(s_{h-1}^{(t)},a_{h-1}^{(t)})}}\right\}.\end{aligned}$$

For every state-action pair (s,a), the minimum in the terms above will be 2 until it is visited at least

$$N_t(s,a) \geq 4\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8H^2N)^{|\mathcal{S}|}\log(N)}{\delta}\right) =: \spadesuit$$

number of times. Therefore,

$$\begin{aligned}&\sum_{t=1}^{N}\sum_{h=1}^{H-1} \xi^{(t)}_{s_h^{(t)},a_h^{(t)}} \\ &\leq 2|\mathcal{S}||\mathcal{A}|\spadesuit + 4\sqrt{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8NH^2)^{|\mathcal{S}|}\log(N))}{\delta}\right)}\sum_{t=1}^{N}\sum_{h=1}^{H-1}\frac{1}{\sqrt{N_t(s_{h-1}^{(t)},a_{h-1}^{(t)})}} \\ &= 2|\mathcal{S}||\mathcal{A}|\spadesuit + 4\sqrt{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8NH^2)^{|\mathcal{S}|}\log(N))}{\delta}\right)}\sum_{s,a\in\mathcal{S}\times\mathcal{A}}\sum_{\ell=1}^{N_N(s,a)}\frac{1}{\sqrt{\ell}} \\ &\overset{(i)}{<} 2|\mathcal{S}||\mathcal{A}|\spadesuit + 8\sqrt{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8NH^2)^{|\mathcal{S}|}\log(N))}{\delta}\right)}\sum_{s,a\in\mathcal{S}\times\mathcal{A}}\sqrt{N_N(s,a)} \\ &\overset{(ii)}{\leq} 2|\mathcal{S}||\mathcal{A}|\spadesuit + 8\sqrt{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8NH^2)^{|\mathcal{S}|}\log(N))}{\delta}\right)|\mathcal{S}||\mathcal{A}|N} \\ &= 8|\mathcal{S}||\mathcal{A}|\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8H^2N)^{|\mathcal{S}|}\log(N)}{\delta}\right) \\ &\qquad + 8\sqrt{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8NH^2)^{|\mathcal{S}|}\log(N))}{\delta}\right)|\mathcal{S}||\mathcal{A}|N} \qquad (40)\end{aligned}$$

where (i) follows since for all $n \in \mathbb{N}$, $\sum_{\ell=1}^{n} \frac{1}{\sqrt{\ell}} < 2\sqrt{n}$, and (ii) follows since $\sum_{s,a \in \mathcal{S}\times\mathcal{A}} N_N(s,a) = N$ and by Jensen's inequality. Plugging this upper bound into inequality (39) we get that

$$\begin{aligned}
\mathcal{R}(N) \leq\ & 8(2H+1)|\mathcal{S}||\mathcal{A}| \cdot \log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H (8H^2N)^{|\mathcal{S}|}\log(N)}{\delta}\right) \\
& + 8(2H+1)\sqrt{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H (8NH^2)^{|\mathcal{S}|}\log(N))}{\delta}\right)|\mathcal{S}||\mathcal{A}|N} \\
& + 4H^2\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)} + \beta_N(\delta)\sqrt{8Nd\max\{\kappa,1\}\log\left(1+\frac{N}{\kappa d}\right)} \\
& + 4\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)} + 1 \qquad (41) \\
=\ & \widetilde{O}\left(\left[H\sqrt{(H+|\mathcal{S}|)|\mathcal{S}||\mathcal{A}|} + H^2 + \sqrt{\kappa d}(d^3+B^{3/2})\right]\sqrt{N} + (H+|\mathcal{S}|)H|\mathcal{S}||\mathcal{A}|\right).
\end{aligned}$$

where the last equality follows since by its definition $\beta_N(\delta) = \widetilde{O}(d^3 + B^{3/2})$ and by simplifying the expression in equation (41). This bound holds with probability $1 - 6N\delta$. Recalling that $\bar{\delta} = 6N\delta$ completes our proof. ■

C Proof of Lemma 3.5

We begin by presenting some additional technical lemmas.

C.1 Additional Technical Results

The first lemma pertains to a pair of positive semi-definite matrices.

Lemma C.1. *If $\mathbf{B} \succeq \mathbf{C} \succ \mathbf{0}$ be $d \times d$ dimensional matrices then,*

$$\sup_{\mathbf{x}\neq 0} \frac{\mathbf{x}^\top \mathbf{B}\mathbf{x}}{\mathbf{x}^\top \mathbf{C}\mathbf{x}} \leq \frac{\det(\mathbf{B})}{\det(\mathbf{C})}.$$

Proof Given any $\mathbf{y} \in \mathbb{R}^d$ let $\mathbf{x} = \mathbf{C}^{-1/2}\mathbf{y}$. Then

$$\sup_{\mathbf{x}\neq 0} \frac{\mathbf{x}^\top \mathbf{B}\mathbf{x}}{\mathbf{x}^\top \mathbf{C}\mathbf{x}} = \sup_{\mathbf{y}\neq 0} \frac{\mathbf{y}^\top \mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}\mathbf{y}}{\|\mathbf{y}\|_2^2} = \left\|\mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}\right\|_{op}$$

by the definition of the operator norm. Recall that by assumption $\mathbf{B} - \mathbf{C} \succeq 0$ therefore $\mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2} - \mathbf{I} \succeq 0$, and hence all the eigenvalues of $\mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}$ are at least 1. Thus,

$$\sup_{\mathbf{x}\neq 0} \frac{\mathbf{x}^\top \mathbf{B}\mathbf{x}}{\mathbf{x}^\top \mathbf{C}\mathbf{x}} \leq \left\|\mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}\right\|_{op} \leq \det(\mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}) = \frac{\det(\mathbf{B})}{\det(\mathbf{C})},$$

where the last equality follows since $\frac{\det(\mathbf{B})}{\det(\mathbf{C})} = \det(\mathbf{C}^{-1/2})\det(\mathbf{B})\det(\mathbf{C}^{-1/2}) = \det(\mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2})$. This completes the proof. ■

Next we present a lemma that establishes guarantees for the EULER algorithm. With some abuse of notation let $r^{\mathbf{v}}(\tau) = \sum_{h\in[H]} r_h^{\mathbf{v}}$ denote the total reward over a trajectory (see definition of $r_h^{\mathbf{v}}$ in equation (13)). Let $V_{\mathbf{v}} := \max_{\pi\in\Pi} \mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^\pi(\cdot|s_1)}\left[r^{\mathbf{v}}(\tau)\right]$ denote the optimal value achieved by any policy when the reward function is $r^{\mathbf{v}}$. The following is a restatement of Lemma 3.4 from [20].

Lemma C.2. *There exists an absolute constant $c > 0$ such that for any $N_{\mathsf{EUL}} > 0$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ the* EULER *algorithm run for N_{EUL} episodes outputs a set of policies*

set $\{\pi^{(\ell)}\}_{\ell=1}^{N_{\mathsf{EUL}}}$ *such that* $U = \mathsf{Unif}(\pi^{(1)}, \cdots, \pi^{(N_{\mathsf{EUL}})})$ *satisfies:*

$$V_{\mathbf{v}} - \mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^U(\cdot|s_1)}\left[r^{\mathbf{v}}(\tau)\right] \le c\left[\sqrt{\frac{|\mathcal{S}||\mathcal{A}|H\log\left(\frac{|\mathcal{S}||\mathcal{A}|N_{\mathsf{EUL}}}{\delta}\right)V_{\mathbf{v}}}{N_{\mathsf{EUL}}}} + \frac{|\mathcal{S}|^2|\mathcal{A}|H^2\log\left(\frac{|\mathcal{S}||\mathcal{A}|N_{\mathsf{EUL}}}{\delta}\right)}{N_{\mathsf{EUL}}}\right].$$

An immediate corollary is the following result.

Corollary C.3. *There exists an absolute constant* C_1 *such that under Assumptions 2.2, 3.3 and 3.4 if* $N_{\mathsf{EUL}} \ge \frac{C_1|\mathcal{S}|^2|\mathcal{A}|H^2\log\left(\frac{|\mathcal{S}||\mathcal{A}|N^2d}{\delta\omega^2}\right)}{\omega^2}$ *then with probability at least* $1-\delta$*, for all* $i \in \left\{1,\ldots,\frac{2d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}\right\}$

$$\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}\left[r^{\mathbf{v}_i}(\tau)\right] \ge \frac{\omega}{2}.$$

Proof Fix an $i \in \left\{1,\ldots,\frac{2d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}\right\}$. By the explorability assumption (Assumption 3.4) we have that $V_{\mathbf{v}_i} \ge \omega$. By Assumption 2.2 since the feature vectors are bounded by 1 we find that

$$\begin{aligned}
V_{\mathbf{v}_i} &= \max_{\pi\in\Pi}\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^\pi(\cdot|s_1)}\left[r^{\mathbf{v}_i}(\tau)\right] \\
&= \max_{\pi\in\Pi}\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^\pi(\cdot|s_1)}\left[\sum_{h=1}^H r_h^{\mathbf{v}_i}(s_h,a_h)\right] \\
&= \max_{\pi\in\Pi}\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^\pi(\cdot|s_1)}\left[\sum_{h=1}^H \mathbf{v}_i^\top\phi_h(s_h,a_h)\right] \\
&= \max_{\pi\in\Pi}\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^\pi(\cdot|s_1)}\left[\mathbf{v}_i^\top\phi(\tau)\right] \le \|\phi(\tau)\|_2\|\mathbf{v}_i\|_2 \le 1
\end{aligned}$$

where the last inequality follows since $\mathbf{v}$ is a unit vector. Thus for any $i \in \left\{1,\ldots,\frac{2d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}\right\}$, because

$$\begin{aligned}
N_{\mathsf{EUL}} &\ge \frac{C_1|\mathcal{S}|^2|\mathcal{A}|H^2\log\left(\frac{|\mathcal{S}||\mathcal{A}|N^2d}{\delta\omega^2}\right)}{\omega^2} \\
&\ge \frac{4c|\mathcal{S}||\mathcal{A}|H\log\left(\frac{2|\mathcal{S}||\mathcal{A}|Nd\log\left(1+\frac{16N}{d\omega^2}\right)}{\delta\log(3/2)}\right)}{\omega}\max\left\{|\mathcal{S}|H, \frac{4c}{\omega}\right\},
\end{aligned}$$

where C_1 is a sufficiently large constant, we have the guarantee that

$$\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}\left[r^{\mathbf{v}_i}(\tau)\right] \ge \omega/2$$

with probability at least $1 - \frac{\delta\log(3/2)}{2d\log\left(1+\frac{16N}{d\omega^2}\right)}$. A union bound completes the proof. ■

The following lemma controls the operator norm of

$$\widehat{\mathbf{a}}_i\widehat{\mathbf{a}}_i^\top - \mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}[\phi(\tau)]\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}[\phi(\tau)]^\top$$

when the number of evaluation episodes N_{EVAL} is sufficiently large.

Lemma C.4. *There exists a positive absolute constant* C_2 *such that for any* $\omega \in (0,1)$ *under Assumption 2.2 if* $N_{\mathsf{EVAL}} \ge \frac{C_2d^3\log^3\left(\frac{Nd^2}{\delta\omega^2}\right)}{\omega^4}$ *then with probability at least* $1-\delta$*, for all* $i \in \left\{1,\ldots,\frac{2d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}\right\}$

$$\left\|\widehat{\mathbf{a}}_i\widehat{\mathbf{a}}_i^\top - \mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}\left[\phi(\tau)\right]\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}\left[\phi(\tau)\right]^\top\right\|_{op} \le \frac{\omega^2}{32d\log\left(d\log\left(1+\frac{16N}{d\omega^2}\right)\right)}.$$

Proof Fix an index $i \in \left\{1, \dots, \frac{2d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}\right\}$. Recall that the trajectories $\tau_i^{(t)}$ are drawn i.i.d. from the distribution $\rho \times \mathbb{P}^{U_i}$. By Assumption 2.2 the absolute value of each entry of $\phi(\tau_i^{(t)})$ is bounded by 1. Thus by applying Hoeffding's inequality to each coordinate and then taking a union bound over all the coordinates we get that

$$\left\|\widehat{\mathbf{a}}_i - \mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}\left[\phi(\tau)\right]\right\|_2^2 = \left\|\frac{1}{N_{\mathsf{EVAL}}}\sum_{t=1}^{N_{\mathsf{EVAL}}}\phi(\tau_i^{(t)}) - \mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}\left[\phi(\tau)\right]\right\|_2^2 \le \frac{c'd\log\left(\frac{2d^2\log\left(1+\frac{16N}{d\omega^2}\right)}{\delta\log(3/2)}\right)}{N_{\mathsf{EVAL}}}$$

with probability at least $1-\delta\log(3/2)/\left(2d\log\left(1+\frac{16N}{d\omega^2}\right)\right)$, where c' is a positive absolute constant. Assume that this event above holds, then by the triangle inequality

$$\begin{aligned}
&\left\|\widehat{\mathbf{a}}_i\widehat{\mathbf{a}}_i^\top - \mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}\left[\phi(\tau)\right]\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}\left[\phi(\tau)\right]^\top\right\|_{op} \\
&\quad\le \left\|\widehat{\mathbf{a}}_i - \mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}\left[\phi(\tau)\right]\right\|_2^2 \\
&\qquad + 2\left\|\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}\left[\phi(\tau)\right]\right\|_2\left\|\widehat{\mathbf{a}}_i - \mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_i}(\cdot|s_1)}\left[\phi(\tau)\right]\right\|_2 \\
&\quad\overset{(i)}{\le} 2\sqrt{\frac{c'd\log\left(\frac{2d^2\log\left(1+\frac{16N}{d\omega^2}\right)}{\delta\log(3/2)}\right)}{N_{\mathsf{EVAL}}}} + \frac{c'd\log\left(\frac{2d^2\log\left(1+\frac{16N}{d\omega^2}\right)}{\delta\log(3/2)}\right)}{N_{\mathsf{EVAL}}} \\
&\quad\overset{(ii)}{\le} \frac{\omega^2}{32d\log\left(d\log\left(1+\frac{16N}{d\omega^2}\right)\right)}
\end{aligned}$$

where (i) follows since $\|\phi(\tau)\| \le 1$, and (ii) holds because since $\omega < 1$ and since

$$N_{\mathsf{EVAL}} \ge \frac{C_2d^3\log^3\left(\frac{Nd^2}{\delta\omega^2}\right)}{\omega^4} \ge \frac{c'(32)^2d^3\log\left(\frac{2d^2\log\left(1+\frac{16N}{d\omega^2}\right)}{\delta\log(3/2)}\right)\log^2\left(d\log(1+\frac{16N}{d\omega^2})\right)}{\omega^4}$$

where C_2 is a large enough positive constant. This shows that the operator norm bound holds for a fixed index i with probability at least $1-\delta\log(3/2)/\left(2d\log\left(1+\frac{16N}{d\omega^2}\right)\right)$. Taking a union bound over all $i \in \left\{1, \dots, \frac{2d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}\right\}$ completes the proof. ■

With these lemmas in place we are now ready to prove Lemma 3.5.

C.2 The Proof

First we restate the lemma.

Lemma 3.5. *There exist positive absolute constants C_1 and C_2 such that, under Assumptions 2.2, 3.3 and 3.4, if Algorithm 2 is run with $N_{\mathsf{EUL}} = \frac{C_1|\mathcal{S}|^2|\mathcal{A}|H^2\log\left(\frac{|\mathcal{S}||\mathcal{A}|N^2d}{\delta\omega^2}\right)}{\omega^2}$ and $N_{\mathsf{EVAL}} = \frac{C_2d^3\log^3\left(\frac{Nd^2}{\delta\omega^2}\right)}{\omega^4}$, and $N > \frac{d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}\left(N_{\mathsf{EUL}}+N_{\mathsf{EVAL}}\right) =: \bar{N}_{\mathsf{exp}}$ then, with probability at least $1-2\delta$, we have $N_{\mathsf{exp}} \le \bar{N}_{\mathsf{exp}}$ and furthermore:*

$$\mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^{\bar{U}}(\cdot|s_1)}\left[\phi(\tau)\phi(\tau)^\top\right] \succeq \frac{\omega^2\log(3/2)}{32d\log\left(d\log\left(1+\frac{16N}{d\omega^2}\right)\right)}\mathbf{I}.$$

Proof Assume the events described in both Corollary C.3 and Lemma C.4 occur. Since N_{EUL} and N_{EVAL} are both appropriately large this happens with probability at least $1-2\delta$.

We shall begin by showing that the number of while loop iterations n_{loop} is bounded by $\frac{d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}$. At any iteration $n \le \frac{2d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}$ by the event in Corollary C.3 we know that

the mixture U_n satisfies

$$\left(\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_n}(\cdot|s_1)}\left[\mathbf{v}_n^\top\phi(\tau)\right]\right)^2 \geq \frac{\omega^2}{4}.$$

Thus,

$$\begin{aligned}
\mathbf{v}_n^\top\mathbf{A}_n\mathbf{v}_n &= \mathbf{v}_n^\top\left(\mathbf{A}_{n-1}+\widehat{\mathbf{a}}_n\widehat{\mathbf{a}}_n^\top\right)\mathbf{v}_n \\
&\geq \mathbf{v}_n^\top\widehat{\mathbf{a}}_n\widehat{\mathbf{a}}_n^\top\mathbf{v}_n \\
&= \left(\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_n}(\cdot|s_1)}\left[\mathbf{v}_n^\top\phi(\tau)\right]\right)^2 + \left(\mathbf{v}_n^\top\widehat{\mathbf{a}}_n\right)^2 - \left(\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_n}(\cdot|s_1)}\left[\phi(\tau)^\top\mathbf{v}_n\right]\right)^2 \\
&\geq \frac{\omega^2}{4} + \left(\mathbf{v}_n^\top\widehat{\mathbf{a}}_n\right)^2 - \left(\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_n}(\cdot|s_1)}\left[\mathbf{v}_n^\top\phi(\tau)\right]\right)^2 \\
&\geq \frac{\omega^2}{4} - \left\|\widehat{\mathbf{a}}_n\widehat{\mathbf{a}}_n^\top - \mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_n}(\cdot|s_1)}\left[\phi(\tau)\right]\mathbb{E}_{s_1\sim\rho,\tau\sim\mathbb{P}^{U_n}(\cdot|s_1)}\left[\phi(\tau)\right]^\top\right\|_{op} \\
&\overset{(i)}{\geq} \frac{\omega^2}{4} - \frac{\omega^2}{32d\log\left(d\log\left(1+\frac{16N}{d\omega^2}\right)\right)} > \frac{3\omega^2}{16},
\end{aligned}$$

where inequality (i) follows by the event in Lemma C.4. Further if $n \leq n_{\mathsf{loop}}$ and the algorithm didn't terminate after iteration $n-1$, we must have

$$\mathbf{v}_n^\top\mathbf{A}_{n-1}\mathbf{v}_n < \frac{\omega^2}{8}.$$

Therefore by Lemma C.1 for any $n \leq \min\left\{n_{\mathsf{loop}}, \frac{2d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}\right\}$ we have

$$\frac{3}{2} = \frac{\frac{3\omega^2}{16}}{\frac{\omega^2}{8}} < \frac{\mathbf{v}_n^\top\mathbf{A}_n\mathbf{v}_n}{\mathbf{v}_n^\top\mathbf{A}_{n-1}\mathbf{v}_n} \leq \frac{\det(\mathbf{A}_n)}{\det(\mathbf{A}_{n-1})}.$$

Thus for any $n \leq \min\left\{n_{\mathsf{loop}}, \frac{2d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}\right\}$,

$$\det(\mathbf{A}_n) > \frac{3}{2}\det(\mathbf{A}_{n-1}) \geq \left(\frac{3}{2}\right)^n\det(\mathbf{A}_0) = \left(\frac{3}{2}\right)^n\left(\frac{\omega^2}{16}\right)^d. \tag{42}$$

The matrix $\mathbf{A}_n$ is obtained as a result of a sequence of rank 1 updates, where each update has its norm bounded ($\|\widehat{\mathbf{a}}_n\|_2 \leq 1$ for all n), so by Lemma A.3:

$$\log\left(\det(\mathbf{A}_n)\right) \leq d\log\left(\frac{\omega^2}{16}+\frac{n}{d}\right). \tag{43}$$

Combining inequalities (42) and (43) we conclude that, for any $n \leq \min\left\{n_{\mathsf{loop}}, \frac{2d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}\right\}$

$$n\log(3/2) + d\log\left(\frac{\omega^2}{16}\right) \leq \log\left(\det(\mathbf{A}_n)\right) \leq d\log\left(\frac{\omega^2}{16}+\frac{n}{d}\right).$$

Therefore, if $n_{\mathsf{loop}} < N$, then then while loop must terminate after at most

$$\begin{aligned}
n_{\mathsf{loop}} &\leq \frac{d\log\left(1+\frac{16n_{\mathsf{loop}}}{d\omega^2}\right)}{\log(3/2)} \\
&\leq \frac{d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}
\end{aligned} \tag{44}$$

loops. To verify that $n_{\mathsf{loop}} < N$, notice that by assumption N is such that

$$\frac{N}{N_{\mathsf{EUL}}+N_{\mathsf{EVAL}}} > \frac{d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}. \tag{45}$$

Therefore inequality (44) is a valid upper bound on n_{loop}.

Thus, we know that the total number of episodes taken by the algorithm to terminate

$$N_{\mathsf{exp}} = n_{\mathsf{loop}} \times (N_{\mathsf{EUL}} + N_{\mathsf{EVAL}}) \leq \frac{d \log\left(1 + \frac{16N}{d\omega^2}\right)}{\log(3/2)} (N_{\mathsf{EUL}} + N_{\mathsf{EVAL}}) = \bar{N}_{\mathsf{exp}}.$$

This proves the first part of the lemma. For the second part notice that for an arbitrary unit vector $\mathbf{v} \in \mathbb{R}^d$

$$\begin{aligned}
&\mathbf{v}^\top \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{\bar{U}}(\cdot|s_1)} \left[\phi(\tau)\phi(\tau)^\top\right] \mathbf{v} \\
&= \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{\bar{U}}(\cdot|s_1)} \left[\left(\mathbf{v}^\top \phi(\tau)\right)^2\right] \\
&\overset{(i)}{=} \frac{1}{n_{\mathsf{loop}}} \sum_{i=1}^{n_{\mathsf{loop}}} \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{U_i}(\cdot|s_1)} \left[\left(\mathbf{v}^\top \phi(\tau)\right)^2\right] \\
&\overset{(ii)}{\geq} \frac{1}{n_{\mathsf{loop}}} \sum_{i=1}^{n_{\mathsf{loop}}} \left(\mathbf{v}^\top \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{U_i}(\cdot|s_1)} \left[\phi(\tau)\right]\right)^2 \\
&= \frac{1}{n_{\mathsf{loop}}} \left[\mathbf{v}^\top \mathbf{A}_{n_{\mathsf{loop}}} \mathbf{v} - \mathbf{v}^\top \mathbf{A}_{n_{\mathsf{loop}}} \mathbf{v} + \sum_{i=1}^{n_{\mathsf{loop}}} \left(\mathbf{v}^\top \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{U_i}(\cdot|s_1)} \left[\phi(\tau)\right]\right)^2\right] \\
&\overset{(iii)}{\geq} \frac{\omega^2}{8n_{\mathsf{loop}}} - \frac{\omega^2}{16 n_{\mathsf{loop}}} + \frac{1}{n_{\mathsf{loop}}} \left[\sum_{i=1}^{n_{\mathsf{loop}}} \left(\left(\mathbf{v}^\top \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{U_i}(\cdot|s_1)} \left[\phi(\tau)\right]\right)^2 - \left(\mathbf{v}^\top \widehat{\mathbf{a}}_i\right)^2\right)\right] \\
&\geq \frac{\omega^2}{16 n_{\mathsf{loop}}} - \max_{i \in n_{\mathsf{loop}}} \left\| \widehat{\mathbf{a}}_i \widehat{\mathbf{a}}_i^\top - \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{U_i}(\cdot|s_1)} \left[\phi(\tau)\right] \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{U_i}(\cdot|s_1)} \left[\phi(\tau)\right]^\top \right\|_{op} \\
&\overset{(iv)}{\geq} \frac{\omega^2 \log(3/2)}{32 d \log\left(d \log\left(1 + \frac{16N}{d\omega^2}\right)\right)}
\end{aligned}$$

where (i) is by the definition of $\bar{U}$ as the uniform mixture over $U_1, \ldots, U_{n_{\mathsf{loop}}}$, (ii) follows by Jensen's inequality, (iii) is because the minimum eigenvalue of $\mathbf{A}_{n_{\mathsf{loop}}}$ is at least $\omega^2/8$ and since $\mathbf{A}_{n_{\mathsf{loop}}} = \frac{\omega^2}{16}\mathbf{I} + \sum_{i=1}^{n_{\mathsf{loop}}} \widehat{\mathbf{a}}_i \widehat{\mathbf{a}}_i^\top$, and finally (iv) is by the upper bound on $n_{\mathsf{loop}} \leq d \log\left(d \log\left(1 + \frac{16N}{d\omega^2}\right)\right)$ established above and by the bound on the operator norm of the difference of the matrices established in Lemma C.4. This wraps up our proof. ■

D Regret Analysis of Algorithm 3 under the explorability assumption

In this algorithm we use the sum-decomposable bonus functions. Throughout this section we assume that Assumptions 2.1, 2.2, 3.3 and 3.4 are in force, and that N_{EXP} and N_{EVAL} are chosen as specified by the statement of Theorem 3.6. We also assume that the number of episodes $N > \bar{N}_{\mathsf{exp}}$ (see its definition in Lemma 3.5). Define the following two quantities that shall be useful in this section

$$t_0 := C_3 \left[\frac{d^2 \log^2(d \log(1 + \frac{16N}{d\omega^2}))}{\omega^4} \sqrt{\log(N/\delta)} + N_{\mathsf{exp}}^{2/3}\right]^{3/2} \tag{46a}$$

$$\Psi_t := \frac{128 d \log\left(d \log\left(1 + \frac{16N}{d\omega^2}\right)\right)}{3 \log(3/2) \omega^2} \cdot \frac{t - N_{\mathsf{exp}}}{t^{2/3} - (N_{\mathsf{exp}} + 1)^{2/3}}, \tag{46b}$$

where C_3 is a large enough positive absolute constant.

D.1 A Sandwich Inequality

As a first step to showing that these bonuses also lead to optimistic value functions we have a sandwich inequality that relates $\|\phi(\tau)\|_{\mathbf{\Sigma}_t^{-1}}$ to the sum decomposable bonus $\sum_{h=1}^H \|\phi_h(s_h, a_h)\|_{\mathbf{\Sigma}_t^{-1}}$.

Lemma D.1. *For any* $\tau \in \Gamma$

$$\|\phi(\tau)\|_{\mathbf{\Sigma}_t^{-1}} \overset{(a)}{\leq} \sum_{h=1}^{H} \|\phi_h(s_h, a_h)\|_{\mathbf{\Sigma}_t^{-1}} \overset{(b)}{\leq} \sqrt{H \frac{\lambda_{\max}(\mathbf{\Sigma}_t)}{\lambda_{\min}(\mathbf{\Sigma}_t)}} \|\phi(\tau)\|_{\mathbf{\Sigma}_t^{-1}}.$$

Proof Since $\phi(\tau) = \sum_{h=1}^{H} \phi_h(s_h, a_h)$ the inequality (a) holds by invoking the triangle inequality. Now to prove inequality (b) note that

$$\begin{aligned}
\sum_{h=1}^{H} \|\phi_h(s_h, a_h)\|_{\mathbf{\Sigma}_t^{-1}} &\leq \sqrt{\lambda_{\max}(\mathbf{\Sigma}_t^{-1})} \sum_{h=1}^{H-1} \|\phi_h(s_h, a_h)\|_2 \\
&\overset{(i)}{\leq} \sqrt{H \lambda_{\max}(\mathbf{\Sigma}_t^{-1})} \sqrt{\sum_{h=1}^{H-1} \|\phi_h(s_h, a_h)\|_2^2} \\
&\overset{(ii)}{=} \sqrt{H \lambda_{\max}(\mathbf{\Sigma}_t^{-1})} \|\phi(\tau)\|_2 \\
&\leq \sqrt{H \frac{\lambda_{\max}(\mathbf{\Sigma}_t^{-1})}{\lambda_{\min}(\mathbf{\Sigma}_t^{-1})}} \|\phi(\tau)\|_{\mathbf{\Sigma}_t^{-1}} \\
&= \sqrt{H \frac{\lambda_{\max}(\mathbf{\Sigma}_t)}{\lambda_{\min}(\mathbf{\Sigma}_t)}} \|\phi(\tau)\|_{\mathbf{\Sigma}_t^{-1}}
\end{aligned}$$

where (i) holds because for any vector $\mathbf{z} \in \mathbb{R}^H$, $\|\mathbf{z}\|_1 \leq \sqrt{H}\|\mathbf{z}\|_2$ and (ii) is a consequence of Assumption 3.4 since ϕ_h and $\phi_{h'}$ are orthogonal for $h \neq h'$ and because ϕ is sum-decomposable by Assumption 3.3. ■

In light of the previous lemma we now establish bounds on the condition number of the matrices $\mathbf{\Sigma}_t^{-1}$ in the next subsection.

D.2 Bound on the Condition Number of $\mathbf{\Sigma}_t$

To bound the condition number we separately upper bound the maximum eigenvalue and lower bound the minimum eigenvalue. Since we have assumed that $\|\phi(\tau)\|_2 \leq 1$, a simple upper bound on the maximum eigenvalue of $\mathbf{\Sigma}_t = \kappa \mathbf{I} + \sum_{q=N_{\mathsf{exp}}+1}^{t} \phi(\tau^{(q)})\phi(\tau^{(q)})^\top$ is

$$\lambda_{\max}(\mathbf{\Sigma}_t) \leq \kappa + (t - N_{\mathsf{exp}}). \tag{47}$$

Let us now derive a lower bound for the smallest eigenvalue. To do this we shall relate the smallest eigenvalue of $\mathbf{\Sigma}_t$ to the smallest eigenvalue of the covariance matrix associated with the exploration policy

$$\bar{\mathbf{\Sigma}} := \mathbb{E}_{s_1 \sim \rho, \tau \sim \mathbb{P}^{\bar{U}}(\cdot|s_1)} \left[\phi(\tau)\phi(\tau)^\top\right].$$

In Lemma 3.5 we derived a high probability lower bound on the minimum eigenvalue of this matrix.

Lemma D.2. *With probability at least* $1 - 3\delta$ *for all* $t \in \{N_{\mathsf{exp}} + 1, \ldots, N\}$*:*

$$\lambda_{\min}(\mathbf{\Sigma}_t) \geq \begin{cases} \kappa + \frac{3\left(t^{2/3} - (N_{\mathsf{exp}}+1)^{2/3}\right)\omega^2 \log(3/2)}{128 d \log\left(d \log\left(1 + \frac{16N}{d\omega^2}\right)\right)} & \text{when } t \geq t_0 \\ \kappa & \text{o.w.,} \end{cases}$$

where t_0 *is defined in equation* (46a).

Proof First let us dispense of the case where $N < t_0$. Since $\mathbf{\Sigma}_t \succeq \kappa \mathbf{I}$ we are done.

Therefore going forward let us assume that $N \geq t_0$. Recall that b_q are the Bernoulli random variables used in Algorithm 3 and that $\mathbb{P}(b_q = 1) = 1/q^{1/3}$. Notice that the following holds

$$\begin{aligned}
\mathbf{\Sigma}_t &= \kappa \mathbf{I} + \sum_{q=N_{\mathsf{exp}}+1}^{t} \phi(\tau^{(q)})\phi(\tau^{(q)})^\top \\
&\succeq \kappa \mathbf{I} + \sum_{q=N_{\mathsf{exp}}+1}^{t} b_q \phi(\tau^{(q)})\phi(\tau^{(q)})^\top \\
&= \kappa \mathbf{I} + \sum_{q=N_{\mathsf{exp}}+1}^{t} \frac{1}{q^{1/3}} \bar{\mathbf{\Sigma}} + \underbrace{\sum_{q=N_{\mathsf{exp}}+1}^{t} \left(b_q \phi(\tau^{(q)})\phi(\tau^{(q)})^\top - \frac{1}{q^{1/3}} \bar{\mathbf{\Sigma}} \right)}_{=:\mathbf{E}_t}.
\end{aligned}$$

Thus we have

$$\begin{aligned}
\lambda_{\min}(\mathbf{\Sigma}_t) &\geq \kappa + \lambda_{\min}(\bar{\mathbf{\Sigma}}) \sum_{q=N_{\mathsf{exp}}+1}^{t} \frac{1}{q^{1/3}} - \lambda_{\max}(\mathbf{E}_t) \\
&\geq \kappa + \lambda_{\min}(\bar{\mathbf{\Sigma}}) \int_{q=N_{\mathsf{exp}}+1}^{t} \frac{1}{q^{1/3}} \, \mathrm{d}q - \lambda_{\max}(\mathbf{E}_t) \\
&= \kappa + \frac{3\left(t^{2/3} - (N_{\mathsf{exp}}+1)^{2/3}\right)}{2} \lambda_{\min}(\bar{\mathbf{\Sigma}}) - \lambda_{\max}(\mathbf{E}_t). \qquad (48)
\end{aligned}$$

First by Lemma 3.5 we know that

$$\lambda_{\min}(\bar{\mathbf{\Sigma}}) \geq \frac{\omega^2 \log(3/2)}{32 d \log\left(d \log\left(1 + \frac{16N}{d\omega^2}\right)\right)} \qquad (49)$$

with probability at least $1 - 2\delta$.

Next to upper bound the maximum eigenvalue of $\mathbf{E}_t$ define the matrix martingale difference sequence

$$\mathbf{D}_q := b_q \phi(\tau^{(q)})\phi(\tau^{(q)})^\top - \frac{1}{q^{1/3}} \bar{\mathbf{\Sigma}}.$$

Observe that $\mathbf{E}_t = \sum_{q=N_{\mathsf{exp}}+1}^{t} \mathbf{D}_q$. We will use the matrix Freedman inequality (Theorem A.2) to upper bound the maximum eigenvalue of $\mathbf{E}_t$. To this end first note that

$$\lambda_{\max}(\mathbf{D}_q) \leq \|\phi(\tau^{(q)})\phi(\tau^{(q)})^\top\|_{op} \leq 1.$$

Further note that

$$\begin{aligned}
&\left\| \sum_{q=N_{\mathsf{exp}}+1}^{t} \mathbb{E}\left[\mathbf{D}_q^2 \,\middle|\, \mathbf{D}_{N_{\mathsf{exp}}+1}, \ldots, \mathbf{D}_{q-1} \right] \right\|_{op} \\
&\leq \sum_{q=N_{\mathsf{exp}}+1}^{t} \left\| \mathbb{E}\left[\mathbf{D}_q^2 \,\middle|\, \mathbf{D}_{N_{\mathsf{exp}}+1}, \ldots, \mathbf{D}_{q-1} \right] \right\|_{op} \\
&= \sum_{q=N_{\mathsf{exp}}+1}^{t} \Bigg\| \mathbb{E}\Bigg[b_q^2 \|\phi(\tau^{(q)})\|_2^2 \phi(\tau^{(q)})\phi(\tau^{(q)})^\top + \frac{\bar{\mathbf{\Sigma}}^2}{q^{2/3}} \\
&\qquad - b_q \left(\phi(\tau^{(q)})\phi(\tau^{(q)})^\top \bar{\mathbf{\Sigma}} + \bar{\mathbf{\Sigma}} \phi(\tau^{(q)})\phi(\tau^{(q)})^\top \right) \,\Bigg|\, \mathbf{D}_{N_{\mathsf{exp}}+1}, \ldots, \mathbf{D}_{q-1} \Bigg] \Bigg\|_{op} \\
&\overset{(i)}{\leq} \sum_{q=N_{\mathsf{exp}}+1}^{t} \left(\frac{1}{q^{1/3}} + \frac{1}{q^{2/3}} + \frac{2}{q^{1/3}} \right) \\
&\leq 4 \sum_{q=N_{\mathsf{exp}}+1}^{t} \frac{1}{q^{1/3}} \leq 4 \int_{q=N_{\mathsf{exp}}}^{t} \frac{1}{q^{1/3}} \, \mathrm{d}q = 6\left(t^{2/3} - N_{\mathsf{exp}}^{2/3} \right)
\end{aligned}$$

where (i) follows since $\mathbb{E}[b_q] = \mathbb{E}[b_q^2] = 1/q^{1/3}$ and because $\|\phi(\tau)\|_2 \leq 1$.

Now we apply Theorem A.2 with the choices

$$\begin{aligned} x &= \frac{3\left(t^{2/3} - (N_{\mathsf{exp}}+1)^{2/3}\right)\omega^2 \log(3/2)}{128 d \log\left(d \log\left(1 + \frac{16N}{d\omega^2}\right)\right)}; \\ V &= 6\left(t^{2/3} - N_{\mathsf{exp}}^{2/3}\right); \\ R &= 1, \end{aligned}$$

to get

$$\mathbb{P}\left[\lambda_{\max}(\mathbf{E}_t) \geq x\right] \leq d \exp\left(-\frac{x^2/2}{V + x/3}\right) \leq d \exp\left(-\frac{x^2}{4V}\right)$$

where the second inequality follows since $V > x/3$. Now by the choice of x and V we know that

$$\mathbb{P}\left[\lambda_{\max}(\mathbf{E}_t) \geq \frac{3\left(t^{2/3} - (N_{\mathsf{exp}}+1)^{2/3}\right)\omega^2 \log(3/2)}{128 d \log\left(d \log\left(1 + \frac{16N}{d\omega^2}\right)\right)}\right] \leq \delta/N$$

whenever

$$t \geq C_3 \left[\frac{d^2 \log^2(d \log(1 + \frac{16N}{d\omega^2}))}{\omega^4} - N_{\mathsf{exp}}^{2/3}\right]^{3/2} = t_0$$

because the constant C_3 is chosen to be large enough. Thus, by a union bound we know that

$$\mathbb{P}\left[\exists t \in \{t_0, \ldots, N\} \ : \ \lambda_{\max}(\mathbf{E}_t) \geq \frac{3\left(t^{2/3} - (N_{\mathsf{exp}}+1)^{2/3}\right)\omega^2 \log(3/2)}{128 d \log\left(d \log\left(1 + \frac{16N}{d\omega^2}\right)\right)}\right] \leq \delta. \tag{50}$$

Combining the inequalities (48), (49) and (50) completes our proof. ■

Next we have a lemma that bounds the condition number

Lemma D.3. *With probability at least* $1 - 3\delta$ *for all* $t \in \{N_{\mathsf{exp}} + 1, \ldots, N\}$

$$\frac{\lambda_{\max}(\mathbf{\Sigma}_t)}{\lambda_{\min}(\mathbf{\Sigma}_t)} \leq \begin{cases} \Psi_t & \text{when } t \geq t_0 \\ 1 + \frac{(t - N_{\mathsf{exp}})}{\kappa} & \text{o.w.}, \end{cases}$$

where t_0 *and* Ψ_N *are defined in equations* (46a) *and* (46b) *respectively.*

Proof The following bound holds with probability at least $1 - 3\delta$ by combing the upper bound on the maximum eigenvalue in inequality (47) with the results of Lemma D.2

$$\frac{\lambda_{\max}(\mathbf{\Sigma}_t)}{\lambda_{\min}(\mathbf{\Sigma}_t)} \leq \begin{cases} \frac{\kappa + (t - N_{\mathsf{exp}})}{\kappa + \frac{3\left(t^{2/3} - (N_{\mathsf{exp}}+1)^{2/3}\right)\omega^2 \log(3/2)}{128 d \log\left(d \log\left(1 + \frac{16N}{d\omega^2}\right)\right)}} & \text{when } t \geq t_0 \\ 1 + \frac{(t - N_{\mathsf{exp}})}{\kappa} & \text{o.w.} \end{cases}$$

Now for any $a, b, c > 0$: $\frac{a+c}{b+c} \leq \frac{a}{b}$ if $a > b$. Therefore we can simplify the expression above in case where $t \geq t_0$ to get

$$\frac{\lambda_{\max}(\mathbf{\Sigma}_t)}{\lambda_{\min}(\mathbf{\Sigma}_t)} \leq \begin{cases} \frac{128 d \log\left(d \log\left(1 + \frac{16N}{d\omega^2}\right)\right)}{3 \log(3/2) \omega^2} \cdot \frac{t - N_{\mathsf{exp}}}{t^{2/3} - (N_{\mathsf{exp}}+1)^{2/3}} & \text{when } t \geq t_0 \\ 1 + \frac{(t - N_{\mathsf{exp}})}{\kappa} & \text{o.w.} \end{cases}$$

By recalling the definition of Ψ_t from above the claim follows. ■

D.3 Definition and Properties of Another "Good Event" $\mathcal{E}^{\mathsf{sd}}_{\mathsf{good}}$

Similar to the proof of Theorem 3.2 the proof of Theorem 3.6 also proceeds by showing that a different favorable event $\mathcal{E}^{\mathsf{sd}}_{\mathsf{good}}$ occurs with high probability. We shall upper bound the regret of Algorithm 3 when this favorable event occurs. Before defining this event we need some additional notation.

Definition D.4. *For all $t \in [N]$, given any policy π define*

$$\bar{V}_t^{\pi,\mathsf{sd}} := \mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^\pi(\cdot|s_1)}\left[\bar{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t,\tau)\right],$$

where recall from equation (11a) that $\bar{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t,\tau) = \min\left\{\mu\left(\widehat{\mathbf{w}}_t^\top\phi(\tau)\right) + \sqrt{\kappa}\beta_t(\delta)\sum_{h=1}^H\|\phi_h(s_h,a_h)\|_{\mathbf{\Sigma}_t^{-1}}, 1\right\}$. *Further, for all episodes $t \in [N]$ also define* $\bar{V}^{(t),\mathsf{sd}} := \bar{V}_t^{\pi^{(t)},\mathsf{sd}}$ *and* $\bar{V}_\star^{(t),\mathsf{sd}} := \bar{V}_t^{\pi_\star,\mathsf{sd}}$.

Also define the value function when the average rewards are $\widetilde{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t,\tau)$ and the transition dynamics are governed by $\widehat{\mathbb{P}}_t$.

Definition D.5. *For any episode $t \in [N]$, given any policy $\pi \in \Pi$ define*

$$\widetilde{V}_t^{\pi,\mathsf{sd}} := \mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}_t^\pi(\cdot|s_1)}\left[\widetilde{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t,\tau)\right] \tag{51}$$

where $\widetilde{\mu}_t^{\mathsf{sd}}$ is defined above in equation (11b). To simplify notation we additionally define $\widetilde{V}^{(t),\mathsf{sd}} := \widetilde{V}_t^{\pi^{(t)},\mathsf{sd}}$ *and* $\widetilde{V}_\star^{(t),\mathsf{sd}} := \widetilde{V}_t^{\pi_\star,\mathsf{sd}}$.

Recall the definition of t_0 from equation (46a) above and consider the following events:

$$\mathcal{E}_1^{\mathsf{sd}} := \left\{\sum_{t=t_0+1}^N (1-b_t)V_\star \le \sum_{t=t_0+1}^N (1-b_t)\widetilde{V}_\star^{(t),\mathsf{sd}}\right\}; \tag{52a}$$

$$\mathcal{E}_2^{\mathsf{sd}} := \left\{\sum_{t=t_0+1}^N (1-b_t)\left(\bar{V}^{(t),\mathsf{sd}} - V^{(t)}\right) \le \beta_N(\delta)(1+\sqrt{H\Psi_N})\sqrt{8Nd\max\{\kappa,1\}\log\left(1+\frac{N}{\kappa d}\right)} + 4\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)}\right\}; \tag{52b}$$

$$\mathcal{E}_3^{\mathsf{sd}} := \left\{\sum_{t=t_0+1}^N (1-b_t)\left(\widetilde{V}^{(t),\mathsf{sd}} - \bar{V}^{(t),\mathsf{sd}}\right) \le (2H+1)\sum_{t=t_0+1}^N\sum_{h=1}^{H-1}\xi^{(t)}_{s_h^{(t)},a_h^{(t)}} + 4H^2\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)} + 1\right\}; \tag{52c}$$

$$\mathcal{E}_4^{\mathsf{sd}} := \left\{\sum_{t=t_0+1}^N b_t \le \left(\frac{20}{3}\log\left(\frac{1}{\delta}\right)\right)^{3/2} + 4N^{2/3}\right\}; \tag{52d}$$

$$\mathcal{E}_5^{\mathsf{sd}} := \left\{N_{\mathsf{EXP}} \le \frac{d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}(N_{\mathsf{EUL}} + N_{\mathsf{EVAL}})\right\}, \tag{52e}$$

where $(s_h^{(t)}, a_h^{(t)})$ is the state-action pair visited at step h during episode t. In the definitions of the events above if $N < t_0 + 1$ and the sums are "empty" then we take their value to be zero.

Lemma D.6. *Define the event* $\mathcal{E}^{\mathsf{sd}}_{\mathsf{good}} := \mathcal{E}_1^{\mathsf{sd}} \cap \mathcal{E}_2^{\mathsf{sd}} \cap \mathcal{E}_3^{\mathsf{sd}} \cap \mathcal{E}_4^{\mathsf{sd}} \cap \mathcal{E}_5^{\mathsf{sd}}$. *If $N > \bar{N}_{\mathsf{exp}}$ then* $\mathbb{P}\left[\mathcal{E}^{\mathsf{sd}}_{\mathsf{good}}\right] \ge 1 - 12N\delta$.

Proof We will show that each of the five events $\mathcal{E}_1^{\mathsf{sd}}$, $\mathcal{E}_2^{\mathsf{sd}}$, $\mathcal{E}_3^{\mathsf{sd}}$, $\mathcal{E}_4^{\mathsf{sd}}$ and $\mathcal{E}_5^{\mathsf{sd}}$ occurs with a high probability and take union bound to prove our claim.

Event $\mathcal{E}_1^{\mathsf{sd}}$: By invoking Lemma B.1 $N - t_0$ times, once per episode, with the choice $\eta = 1$ we get

$$\begin{aligned}
\sum_{t=t_0}^{N}(1-b_t)V_\star &= \sum_{t=t_0}^{N}(1-b_t)\mathbb{E}_{s_1\sim\rho,\ \tau\sim\mathbb{P}^{\pi_\star}(\cdot|s_1)}\left[\mu(\mathbf{w}_\star^\top\phi(\tau))\right] \\
&\le \sum_{t=t_0}^{N}(1-b_t)\mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}_t^{\pi_\star}(\cdot|s_1)}\left[\mu(\mathbf{w}_\star^\top\phi(\tau)) + \sum_{h=1}^{H-1}\bar{\xi}_{s_h,a_h}^{(t)}(1)\right] \\
&\le \sum_{t=t_0}^{N}(1-b_t)\mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}_t^{\pi_\star}(\cdot|s_1)}\left[\mu(\mathbf{w}_\star^\top\phi(\tau)) + \sum_{h=1}^{H-1}\xi_{s_h,a_h}^{(t)}\right] \qquad (53) \\
&\qquad\qquad \text{(by the definition of } \xi_{s_h,a_h}^{(t)} \text{ in equation (7))}
\end{aligned}$$

with probability at least $1 - N\delta$. Recall the definition of the event $\mathcal{E}_\delta$ from equation (5) and observe that it occurs with probability at least $1-\delta$ by Lemma 3.1. Under event $\mathcal{E}_\delta$ for any $t \in \{t_0, \dots, N\}$ and any $\tau \in \Gamma$

$$\begin{aligned}
\mu(\mathbf{w}_\star^\top\phi(\tau)) = \min\left\{\mu(\mathbf{w}_\star^\top\phi(\tau)), 1\right\} &\le \min\left\{\mu(\widehat{\mathbf{w}}_t^\top\phi(\tau)) + \sqrt{\kappa}\beta_t(\delta)\|\phi(\tau)\|_{\mathbf{\Sigma}_t^{-1}}, 1\right\} \\
&\le \min\left\{\mu(\widehat{\mathbf{w}}_t^\top\phi(\tau)) + \sqrt{\kappa}\beta_t(\delta)\sum_{h=1}^{H}\|\phi_h(s_h,a_h)\|_{\mathbf{\Sigma}_t^{-1}}, 1\right\} \\
&\qquad \text{(by Lemma D.1)} \\
&= \bar{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t, \tau).
\end{aligned}$$

Therefore by a union bound over $\mathcal{E}_\delta$ and the event where inequality (53) holds we infer that

$$\sum_{t=t_0}^{N}(1-b_t)V_\star \le \sum_{t=t_0}^{N}(1-b_t)\mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}_t^{\pi_\star}(\cdot|s_1)}\left[\bar{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t,\tau) + \sum_{h=1}^{H-1}\xi_{s_h,a_h}^{(t)}\right] = \sum_{t=1}^{N}(1-b_t)\widetilde{V}_\star^{(t),\mathsf{sd}},$$

with probability at least $1 - (N+1)\delta$.

Event $\mathcal{E}_2^{\mathsf{sd}}$: Assume that the event $\mathcal{E}_\delta$ occurs and also that for all $t \in \{t_0, \dots, N\}$

$$\frac{\lambda_{\max}(\mathbf{\Sigma}_t)}{\lambda_{\min}(\mathbf{\Sigma}_t} \le \Psi_t. \qquad (54)$$

The results of Lemma 3.1 and Lemma D.3 along with a union bound guarantee that this happens with probability at least $1 - 4\delta$.

Consider the following martingale difference sequence

$$D_t := (1-b_t)\left(\bar{V}^{(t),\mathsf{sd}} - V^{(t)} - \left[\bar{\mu}_t^{\mathsf{sd}}\left(\widehat{\mathbf{w}}_t, \tau^{(t)}\right) - \mu\left(\mathbf{w}_\star^\top\phi(\tau^{(t)})\right)\right]\right).$$

Note that $|D_t| \le 2$ since both $\bar{\mu}_t^{\mathsf{sd}}$ and μ take values between 0 and 1. Therefore, by applying Lemma A.1 we have that

$$\begin{aligned}
&\sum_{t=t_0}^{N}(1-b_t)\left(\bar{V}^{(t),\mathsf{sd}} - V^{(t)}\right) \\
&\le \sum_{t=t_0}^{N}(1-b_t)\left(\bar{\mu}_t^{\mathsf{sd}}\left(\widehat{\mathbf{w}}_t, \tau^{(t)}\right) - \mu\left(\mathbf{w}_\star^\top\phi(\tau^{(t)})\right)\right) + 4\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)} \qquad (55)
\end{aligned}$$

with probability at least $1-\delta$. Let us now upper bound the sum in the RHS above

$$\begin{aligned}
&\sum_{t=t_0}^{N}(1-b_t)\bar{\mu}_t^{\mathsf{sd}}\left(\widehat{\mathbf{w}}_t,\tau^{(t)}\right)-\mu\left(\mathbf{w}_\star^\top\phi(\tau^{(t)})\right)\\
&\overset{(i)}{=}\sum_{t=t_0}^{N}(1-b_t)\left(\min\left\{\mu\left(\widehat{\mathbf{w}}_t^\top\phi(\tau^{(t)}\right)+\sqrt{\kappa}\beta_t(\delta)\sum_{h=1}^{H}\|\phi_h(s_h,a_h)\|_{\mathbf{\Sigma}_t^{-1}},1\right\}-\min\left\{\mu\left(\mathbf{w}_\star^\top\phi(\tau^{(t)})\right),1\right\}\right)\\
&\overset{(ii)}{\leq}\sum_{t=t_0}^{N}\left|\mu\left(\widehat{\mathbf{w}}_t^\top\phi(\tau^{(t)}\right)+\sqrt{\kappa}\beta_t(\delta)\sum_{h=1}^{H}\|\phi_h(s_h,a_h)\|_{\mathbf{\Sigma}_t^{-1}}-\mu\left(\mathbf{w}_\star^\top\phi(\tau^{(t)})\right)\right|\\
&\overset{(iii)}{\leq}2\sqrt{\kappa}\sum_{t=t_0}^{N}\beta_t(\delta)\left(\|\phi(\tau^{(t)})\|_{\mathbf{\Sigma}_t^{-1}}+\sum_{h=1}^{H}\|\phi_h(s_h,a_h)\|_{\mathbf{\Sigma}_t^{-1}}\right)\\
&\overset{(iv)}{\leq}2\sqrt{\kappa}\beta_N(\delta)\sum_{t=t_0}^{N}\left(\|\phi(\tau^{(t)})\|_{\mathbf{\Sigma}_t^{-1}}+\sum_{h=1}^{H}\|\phi_h(s_h,a_h)\|_{\mathbf{\Sigma}_t^{-1}}\right)\\
&\overset{(v)}{\leq}2\sqrt{\kappa}\beta_N(\delta)\sum_{t=t_0}^{N}\left(1+\sqrt{H\frac{\lambda_{\max}(\mathbf{\Sigma}_t)}{\lambda_{\min}(\mathbf{\Sigma}_t)}}\right)\|\phi(\tau^{(t)})\|_{\mathbf{\Sigma}_t^{-1}}\\
&\overset{(vi)}{\leq}2\sqrt{\kappa}\beta_N(\delta)\left(1+\sqrt{H\Psi_N}\right)\sum_{t=t_0}^{N}\|\phi(\tau^{(t)})\|_{\mathbf{\Sigma}_t^{-1}}
\end{aligned}$$

where (i) follows by the definition of $\bar{\mu}_t^{\mathsf{sd}}$ and since μ is bounded between 0 and 1, (ii) follows since for the function $z\mapsto\min\{z,1\}$ is 1-Lipschitz and since $1-b_t\in\{0,1\}$, (iii) follows since we have assumed that the event $\mathcal{E}_\delta$ occurs which provides the bound $|\mu\left(\widehat{\mathbf{w}}_t^\top\phi(\tau^{(t)}\right)-\mu\left(\mathbf{w}_\star^\top\phi(\tau^{(t)})\right)|\leq\sqrt{\kappa}\beta_t(\delta)\|\phi(\tau^{(t)})\|_{\mathbf{\Sigma}_t^{-1}}$, (iv) follows since $\beta_t(\delta)$ is an increasing function of t, (v) follows by invoking Lemma D.1 and finally (vi) follows since we have assumed a bound on the condition number of $\mathbf{\Sigma}_t$ in inequality (54) and because $\Psi_N>\Psi_t$.

Continuing, since for any vector $\mathbf{z}\in\mathbb{R}^N$ $\|\mathbf{z}\|_1\leq\sqrt{N}\|\mathbf{z}\|_2$, thus

$$\begin{aligned}
&\sum_{t=t_0}^{N}(1-b_t)\left(\bar{\mu}_t^{\mathsf{sd}}\left(\widehat{\mathbf{w}}_t,\tau^{(t)}\right)-\mu\left(\mathbf{w}_\star^\top\phi(\tau^{(t)})\right)\right)\\
&\qquad\leq 2\sqrt{\kappa}\beta_N(\delta)\left(1+\sqrt{H\Psi_N}\right)\sqrt{N}\sqrt{\sum_{t=1}^{N}\|\phi(\tau^{(t)})\|^2_{\mathbf{\Sigma}_t^{-1}}}\\
&\qquad\leq\beta_N(\delta)\left(1+\sqrt{H\Psi_N}\right)\sqrt{8Nd\max\{\kappa,1\}\log\left(1+\frac{N}{\kappa d}\right)}
\end{aligned}$$

where the final inequality follows by invoking the determinant lemma (Lemma A.3) from above.

A union bound over the event $\mathcal{E}_\delta$, the event where the condition number of $\mathbf{\Sigma}_t$ is bounded and the event where inequality (55) holds proves that this bound holds with probability at least $1-5\delta$.

Event $\mathcal{E}_3^{\mathsf{sd}}$: By mirroring the proof on the bound on the probability of the event $\mathcal{E}_3$ in Lemma B.5 we can show that $\mathbb{P}\left[\mathcal{E}_3^{\mathsf{sd}}\right]\geq 1-(N+1)\delta$.

Event $\mathcal{E}_4^{\mathsf{sd}}$: On applying Theorem A.2 with the martingale difference sequence $b_t-1/t^{1/3}$ we know that with probability at least $1-\delta$:

$$\sum_{t=1}^{N}b_t\leq 4N^{2/3}$$

if $N \geq \left(\frac{20}{3}\log\left(\frac{1}{\delta}\right)\right)^{3/2}$. Thus, with probability at least $1-\delta$

$$\sum_{t=1}^{N} b_t \leq \left(\frac{20}{3}\log\left(\frac{1}{\delta}\right)\right)^{3/2} + 4N^{2/3}.$$

In other words $\mathbb{P}[\mathcal{E}_4^{\mathsf{sd}}] \geq 1-\delta$.

Event $\mathcal{E}_5^{\mathsf{sd}}$: By invoking Lemma 3.5 it immediately follows that $\mathbb{P}[\mathcal{E}_5^{\mathsf{sd}}] \geq 1-2\delta$.

Union bound over the five events: A union bound over the five events shows that

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}_{\mathsf{good}}^{\mathsf{sd}}\right] &\geq 1 - \mathbb{P}[(\mathcal{E}_1^{\mathsf{sd}})^c] - \mathbb{P}[(\mathcal{E}_2^{\mathsf{sd}})^c] - \mathbb{P}[(\mathcal{E}_3^{\mathsf{sd}})^c] - \mathbb{P}[(\mathcal{E}_4^{\mathsf{sd}})^c] - \mathbb{P}[(\mathcal{E}_5^{\mathsf{sd}})^c] \\ &\geq 1 - (2N+10)\delta \geq 1 - 12N\delta,\end{aligned}$$

which completes the proof. ∎

D.4 Proof of Theorem 3.6

Recall the statement of the theorem.

Theorem 3.6. *For any $\bar{\delta} \in (0,1]$, set $\delta = \bar{\delta}/(12N)$. Under Assumptions 2.1, 2.2, 3.3 and 3.4, and for all $N > \bar{N}_{\mathsf{exp}}$ (see its definition in Lemma 3.5) if Algorithm 3 is run with the parameters N_{EUL} and N_{EVAL} set as specified in Lemma 3.5 then its regret is upper bounded as follows:*

$$\mathcal{R}(N) \leq \widetilde{O}\left(\frac{\sqrt{\kappa H}d}{\omega}(d^3 + B^{3/2})N^{2/3} + \left[H\sqrt{(H+|\mathcal{S}|)|\mathcal{S}||\mathcal{A}|} + H^2\right]\sqrt{N} \right.$$
$$\left. + (H+|\mathcal{S}|)H|\mathcal{S}||\mathcal{A}| + \frac{d^2}{\omega^2}\left(\frac{d^2}{\omega^2} + |\mathcal{S}|^2|\mathcal{A}|H^2\right)\right),$$

with probability at least $1-\bar{\delta}$.

Proof Let us assume that the event $\mathcal{E}_{\mathsf{good}}^{\mathsf{sd}}$ defined in Lemma D.6 occurs. By Lemma D.6 we know that $\mathbb{P}\left[\mathcal{E}_{\mathsf{good}}\right] \geq 1 - 12N\delta$. First we decompose the regret as follows:

$$\begin{aligned}\mathcal{R}(N) &= \sum_{t=1}^{N} V_\star - V^{(t)} \\ &= \sum_{t=1}^{t_0} V_\star - V^{(t)} + \sum_{t=t_0+1}^{N} V_\star - V^{(t)} \\ &= \sum_{t=1}^{t_0} V_\star - V^{(t)} + \sum_{t=t_0+1}^{N} b_t(V_\star - V^{(t)}) + \sum_{t=t_0+1}^{N}(1-b_t)(V_\star - V^{(t)}).\end{aligned}$$

Now since $V_\star - V^{(t)}$ is bounded between 0 and 1 we know that

$$\begin{aligned}
\mathcal{R}(N) &\leq t_0 + \sum_{t=t_0+1}^{N} b_t + \sum_{t=t_0+1}^{N} (1-b_t)(V_\star - V^{(t)}) \\
&\overset{(i)}{\leq} t_0 + \left(\frac{20}{3}\log\left(\frac{1}{\delta}\right)\right)^{3/2} + 4N^{2/3} + \sum_{t=t_0+1}^{N} (1-b_t)(V_\star - V^{(t)}) \\
&\overset{(ii)}{\leq} C_3 \left[\frac{d^2 \log^2(d\log(1+\frac{16N}{d\omega^2}))}{\omega^4}\sqrt{\log(N/\delta)} + N_{\mathsf{exp}}^{2/3}\right]^{3/2} + \left(\frac{20}{3}\log\left(\frac{1}{\delta}\right)\right)^{3/2} \\
&\quad + 4N^{2/3} + \sum_{t=t_0+1}^{N} (1-b_t)(V_\star - V^{(t)}) \\
&\overset{(iii)}{\leq} C_3 \left[\frac{d^2 \log^2(d\log(1+\frac{16N}{d\omega^2}))}{\omega^4}\sqrt{\log(N/\delta)} + \left(\frac{d\log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)}(N_{\mathsf{EUL}} + N_{\mathsf{EVAL}})\right)^{2/3}\right]^{3/2} \\
&\quad + \left(\frac{20}{3}\log\left(\frac{1}{\delta}\right)\right)^{3/2} + 4N^{2/3} \\
&\quad + \sum_{t=t_0+1}^{N} (1-b_t)(V_\star - V^{(t)})
\end{aligned} \tag{56}$$

where (i) follows by the definition of the event $\mathcal{E}_4^{\mathsf{sd}}$, (ii) is by the definition of t_0 in equation (46a) and (iii) follows by the definition of $\mathcal{E}_5^{\mathsf{sd}}$ that bounds N_{EXP}. It remains to bound the last term in the RHS above. Going forward let us assume that $N \geq t_0 + 1$, else we are done. To bound this term note that by the definition of the event $\mathcal{E}_1^{\mathsf{sd}}$ we know that

$$\sum_{t=t_0+1}^{N} (1-b_t)(V_\star - V^{(t)}) \leq \sum_{t=t_0+1}^{N} (1-b_t)\left(\widetilde{V}_\star^{(t),\mathsf{sd}} - V^{(t)}\right).$$

By the definition of the policy $\pi^{(t)}$ (see equation (14)) we have that

$$\widetilde{V}_\star^{(t),\mathsf{sd}} = \mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}_t^{\pi_\star}(\cdot|s_1)}\left[\widetilde{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t,\tau)\right] \leq \mathbb{E}_{s_1\sim\rho,\ \tau\sim\widehat{\mathbb{P}}_t^{\pi^{(t)}}(\cdot|s_1)}\left[\widetilde{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t,\tau)\right] = \widetilde{V}^{(t),\mathsf{sd}}.$$

Thus,

$$\sum_{t=t_0+1}^{N} (1-b_t)(V_\star - V^{(t)}) \leq \sum_{t=t_0+1}^{N} (1-b_t)\left(\widetilde{V}^{(t),\mathsf{sd}} - V^{(t)}\right).$$

Under event $\mathcal{E}_2^{\mathsf{sd}}$ we know that

$$\begin{aligned}
&\sum_{t=t_0+1}^{N} (1-b_t)\left(\bar{V}^{(t),\mathsf{sd}} - V^{(t)}\right) \\
&\leq \beta_N(\delta)\left(1+\sqrt{H\Psi_N}\right)\sqrt{8Nd\max\{\kappa,1\}\log\left(1+\frac{N}{\kappa d}\right)} + 4\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)}.
\end{aligned}$$

By combining the previous two inequalities we find that

$$\begin{aligned}
&\sum_{t=t_0+1}^{N} (1-b_t)(V_\star - V^{(t)}) \\
&\leq \sum_{t=t_0+1}^{N} (1-b_t)\left(\widetilde{V}^{(t),\mathsf{sd}} - \bar{V}^{(t),\mathsf{sd}}\right) \\
&\quad + \beta_N(\delta)\left(1+\sqrt{H\Psi_N}\right)\sqrt{8Nd\max\{\kappa,1\}\log\left(1+\frac{N}{\kappa d}\right)} + 4\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)}.
\end{aligned}$$

Finally under event $\mathcal{E}_3^{\mathsf{sd}}$ we have a bound on the first term on the right hand side above, this leads to the bound

$$\begin{aligned}
&\sum_{t=t_0+1}^{N}(1-b_t)(V_\star - V^{(t)}) \\
&\leq (2H+1)\sum_{t=t_0+1}^{N}\sum_{h=1}^{H-1}\xi^{(t)}_{s_h^{(t)},a_h^{(t)}} + 4H^2\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)} \\
&+ \beta_N(\delta)\left(1+\sqrt{H\Psi_N}\right)\sqrt{8Nd\max\{\kappa,1\}\log\left(1+\frac{N}{\kappa d}\right)} + 4\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)} + 1. \quad (57)
\end{aligned}$$

It remains to bound the term $\sum_{t=t_0+1}^{N}\sum_{h=1}^{H-1}\xi^{(t)}_{s_h^{(t)},a_h^{(t)}}$. By mirroring the logic used to arrive at inequality (40) we can show that

$$\begin{aligned}
\sum_{t=t_0+1}^{N}\sum_{h=1}^{H-1}\xi^{(t)}_{s_h^{(t)},a_h^{(t)}} \leq 8|\mathcal{S}||\mathcal{A}|\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8H^2N)^{|\mathcal{S}|}\log(N)}{\delta}\right) \\
+ 8\sqrt{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8NH^2)^{|\mathcal{S}|}\log(N))}{\delta}\right)|\mathcal{S}||\mathcal{A}|N}.
\end{aligned}$$

Plugging this upper bound into inequality (57) we get

$$\begin{aligned}
&\sum_{t=t_0+1}^{N}(1-b_t)(V_\star - V^{(t)}) \\
&\leq 8(2H+1)|\mathcal{S}||\mathcal{A}|\cdot\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8H^2N)^{|\mathcal{S}|}\log(N)}{\delta}\right) \\
&\quad + 8(2H+1)\sqrt{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H(8NH^2)^{|\mathcal{S}|}\log(N))}{\delta}\right)|\mathcal{S}||\mathcal{A}|N} \\
&\quad + 4H^2\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)} + \beta_N(\delta)\left(1+\sqrt{H\Psi_N}\right)\sqrt{8Nd\max\{\kappa,1\}\log\left(1+\frac{N}{\kappa d}\right)} \\
&\quad + 4\sqrt{N\log\left(\frac{6\log(N)}{\delta}\right)} + 1.
\end{aligned}$$

Now finally, by using this upper bound in inequality (56) we find that

$$
\begin{aligned}
&\mathcal{R}(N) \\
&\leq C_3 \left[\frac{d^2 \log^2(d \log(1+\frac{16N}{d\omega^2}))}{\omega^4} \sqrt{\log(N/\delta)} + \left(\frac{d \log\left(1+\frac{16N}{d\omega^2}\right)}{\log(3/2)} (N_{\mathsf{EUL}} + N_{\mathsf{EVAL}}) \right)^{2/3} \right]^{3/2} \\
&\quad + \left(\frac{20}{3} \log\left(\frac{1}{\delta}\right) \right)^{3/2} + 4N^{2/3} \\
&\quad + 8(2H+1)|\mathcal{S}||\mathcal{A}| \cdot \log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H (8H^2 N)^{|\mathcal{S}|} \log(N)}{\delta} \right) \\
&\quad + 8(2H+1) \sqrt{\log\left(\frac{6(|\mathcal{S}||\mathcal{A}|H)^H (8NH^2)^{|\mathcal{S}|} \log(N))}{\delta} \right) |\mathcal{S}||\mathcal{A}|N} \\
&\quad + 4H^2 \sqrt{N \log\left(\frac{6\log(N)}{\delta} \right)} + \beta_N(\delta) \left(1 + \sqrt{H\Psi_N}\right) \sqrt{8Nd \max\{\kappa, 1\} \log\left(1 + \frac{N}{\kappa d}\right)} \\
&\quad + 4\sqrt{N \log\left(\frac{6\log(N)}{\delta} \right)} + 1 \qquad (58) \\
&= \widetilde{O}\left(\frac{\sqrt{\kappa H} d}{\omega} (d^3 + B^{3/2}) N^{2/3} + \left[H\sqrt{(H+|\mathcal{S}|)|\mathcal{S}||\mathcal{A}|} + H^2 \right] \sqrt{N} \right. \\
&\qquad \left. + (H+|\mathcal{S}|)H|\mathcal{S}||\mathcal{A}| + \frac{d^2}{\omega^2} \left(\frac{d^2}{\omega^2} + |\mathcal{S}|^2 |\mathcal{A}| H^2 \right) \right)
\end{aligned}
$$

where the last equality follows since by their definitions

$$
\begin{aligned}
N_{\mathsf{EUL}} &= \widetilde{\Theta}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|H^2}{\omega^2} \right); \qquad N_{\mathsf{EVAL}} = \widetilde{\Theta}\left(\frac{d^3}{\omega^4} \right); \\
\beta_N(\delta) &= \widetilde{O}\left(d^3 + B^{3/2} \right); \qquad \Psi_N = \widetilde{O}\left(\frac{dN^{1/3}}{\omega^2} \right),
\end{aligned}
$$

and by simplifying the expression in equation (58). This bound holds with probability $1 - 12N\delta$. Recalling that $\bar{\delta} = 12N\delta$ completes our proof. ■

E A Dynamic Programming Approach to Approximate $\pi^{(t)}$

In this section we present a computationally efficient dynamic programming algorithm that can be used to approximate the policy $\pi^{(t)}$ that is defined in equation (14) in Algorithm 3. We will also provide a proof for Proposition 3.7.

To avoid clashes of notation with the other sections of the paper we denote policies using θ here. We assume that we are given a transition dynamics model $\bar{\mathbb{P}}$, a vector $\mathbf{w} \in \mathbb{R}^d$, feature maps $\{\phi_h\}_{h\in[H]}$, a positive semi-definite matrix $\mathbf{\Sigma}$ and a bonus function $b_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for every $h \in [H]$. Also assume that there exists $\zeta > 0$ such that $\mathbf{w}^\top \phi(\tau) \in [-\zeta, \zeta]$, $\sum_h \|\phi_h(s_h, a_h)\|_{\mathbf{\Sigma}^{-1}} \in [0, \zeta]$ and $\sum_h b_h(s_h, a_h) \in [0, \zeta]$ for all $\tau \in \Gamma$. Finally let $w_h(s,a) := \mathbf{w}^\top \phi_h(s,a)$ and $v_h(s,a) := \|\phi_h(s,a)\|_{\mathbf{\Sigma}^{-1}}$.

Given a policy θ and an initial state s_1 define an optimistic value-function with this vector $\mathbf{w}$, feature maps, positive semi-definite matrix $\mathbf{\Sigma}$ and bonuses $\{b_h\}_{h\in[H]}$ as

$$\begin{aligned}
&\bar{V}^{\theta}_{\mathsf{opt}} \\
&:= \mathbb{E}_{s_1\sim\rho,\tau\sim\bar{\mathbb{P}}^{\theta}(\cdot|s_1)}\left[\min\left\{\mu\left(\mathbf{w}^{\top}\phi(\tau)\right)+\sum_{h=1}^{H}\|\phi_h(s_h,a_h)\|_{\mathbf{\Sigma}^{-1}},1\right\}+\sum_{h=1}^{H}b_h(s_h,a_h)\right] \\
&= \mathbb{E}_{s_1\sim\rho,\ \tau\sim\bar{\mathbb{P}}^{\theta}(\cdot|s_1)}\left[\min\left\{\mu\left(\sum_{h=1}^{H}w_h(s_h,a_h)\right)+\sum_{h=1}^{H}v_h(s_h,a_h),1\right\}+\sum_{h=1}^{H}b_h(s_h,a_h)\right].
\end{aligned}$$

Define the optimal policy with respect to this optimistic value function:

$$\theta_{\star}\in\arg\max_{\theta\in\Pi}\bar{V}^{\theta}_{\mathsf{opt}}.$$

Our goal is to find an ε-optimal policy $\widehat{\theta}=(\widehat{\theta}_1,\ldots,\widehat{\theta}_H)$ that satisfies

$$\bar{V}^{\theta_{\star}}_{\mathsf{opt}}-\bar{V}^{\widehat{\theta}}_{\mathsf{opt}}\le\varepsilon.$$

Also define the conditional optimistic value-function at any step $h\in[H]$:

$$\begin{aligned}
\bar{V}^{\theta}_h(s,\tau'_{h-1}) := \mathbb{E}_{\tau\sim\bar{\mathbb{P}}^{\theta}}\Bigg[\min\Bigg\{\mu\left(\sum_{\ell=1}^{H}w_\ell(s_\ell,a_\ell)\right)+\sum_{\ell=1}^{H}v_\ell(s_\ell,a_\ell),1\Bigg\} \\
+\sum_{\ell=1}^{H}b_\ell(s_\ell,a_\ell)\ \Bigg|\ s_h=s,\tau_{h-1}=\tau'_{h-1}\Bigg]. \qquad (59)
\end{aligned}$$

Define $m:=\left\lceil\frac{\zeta-(-\zeta)}{\varepsilon/(6H^2)}\right\rceil=\left\lceil\frac{12H^2\zeta}{\varepsilon}\right\rceil$ intervals

$$\begin{aligned}
\psi_j &:= \left[-\zeta+\frac{(j-1)\varepsilon}{6H^2},-\zeta+\frac{j\varepsilon}{6H^2}\right), \quad \text{if } j\in\{1,\ldots,m-1\} \\
\text{and,}\quad \psi_m &:= \left[-\zeta+\frac{(m-1)\varepsilon}{6H^2},\zeta\right].
\end{aligned}$$

The centers of these intervals are $\nu_j:=-\zeta+\frac{(j-\frac{1}{2})\varepsilon}{6H^2}$ for every $j\in[m]$. Define a map $\sigma:[-\zeta,\zeta]\to\{1,\ldots,m\}$ that maps each x to the index of interval that x lies in,

$$\sigma(x)=j,\quad \text{if } x\in\psi_j.$$

Our dynamic programming approach will require us to define tensors $\widehat{a}_h$ and $\widehat{V}_h$ for every $h\in[H]$. Given any quartet $(s,i,j,k)\in\mathcal{S}\times[m]\times[m]\times[m]$ define the following at the final step H

$$\begin{aligned}
\widehat{a}_H(s,i,j,k) &\in \arg\max_{a\in\mathcal{A}}\left\{\min\left\{\mu\left(\nu_i+w_H(s,a)\right)+\nu_j+v_H(s,a),1\right\}+\nu_k+b_H(s,a)\right\}; \\
\widehat{V}_H(s,i,j,k) &:= \max_{a\in\mathcal{A}}\left\{\min\left\{\mu\left(\nu_i+w_H(s,a)\right)+\nu_j+v_H(s,a),1\right\}+\nu_k+b_H(s,a)\right\}.
\end{aligned}$$

The action $\widehat{a}_H(s,i,j,k)$ is the optimal action when the state is s and the "histories" $\sum_{h=1}^{H-1}w_h(s_h,a_h)$, $\sum_{h=1}^{H-1}v_h(s_h,a_h)$ and $\sum_{h=1}^{H-1}b_h(s_h,a_h)$ are equal to ν_i, ν_j and ν_k respectively. Further, the tensor $\widehat{V}_H(s,i,j,k)$ stores the value of the conditional value function when this optimal action is taken given this quartet. Also recursively define the following in the preceding steps:

$$\begin{aligned}
\widehat{a}_h(s,i,j,k) &\in \arg\max_{a\in\mathcal{A}}\mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\left[\widehat{V}_{h+1}(s',\sigma(w_h(s,a)+\nu_i),\sigma(v_h(s,a)+\nu_j),\sigma(b_h(s,a)+\nu_k))\right]; \\
\widehat{V}_h(s,i,j,k) &:= \max_{a\in\mathcal{A}}\mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\left[\widehat{V}_{h+1}(s',\sigma(w_h(s,a)+\nu_i),\sigma(v_h(s,a)+\nu_j),\sigma(b_h(s,a)+\nu_k))\right].
\end{aligned}$$

At the initial step $h=1$ the expectation over the states $s'\sim\bar{\mathbb{P}}(\cdot|s,a)$ in the definition above is replaced by the expectation over the initial state $s_1\sim\rho$.

To construct $\widehat{\theta}$, our strategy will be to use these near-optimal actions $(\widehat{a}_h)$ at every step over this "$\frac{\varepsilon}{6H^2}$-net" of representative histories that is defined. Then given any state and sub-trajectory we will map this state and sub-trajectory to its nearest neighbor in the net of histories and play the near-optimal action corresponding to this neighbor. To this end define the maps

$$i_h(\tau_{h-1}) := \sigma\left(\sum_{\ell=1}^{h-1} w_\ell(s_\ell, a_\ell)\right), \tag{60a}$$

$$j_h(\tau_{h-1}) := \sigma\left(\sum_{\ell=1}^{h-1} v_\ell(s_\ell, a_\ell)\right), \quad \text{and} \tag{60b}$$

$$k_h(\tau_{h-1}) := \sigma\left(\sum_{\ell=1}^{h-1} b_\ell(s_\ell, a_\ell)\right). \tag{60c}$$

At times we will use i_h, j_h and k_h as shorthand for $i_h(\tau_{h-1}), j_h(\tau_{h-1})$ and $k_h(\tau_{h-1})$ respectively. Given a state s and sub-trajectory τ_{h-1} the policy at step $h \in [H]$, $\widehat{\theta}_h(\cdot|s, \tau_{h-1})$ puts all of its mass on the action

$$\widehat{a}_h\left(s_h, i_h(\tau_{h-1}), j_h(\tau_{h-1}), k_h(\tau_{h-1})\right)$$

(where we break ties among actions arbitrarily). Given a policy θ, let $\theta_{h:H} = (\theta_h, \ldots, \theta_H)$ denote the set of policies from step h onward. Let $\tilde{P}^{\theta_{h:H}}(\cdot|s)$ denote the distribution of the trajectory in the steps $h, \ldots, H$ given that the state at step $h-1$ was s. Finally define the extended conditional value-functions for the policy $\widehat{\theta}$ to be

$$\begin{aligned}
&\breve{V}_h^{\widehat{\theta}_{h+1:H}}(s, \alpha, \beta, \gamma) := \\
&\mathbb{E}_{\tau \sim \tilde{\mathbb{P}}^{\widehat{\theta}_{h+1:H}}(\cdot|s)}\Bigg[\min\Bigg\{\mu\left(\alpha + w_h(s, \widehat{a}_h(s, \sigma(\alpha), \sigma(\beta), \sigma(\gamma))) + \sum_{\ell=h+1}^{H} w_\ell(s_\ell, a_\ell)\right) \\
&\qquad\qquad +\beta + v_h(s, \widehat{a}_h(s, \sigma(\alpha), \sigma(\beta), \sigma(\gamma))) + \sum_{\ell=h+1}^{H} v_\ell(s_\ell, a_\ell), 1\Bigg\} \\
&\qquad\qquad\qquad +\gamma + b_h(s, \widehat{a}_h(s, \sigma(\alpha), \sigma(\beta), \sigma(\gamma))) + \sum_{\ell=h+1}^{H} b_\ell(s_\ell, a_\ell)\Bigg]
\end{aligned}$$

for any $h \in [H]$, $s \in \mathcal{S}$, $\alpha \in [-\zeta, \zeta]$, $\beta \in [0, \zeta]$ and $\gamma \in [0, \zeta]$. In the definition above the expectation is over the steps $h+1, \ldots, H$. The extended value function is the definition of the conditional value function by using the summary of the history: α, β and γ.

E.1 The Policy $\widehat{\theta}$ is ε-Optimal

The following lemma shows that the policy $\widehat{\theta}$ is ε-optimal and can be found efficiently. We shall use this lemma to prove Proposition 3.7 below.

Lemma E.1. *The policy $\widehat{\theta}$ satisfies*

$$\bar{V}_{\mathsf{opt}}^{\theta_\star} - \bar{V}_{\mathsf{opt}}^{\widehat{\theta}} \leq \varepsilon.$$

Furthermore the policy $\widehat{\theta}$ can be found in $\mathsf{poly}\left(|\mathcal{S}|, |\mathcal{A}|, H, \zeta, \frac{1}{\varepsilon}\right)$ *time and memory.*

Proof The proof shall proceed in two steps. First, we shall show via an inductive argument that certain properties are satisfied at all steps. In the second part we will use these properties to prove the lemma.

Part I: The inductive hypothesis. The induction will be over the steps $H, \ldots, 1$. We shall inductively show that:

(a) For any $s \in \mathcal{S}$, $\alpha \in [-\zeta, \zeta]$, $\beta \in [0, \zeta]$ and $\gamma \in [0, \zeta]$:

$$\left|\breve{V}_h^{\widehat{\theta}_{h:H}}(s, \alpha, \beta, \gamma) - \widehat{V}_h(s, \sigma(\alpha), \sigma(\beta), \sigma(\gamma))\right| \leq \frac{(H+1-h)\varepsilon}{2H^2};$$

(b) for any $s \in \mathcal{S}$ and $\tau_{h-1} \in \Gamma_{h-1}$

$$\max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s,a)} \left[\bar{V}_{h+1}^{\widehat{\theta}_{h+1:H}}(s', \{s, a, \tau_{h-1}\}) \right] - \bar{V}_h^{\widehat{\theta}_{h:H}}(s, \tau_{h-1}) \leq \frac{(H+1-h)\varepsilon}{H^2};$$

(c) given the tensor $\widehat{V}_{h+1}$ it is possible to find $\widehat{a}_h(s,i,j,k)$ and $\widehat{V}_h(s,i,j,k)$ for all quartets using poly $\left(|\mathcal{S}|, |\mathcal{A}|, H, \zeta, \frac{1}{\varepsilon}\right)$ time and memory.

Note that $\max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s,a)} \left[\bar{V}_{h+1}^{\widehat{\theta}_{h+1:H}}(s', \{s, a, \tau_{h-1}\}) \right]$ corresponds to the conditional-value (see the Definition of $\bar{V}$ in equation (59)) of taking the best action at step h when the policy for the future steps is $\widehat{\theta}_{h+1:H}$.

Base case: The base case of the induction is at step H.

*Part (a):*Fix an α, β and γ and define the shorthand $\widehat{a}_H := \widehat{a}_H(s, \sigma(\alpha), \sigma(\beta), \sigma(\gamma))$. By the definition of $\breve{V}_H$, $\widehat{V}_H$ and the policy $\widehat{\theta}$ we have

$$\begin{aligned}
&\left| \breve{V}_H^{\widehat{\theta}_H}(s, \alpha, \beta, \gamma) - \widehat{V}_H(s, \sigma(\alpha), \sigma(\beta), \sigma(\gamma)) \right| \\
&\quad = \Bigg| \min \Big\{ \mu\left(\alpha + w_H(s, \widehat{a}_h)\right) + \beta + v_H(s, \widehat{a}_H), 1 \Big\} + \gamma + b_H(s, \widehat{a}_H) \\
&\qquad\quad - \min \left\{ \mu\left(\nu_{\sigma(\alpha)} + w_H(s, \widehat{a}_H)\right) + \nu_{\sigma(\beta)} + v_H(s, \widehat{a}_H), 1 \right\} + \nu_{\sigma(\gamma)} + b_H(s, \widehat{a}_H) \Bigg| \\
&\quad \overset{(i)}{\leq} \left| \mu\left(\alpha + w_H(s, \widehat{a}_h)\right) - \mu\left(\nu_{\sigma(\alpha)} + w_H(s, \widehat{a}_H)\right) \right| + |\beta - \nu_{\sigma(\beta)}| + |\gamma - \nu_{\sigma(\gamma)}| \\
&\quad \overset{(ii)}{\leq} \left| \alpha - \nu_{\sigma(\alpha)} \right| + |\beta - \nu_{\sigma(\beta)}| + |\gamma - \nu_{\sigma(\gamma)}| \overset{(iii)}{\leq} 3 \times \frac{\varepsilon}{6H^2} = \frac{\varepsilon}{2H^2}
\end{aligned}$$

where (i) follows since the function $z \mapsto \min(z, 1)$ is 1-Lipschitz and by the triangle inequality, and (ii) follows since μ is 1-Lipschitz, and (iii) follows by the definition of the function σ, that projects a number onto a grid with granularity $\varepsilon/(6H^2)$.

Part (b): An episode terminates at the end of step H, therefore we define $\bar{V}_{H+1}(s', \tau_H) := \min \left\{ \mu\left(\sum_{h=1}^H w_h(s_h, a_h)\right) + \sum_{h=1}^H v_h(s_h, a_h), 1 \right\} + \sum_{h=1}^H b_h(s_h, a_h)$. Thus, by the definition of the extended conditional value function $\breve{V}_H$

$$\begin{aligned}
&\max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s,a)} \left[\bar{V}_{H+1}(s', \{s, a, \tau_{H-1}\}) \right] - \bar{V}_H^{\widehat{\theta}_H}(s, \tau_{H-1}) \\
&= \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s,a)} \left[\bar{V}_{H+1}(s', \{s, a, \tau_{H-1}\}) \right] - \breve{V}_H^{\widehat{\theta}_H} \left(s, \sum_{\ell=1}^{H-1} w_\ell(s_\ell, a_\ell), \sum_{\ell=1}^{H-1} v_\ell(s_\ell, a_\ell), \sum_{\ell=1}^{H-1} b_\ell(s_\ell, a_\ell) \right) \\
&\overset{(i)}{=} \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s,a)} \left[\bar{V}_{H+1}(s', \{s, a, \tau_{H-1}\}) \right] - \widehat{V}_H^{\widehat{\theta}_H}(s, i_H, j_H, k_H) \\
&\qquad\qquad + \widehat{V}_H^{\widehat{\theta}_H}(s, i_H, j_H, k_H) - \breve{V}_H^{\widehat{\theta}_H} \left(s, \sum_{\ell=1}^{H-1} w_\ell(s_\ell, a_\ell), \sum_{\ell=1}^{H-1} v_\ell(s_\ell, a_\ell), \sum_{\ell=1}^{H-1} b_\ell(s_\ell, a_\ell) \right) \\
&\overset{(ii)}{\leq} \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s,a)} \left[\bar{V}_{H+1}(s', \{s, a, \tau_{H-1}\}) \right] - \widehat{V}_H^{\widehat{\theta}_H}(s, i_H, j_H, k_H) + \frac{\varepsilon}{2H^2}
\end{aligned}$$

where in (i) recall the definitions of i_H, j_H and k_H from above in equations (60a)-(60c), and (ii) follows by Part (a) of the induction hypothesis. Continuing

$$
\begin{aligned}
&\max_{a\in\mathcal{A}} \mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\left[\bar{V}_{H+1}(s',\{s,a,\tau_{H-1}\})\right] - \bar{V}_H^{\widehat{\theta}_H}(s,\tau_{H-1}) \\
&\overset{(i)}{\leq} \max_{a\in\mathcal{A}} \Bigg\{ \min\left\{\mu\left(\sum_{h=1}^{H-1} w_h(s_h,a_h) + w_H(s,a)\right) + \sum_{h=1}^{H-1} v_h(s_h,a_h) + v_H(s,a), 1\right\} \\
&\qquad\qquad + \sum_{h=1}^{H-1} b_h(s_h,a_h) + b_H(s,a)\Bigg\} \\
&\qquad - \max_{a'\in\mathcal{A}} \left\{\min\left\{\mu\left(\nu_{i_H} + w_H(s,a') + \nu_{j_H} + v_H(s,a')\right), 1\right\} + \nu_{k_H} + b_H(s,a')\right\} + \frac{\varepsilon}{2H^2} \\
&\leq \max_{a\in\mathcal{A}} \Bigg\{ \min\left\{\mu\left(w_H(s,a) + \sum_{h=1}^{H-1} w_h(s_h,a_h)\right) + \sum_{h=1}^{H-1} v_h(s_h,a_h) + v_H(s,a), 1\right\} \\
&\qquad\qquad + \sum_{h=1}^{H-1} b_h(s_h,a_h) + b_H(s,a) \\
&\qquad\qquad - \min\left\{\mu\left(w_H(s,a) + \nu_{i_H}\right) + v_H(s,a) + \nu_{j_H}, 1\right\} - \nu_{k_H} - b_H(s,a)\Bigg\} + \frac{\varepsilon}{2H^2} \\
&\overset{(ii)}{\leq} \left|\sum_{h=1}^{H-1} w_h(s_h,a_h) - \nu_{i_H}\right| + \left|\sum_{h=1}^{H-1} v_h(s_h,a_h) - \nu_{j_H}\right| + \left|\sum_{h=1}^{H-1} b_h(s_h,a_h) - \nu_{k_H}\right| + \frac{\varepsilon}{2H^2} \overset{(iii)}{\leq} \frac{\varepsilon}{H^2}.
\end{aligned}
$$

where (i) follows by the definition of $\widehat{V}_H(s,i_H,j_H,k_H)$, (ii) follows because the functions $z \mapsto \min\{z,1\}$ and $z \mapsto \frac{1}{1+\exp(-z)}$ are both 1-Lipschitz, and (iii) follows by the definition of the maps i_H, j_H and k_H, and the intervals ψ_j. This proves the second part of the inductive hypothesis in the base case.

Part (c): Let's show $\widehat{a}_H(s,i,j,k)$ and $\widehat{V}_H(s,i,j,k)$ can be computed efficiently. Fix a quartet $(s,i,j,k) \in \mathcal{S}\times[m]\times[m]\times[m]$. Then the values

$$
\begin{aligned}
\widehat{a}_H(s,i,j,k) &\in \arg\max_{a\in\mathcal{A}} \left\{\min\left\{\mu\left(\nu_i + w_H(s,a)\right) + \nu_j + v_H(s,a), 1\right\} + \nu_k + b_H(s,a)\right\} \\
\widehat{V}_H(s,i,j,k) &= \max_{a\in\mathcal{A}} \left\{\min\left\{\mu\left(\nu_i + w_H(s,a)\right) + \nu_j + v_H(s,a), 1\right\} + \nu_k + b_H(s,a)\right\}
\end{aligned}
$$

can be found using $\mathsf{poly}(|\mathcal{A}|)$ time and memory. Therefore, the entire tensor can be found using $|\mathcal{S}|m^3 \times \mathsf{poly}(|\mathcal{A}|) = \mathsf{poly}(|\mathcal{S}|,|\mathcal{A}|,H,\zeta,\frac{1}{\varepsilon})$ time and memory.

Induction step: Assume that the induction hypothesis holds at the steps $H,\ldots,h+1$. We will now prove that each part of the induction hypothesis also holds at the step $h \geq 1$.

Part (a): Fix an α, β and γ and let's define the shorthand $\widehat{a}_h = \widehat{a}_h(s, \sigma(\alpha), \sigma(\beta), \sigma(\gamma))$. Hence[2],

$$
\begin{aligned}
&\left|\breve{V}_h^{\widehat{\theta}_{h:H}}(s, \alpha, \beta, \gamma) - \widehat{V}_h(s, \sigma(\alpha), \sigma(\beta), \sigma(\gamma))\right| \\
&= \Bigg|\mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s, \widehat{a}_h(s, \tau_{h-1}))}\Big[\breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \alpha + w_h(s, \widehat{a}_h), \beta + v_h(s, \widehat{a}_h), \gamma + b_h(s, \widehat{a}_h)\right) \\
&\qquad\qquad - \widehat{V}_{h+1}\left(s', \sigma\left(w_h(s, \widehat{a}_h) + \nu_{\sigma(\alpha)}\right), \sigma\left(v_h(s, \widehat{a}_h) + \nu_{\sigma(\beta)}\right), \sigma\left(b_h(s, \widehat{a}_h) + \nu_{\sigma(\gamma)}\right)\right)\Big]\Bigg| \\
&= \Bigg|\mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s, \widehat{a}_h(s, \tau_{h-1}))}\Big[\breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \alpha + w_h(s, \widehat{a}_h), \beta + v_h(s, \widehat{a}_h), \gamma + b_h(s, \widehat{a}_h)\right) \\
&\qquad\qquad - \breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', w_h(s, \widehat{a}_h) + \nu_{\sigma(\alpha)}, v_h(s, \widehat{a}_h) + \nu_{\sigma(\beta)}, b_h(s, \widehat{a}_h) + \nu_{\sigma(\gamma)}\right) \\
&\qquad\qquad + \breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', w_h(s, \widehat{a}_h) + \nu_{\sigma(\alpha)}, v_h(s, \widehat{a}_h) + \nu_{\sigma(\beta)}, b_h(s, \widehat{a}_h) + \nu_{\sigma(\gamma)}\right) \\
&\qquad\qquad - \widehat{V}_{h+1}\left(s', \sigma\left(w_h(s, \widehat{a}_h) + \nu_{\sigma(\alpha)}\right), \sigma\left(v_h(s, \widehat{a}_h) + \nu_{\sigma(\beta)}\right), \sigma\left(b_h(s, \widehat{a}_h) + \nu_{\sigma(\gamma)}\right)\right)\Big]\Bigg| \\
&\leq \Bigg|\mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s, \widehat{a}_h(s, \tau_{h-1}))}\Big[\breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \alpha + w_h(s, \widehat{a}_h), \beta + v_h(s, \widehat{a}_h), \gamma + b_h(s, \widehat{a}_h)\right) \\
&\qquad - \breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', w_h(s, \widehat{a}_h) + \nu_{\sigma(\alpha)}, v_h(s, \widehat{a}_h) + \nu_{\sigma(\beta)}, b_h(s, \widehat{a}_h) + \nu_{\sigma(\gamma)}\right)\Big]\Bigg| + \frac{(H-h)\varepsilon}{2H^2}
\end{aligned} \tag{61}
$$

where the last inequality follows by Part (a) of the inductive hypothesis at step $h+1$. Let us now bound

$$
\begin{aligned}
&\left|\breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \alpha + w_h(s, \widehat{a}_h), \beta + b_h(s, \widehat{a}_h)\right) - \breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', w_h\left(s, \widehat{a}_h\right) + \nu_{\sigma(\alpha)}, b_h(s, \widehat{a}_h) + \nu_{\sigma(\beta)}\right)\right| \\
&= \Bigg|\mathbb{E}_{\tau \sim \bar{\mathbb{P}}^{\widehat{\theta}_{h+1:H}}}\Bigg[\min\left\{\mu\left(\alpha + w_h(s, \widehat{a}_h) + \sum_{\ell=h+1}^{H} w_\ell(s_\ell, a_\ell)\right) + \beta + v_h(s, \widehat{a}_h) + \sum_{\ell=h+1}^{H} v_\ell(s_\ell, a_\ell), 1\right\} \\
&\qquad + \gamma + b_h(s, \widehat{a}_h) + \sum_{\ell=h+1}^{H} b_\ell(s_\ell, a_\ell) \;\Bigg|\; s_{h+1} = s', \tau_h = \{s, \widehat{a}_h, \tau_{h-1}\}\Bigg] \\
&- \mathbb{E}_{\tau \sim \bar{\mathbb{P}}^{\widehat{\theta}_{h+1:H}}}\Bigg[\min\left\{\mu\left(\nu_{\sigma(\alpha)} + w_h(s, \widehat{a}_h) + \sum_{\ell=h+1}^{H} w_\ell(s_\ell, a_\ell)\right) + \nu_{\sigma(\beta)} + v_h(s, \widehat{a}_h) + \sum_{\ell=h+1}^{H} v_\ell(s_\ell, a_\ell), 1\right\} \\
&\qquad + \nu_{\sigma(\gamma)} + b_h(s, \widehat{a}_h) + \sum_{\ell=h+1}^{H} b_\ell(s_\ell, a_\ell) \;\Bigg|\; s_{h+1} = s', \tau_h = \{s, \widehat{a}_h, \tau_{h-1}\}\Bigg]\Bigg|.
\end{aligned}
$$

Since the functions $z \mapsto \min\{z, 1\}$ and $z \mapsto \frac{1}{1+\exp(-z)}$ are 1-Lipschitz, therefore

$$
\begin{aligned}
&\left|\breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \alpha + w_h(s, \widehat{a}_h), \beta + b_h(s, \widehat{a}_h)\right) - \breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', w_h\left(s, \widehat{a}_h\right) + \nu_{\sigma(\alpha)}, b_h(s, \widehat{a}_h) + \nu_{\sigma(\beta)}\right)\right| \\
&\qquad \leq |\alpha - \nu_{\sigma(\alpha)}| + |\beta - \nu_{\sigma(\beta)}| + |\gamma - \nu_{\sigma(\gamma)}| \leq \frac{\varepsilon}{2H^2}.
\end{aligned} \tag{62}
$$

This combined with inequality (61) shows that

$$
\left|\breve{V}_h^{\widehat{\theta}_{h:H}}(s, \alpha, \beta) - \widehat{V}_h(s, \sigma(\alpha), \sigma(\beta))\right| \leq \frac{(H+1-h)\varepsilon}{2H^2}
$$

and completes the proof of the first part of the induction step.

[2] In the arguments that follow when $h = 1$, the outer expectation $\mathbb{E}_{s' \sim \bar{P}(\cdot|s, \widehat{a}_h(s, \tau_{h-1}))}$ is replaced by $\mathbb{E}_{s_1 \sim \rho}$ however the same arguments remain unchanged.

Part (b): Here let $\widehat{a}_h$ be shorthand for $\widehat{a}_h(s, \sigma(\sum_{\ell=1}^{h-1} w_\ell(s_\ell, a_\ell)), \sigma(\sum_{\ell=1}^{h-1} v_\ell(s_\ell, a_\ell)), \sigma(\sum_{\ell=1}^{h-1} b_\ell(s_\ell, a_\ell)))$. Since the policy $\widehat{\theta}_h$ picks the action $\widehat{a}_h$

$$
\begin{aligned}
&\bar{V}_h^{\widehat{\theta}_{h:H}}(s, \tau_{h-1}) \\
&\quad = \mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s,\widehat{a}_h)} \left[\bar{V}_{h+1}^{\widehat{\theta}_{h+1:H}}(s', \{s, \widehat{a}_h, \tau_{h-1}\}) \right] \\
&\quad = \mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s,\widehat{a}_h)} \Bigg[\bar{V}_{h+1}^{\widehat{\theta}_{h+1:H}}(s', \{s, \widehat{a}_h, \tau_{h-1}\}) \\
&\qquad\quad - \breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', w_h(s, \widehat{a}_h) + \nu_{i_h}, v_h(s, \widehat{a}_h) + \nu_{j_h}, b_h(s, \widehat{a}_h) + \nu_{k_h}\right) \Bigg] \\
&\qquad + \mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s,\widehat{a}_h)} \left[\breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', w_h(s, \widehat{a}_h) + \nu_{i_h}, v_h(s, \widehat{a}_h) + \nu_{j_h}, b_h(s, \widehat{a}_h) + \nu_{k_h}\right) \right]. \quad (63)
\end{aligned}
$$

We know that the difference of the first two terms in the expectation above

$$
\begin{aligned}
&\bar{V}_{h+1}^{\widehat{\theta}_{h+1:H}}(s', \{s, \widehat{a}_h, \tau_{h-1}\}) - \breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', w_h(s, \widehat{a}_h) + \nu_{i_h}, v_h(s, \widehat{a}_h) + \nu_{j_h}, b_h(s, \widehat{a}_h) + \nu_{k_h}\right) \\
&= \mathbb{E}_{\tau \sim \bar{\mathbb{P}}^{\widehat{\theta}_{h+1:H}}} \Bigg[\min\left\{ \mu\Big(\sum_{\ell=1}^{h-1} w_\ell(s_\ell, a_\ell) + w_h(s, \widehat{a}_h) + \sum_{\ell=h+1}^{H} w_\ell(s_\ell, a_\ell) \Big) \right. \\
&\qquad\qquad \left. + \sum_{\ell=1}^{h-1} v_\ell(s_\ell, a_\ell) + v_h(s, \widehat{a}_h) + \sum_{\ell=h+1}^{H} v_\ell(s_\ell, a_\ell), 1 \right\} \\
&\qquad\quad + \sum_{\ell=1}^{h-1} b_\ell(s_\ell, a_\ell) + b_h(s, \widehat{a}_h) + \sum_{\ell=h+1}^{H} b_\ell(s_\ell, a_\ell) \\
&\qquad - \min\left\{ \mu\Big(\nu_{i_h} + w_h(s, \widehat{a}_h) + \sum_{\ell=h+1}^{H} w_\ell(s_\ell, a_\ell) \Big) \right. \\
&\qquad\qquad \left. + \nu_{j_h} + v_h(s, \widehat{a}_h) + \sum_{\ell=h+1}^{H} v_\ell(s_\ell, a_\ell), 1 \right\} \\
&\qquad - \nu_{k_h} - b_h(s, \widehat{a}_h) - \sum_{\ell=h+1}^{H} b_\ell(s_\ell, a_\ell) \;\Big|\; s, \widehat{a}_h, \tau_{h-1} \Bigg] \\
&\overset{(i)}{\geq} - \left| \sum_{\ell=1}^{h-1} w_\ell(s_\ell, a_\ell) - \nu_{i_h} \right| - \left| \sum_{\ell=1}^{h-1} v_\ell(s_\ell, a_\ell) - \nu_{j_h} \right| - \left| \sum_{\ell=1}^{h-1} b_\ell(s_\ell, a_\ell) - \nu_{k_h} \right| \overset{(ii)}{\geq} -\frac{\varepsilon}{2H^2}.
\end{aligned}
$$

where (i) follows since the functions $z \mapsto \min\{z, 1\}$ and $z \mapsto \frac{1}{1+\exp(-z)}$ are 1-Lipschitz and (ii) follows since ν_{i_h}, ν_{j_h} and ν_{k_h} are the nearest neighbors of $\sum_{\ell=1}^{h-1} w_\ell(s_\ell, a_\ell)$, $\sum_{\ell=1}^{h-1} v_\ell(s_\ell, a_\ell)$ and $\sum_{\ell=1}^{h-1} b_\ell(s_\ell, a_\ell)$ respectively in the $\frac{\varepsilon}{6H^2}$ grid. This previous inequality combined with equation (63) yields

$$
\begin{aligned}
&\bar{V}_h^{\widehat{\theta}_{h:H}}(s, \tau_{h-1}) \\
&\qquad \geq \mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s,\widehat{a}_h)} \left[\breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', w_h(s, \widehat{a}_h) + \nu_{i_h}, v_h(s, \widehat{a}_h) + \nu_{j_h}, b_h(s, \widehat{a}_h) + \nu_{k_h}\right) \right] - \frac{\varepsilon}{2H^2}.
\end{aligned}
$$

This relates the true conditional-value function to the extended value function $\breve{V}$. We will now continue further to relate the true conditional-value function to the surrogate $\widehat{V}$ that we can compute

on the grid of histories. Continuing from the previous display above we get

$$
\begin{aligned}
&\bar{V}_h^{\widehat{\theta}_{h:H}}(s,\tau_{h-1}) \\
&\geq \mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,\widehat{a}_h)}\Big[\check{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', w_h(s,\widehat{a}_h)+\nu_{i_h}, v_h(s,\widehat{a}_h)+\nu_{j_h}, b_h(s,\widehat{a}_h)+\nu_{k_h}\right) \\
&\quad - \widehat{V}_{h+1}\left(s', \sigma\left(\nu_{i_h}+w_h(s,\widehat{a}_h)\right), \sigma\left(\nu_{j_h}+v_h(s,\widehat{a}_h)\right), \sigma\left(\nu_{k_h}+b_h(s,\widehat{a}_h)\right)\right)\Big] \\
&\quad + \mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,\widehat{a}_h)}\left[\widehat{V}_{h+1}\left(s', \sigma\left(\nu_{i_h}+w_h(s,\widehat{a}_h)\right), \sigma\left(\nu_{j_h}+v_h(s,\widehat{a}_h)\right), \sigma\left(\nu_{k_h}+b_h(s,\widehat{a}_h)\right)\right)\right] - \frac{\varepsilon}{2H^2} \\
&\overset{(i)}{\geq} \mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,\widehat{a}_h)}\left[\widehat{V}_{h+1}\left(s', \sigma\left(\nu_{i_h}+w_h(s,\widehat{a}_h)\right), \sigma\left(\nu_{j_h}+v_h(s,\widehat{a}_h)\right), \sigma\left(\nu_{k_h}+b_h(s,\widehat{a}_h)\right)\right)\right] \\
&\qquad\qquad - \frac{(H-h)\varepsilon}{2H^2} - \frac{\varepsilon}{2H^2} \\
&\overset{(ii)}{=} \max_{a\in\mathcal{A}} \mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\left[\widehat{V}_{h+1}\left(s', \sigma\left(\nu_{i_h}+w_h(s,a)\right), \sigma\left(\nu_{j_h}+v_h(s,a)\right), \sigma\left(\nu_{k_h}+b_h(s,a)\right)\right)\right] \\
&\qquad\qquad - \frac{(H+1-h)\varepsilon}{2H^2}
\end{aligned} \tag{64}
$$

where (i) follows by using the first part of the induction hypothesis at step $h+1$ and (ii) follows by the definition of $\widehat{a}_h$. With this lower bound in place let us now establish a bound on the quantity of interest

$$
\begin{aligned}
&\max_{a\in\mathcal{A}} \mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\left[\bar{V}_{h+1}^{\widehat{\theta}_{h+1:H}}(s', \{s,a,\tau_{h-1}\})\right] - \bar{V}_h^{\widehat{\theta}_{h:H}}(s,\tau_{h-1}) \\
&\overset{(i)}{\leq} \max_{a\in\mathcal{A}}\left\{\mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\left[\bar{V}_{h+1}^{\widehat{\theta}_{h+1:H}}(s', \{s,a,\tau_{h-1}\})\right]\right\} \\
&\qquad - \max_{a\in\mathcal{A}}\left\{\mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\left[\widehat{V}_{h+1}\left(s', \sigma\left(\nu_{i_h}+w_h(s,a)\right), \sigma\left(\nu_{j_h}+v_h(s,a)\right), \sigma\left(\nu_{k_h}+b_h(s,a)\right)\right)\right]\right\} \\
&\qquad\qquad + \frac{(H+1-h)\varepsilon}{2H^2} \\
&\leq \max_{a\in\mathcal{A}}\Big\{\mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\Big[\bar{V}_{h+1}^{\widehat{\theta}_{h+1:H}}(s', \{s,a,\tau_{h-1}\}) \\
&\qquad\qquad - \widehat{V}_{h+1}\left(s', \sigma\left(\nu_{i_h}+w_h(s,a)\right), \sigma\left(\nu_{j_h}+v_h(s,a)\right), \sigma\left(\nu_{k_h}+b_h(s,a)\right)\right)\Big]\Big\} \\
&\qquad\qquad + \frac{(H+1-h)\varepsilon}{2H^2}
\end{aligned} \tag{65}
$$

where (i) follows by invoking inequality (64). Note that by its definition

$$
\begin{aligned}
&\check{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \sum_{\ell=1}^{h-1} w_\ell(s_\ell,a_\ell) + w_h(s,a), \sum_{\ell=1}^{h-1} v_\ell(s_\ell,a_\ell) + v_h(s,a), \sum_{\ell=1}^{h-1} b_\ell(s_\ell,a_\ell) + b_h(s,a)\right) \\
&\qquad\qquad = \bar{V}_{h+1}^{\widehat{\theta}_{h+1:H}}(s', \{s,a,\tau_{h-1}\}),
\end{aligned}
$$

therefore continuing from inequality (65)

$$
\begin{aligned}
&\max_{a\in\mathcal{A}} \mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\left[\bar{V}_{h+1}^{\widehat{\theta}_{h+1:H}}(s',\{s,a,\tau_{h-1}\})\right] - \bar{V}_{h}^{\widehat{\theta}_{h:H}}(s,\tau_{h-1}) \\
&\leq \max_{a\in\mathcal{A}}\Bigg\{\mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\Bigg[\\
&\qquad \breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \sum_{\ell=1}^{h-1} w_\ell(s_\ell,a_\ell) + w_h(s,a), \sum_{\ell=1}^{h-1} v_\ell(s_\ell,a_\ell) + v_h(s,a), \sum_{\ell=1}^{h-1} b_\ell(s_\ell,a_\ell) + b_h(s,a)\right) \\
&\qquad\qquad - \widehat{V}_{h+1}\left(s', \sigma\left(\nu_{i_h} + w_h(s,a)\right), \sigma\left(\nu_{j_h} + v_h(s,a)\right), \sigma\left(\nu_{k_h} + b_h(s,a)\right)\right)\Bigg]\Bigg\} \\
&\qquad\qquad\qquad + \frac{(H+1-h)\varepsilon}{2H^2} \\
&\leq \max_{a\in\mathcal{A}}\Bigg\{\mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\Bigg[\\
&\qquad \breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \sum_{\ell=1}^{h-1} w_\ell(s_\ell,a_\ell) + w_h(s,a), \sum_{\ell=1}^{h-1} v_\ell(s_\ell,a_\ell) + v_h(s,a), \sum_{\ell=1}^{h-1} b_\ell(s_\ell,a_\ell) + b_h(s,a)\right) \\
&\qquad\qquad - \breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \nu_{i_h} + w_h(s,a), \nu_{j_h} + v_h(s,a), \nu_{k_h} + b_h(s,a)\right) \\
&\qquad\qquad + \breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \nu_{i_h} + w_h(s,a), \nu_{j_h} + v_h(s,a), \nu_{k_h} + b_h(s,a)\right) \\
&\qquad\qquad - \widehat{V}_{h+1}\left(s', \sigma\left(\nu_{i_h} + w_h(s,a)\right), \sigma\left(\nu_{j_h} + v_h(s,a)\right), \sigma\left(\nu_{k_h} + b_h(s,a)\right)\right)\Bigg]\Bigg\} \\
&\qquad\qquad\qquad + \frac{(H+1-h)\varepsilon}{2H^2} \\
&\overset{(i)}{\leq} \max_{a\in\mathcal{A}}\Bigg\{\mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\Bigg[\breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \nu_{i_h} + w_h(s,a), \nu_{j_h} + v_h(s,a), \nu_{k_h} + b_h(s,a)\right) \\
&\qquad\qquad - \widehat{V}_{h+1}\left(s', \sigma\left(\nu_{i_h} + w_h(s,a)\right), \sigma\left(\nu_{j_h} + v_h(s,a)\right), \sigma\left(\nu_{k_h} + b_h(s,a)\right)\right)\Bigg]\Bigg\} \\
&\qquad\qquad + \frac{(H+1-h)\varepsilon}{2H^2} + \frac{\varepsilon}{2H^2} \\
&\overset{(ii)}{\leq} \frac{(H-h)\varepsilon}{2H^2} + \frac{(H+2-h)\varepsilon}{2H^2} = \frac{(H+1-h)\varepsilon}{H^2}.
\end{aligned}
$$

where (i) follows by bounding

$$
\begin{aligned}
&\breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \sum_{\ell=1}^{h-1} w_\ell(s_\ell,a_\ell) + w_h(s,a), \sum_{\ell=1}^{h-1} v_\ell(s_\ell,a_\ell) + v_h(s,a), \sum_{\ell=1}^{h-1} b_\ell(s_\ell,a_\ell) + b_h(s,a)\right) \\
&\qquad - \breve{V}_{h+1}^{\widehat{\theta}_{h+1:H}}\left(s', \nu_{i_h} + w_h(s,a), \nu_{j_h} + v_h(s,a), \nu_{k_h} + b_h(s,a)\right) \leq \frac{\varepsilon}{2H^2}
\end{aligned}
$$

using the same logic as we used above to arrive at inequality (62), and (ii) follows by using the Part (a) of the inductive hypothesis at step $h+1$. This proves the second part of the inductive hypothesis.

*Part (c):*Recall the definition of

$$
\begin{aligned}
&\widehat{a}_h(s,i,j,k) \in \arg\max_{a\in\mathcal{A}} \mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\left[\widehat{V}_{h+1}(s', \sigma(w_h(s,a)+\nu_i), \sigma(v_h(s,a)+\nu_j), \sigma(b_h(s,a)+\nu_k))\right], \\
&\widehat{V}_h(s,i,j,k) := \max_{a\in\mathcal{A}} \mathbb{E}_{s'\sim\bar{\mathbb{P}}(\cdot|s,a)}\left[\widehat{V}_{h+1}(s', \sigma(w_h(s,a)+\nu_i), \sigma(v_h(s,a)+\nu_j), \sigma(b_h(s,a)+\nu_k))\right].
\end{aligned}
$$

For a fixed quartet (s,i,j,k) to calculate $\widehat{a}_h(s,i,j,k)$ it is possible to first calculate

$$
\widehat{V}_{h+1}(s', \sigma(w_h(s,a)+\nu_i), \sigma(w_h(s,a)+\nu_j), \sigma(b_h(s,a)+\nu_k))
$$

for all $s' \in \mathcal{S}$ and all $a \in \mathcal{A}$. Since we already have access to the tensor the entire $\widehat{V}_{h+1}$ this takes $\mathsf{poly}(|\mathcal{S}||\mathcal{A}|)$ time and memory. Once we have calculated this it is possible to use this to calculate

$$\mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s,a)} \left[\widehat{V}_{h+1}(s', \sigma(w_h(s,a) + \nu_i), \sigma(b_h(s,a) + \nu_j)) \right]$$

for all choices of a (we can do this since we have access to the distribution $\bar{\mathbb{P}}$) using $\mathsf{poly}(|\mathcal{S}|, |\mathcal{A}|)$ time and memory. After we have enumerated this value for all $a \in \mathcal{A}$ we can identify $\widehat{a}_h(s,i,j,k)$ and $\widehat{V}_h(s,i,j,k)$ for this quartet. There are $|\mathcal{S}|m^3 = |\mathcal{S}| \left(\left\lceil \frac{12H^2\zeta}{\varepsilon} \right\rceil \right)^3$ quartets. Therefore it is possible to calculate both these tensors using $\mathsf{poly}\left(|\mathcal{S}|, |\mathcal{A}|, H, \zeta, \frac{1}{\varepsilon}\right)$ time and memory, which proves our claim.

This completes the proof of all parts of the induction hypothesis.

Part II: Using the induction hypothesis to prove the lemma. We begin by proving that the policy $\widehat{\theta}$ can be found efficiently. To see this, notice that at every step the policy $\widehat{\theta}$ only requires to know the tensor of actions $\widehat{a}_h$. Starting from $h = H$, we have shown that each $\widehat{a}_h$ can be computed using $\mathsf{poly}\left(|\mathcal{S}|, |\mathcal{A}|, H, \zeta, \frac{1}{\varepsilon}\right)$ time and memory. Thus, all H of these tensors can be found using $\mathsf{poly}\left(|\mathcal{S}|, |\mathcal{A}|, H, \zeta, \frac{1}{\varepsilon}\right)$ time and memory.

Now let's prove that

$$\bar{V}^{\theta_\star} - \bar{V}^{\widehat{\theta}} \leq \varepsilon.$$

Define a policy $\theta_{\star h} := \left(\theta_{\star 1}, \ldots, \theta_{\star h}, \widehat{\theta}_{h+1}, \widehat{\theta}_H\right)$ for $h \in \{0, \ldots, H\}$. Therefore,

$$\bar{V}^{\theta_\star} - \bar{V}^{\widehat{\theta}} = \sum_{h=H}^{1} \bar{V}^{\theta_{\star h}} - \bar{V}^{\theta_{\star h-1}}. \tag{66}$$

Consider any term in this decomposition above,

$$\begin{aligned}
&\bar{V}^{\theta_{\star h}} - \bar{V}^{\theta_{\star h-1}} \\
&= \mathbb{E}_{s_1 \sim \rho,\, \tau_{h-1} \sim \bar{\mathbb{P}}^{\theta_\star 1:h-1}} \left[\mathbb{E}_{s_h \sim \bar{\mathbb{P}}(\cdot|s_{h-1}, a_{h-1})} \left[\max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim \bar{\mathbb{P}}(\cdot|s_h, a)} \left[\bar{V}_{h+1}^{\widehat{\theta}_{h+1:H}}(s', \{s_h, a, \tau_{h-1}\}) \right] \right. \right. \\
&\qquad \left. \left. - \bar{V}_h^{\widehat{\theta}_{h:H}}(s_h, \tau_{h-1}) \right] \right]
\end{aligned}$$

where the outer expectation $\mathbb{E}_{\tau_{h-1} \sim \bar{\mathbb{P}}^{\theta_\star 1:h-1}}$ is over the randomness in the first $h-1$ round where the policy is $(\theta_{\star 1}, \ldots, \theta_{\star h-1})$ and the initial state is s_1. Now by invoking the second part of the induction hypothesis to bound the RHS in the display above we get

$$\bar{V}^{\theta_{\star h}} - \bar{V}^{\theta_{\star h-1}} \leq \frac{(H+1-h)\varepsilon}{H^2}.$$

Plugging this into equation (66) we conclude that

$$\bar{V}^{\theta_\star} - \bar{V}^{\widehat{\theta}} \leq \frac{\varepsilon}{2H^2} \sum_{h=1}^{H} (H+1-h) < \frac{\varepsilon}{2H^2} \sum_{h=1}^{H} (H+1) \leq \varepsilon$$

completing our proof. ■

E.2 Proof of Proposition 3.7

Recall the statement of the proposition from above.

Proposition 3.7. *For any $t \in [N]$ define $\widetilde{V}_t^{\mathsf{sd}}(\pi) := \mathbb{E}_{s_1 \sim \rho,\, \tau \sim \widehat{\mathbb{P}}_t^{\pi}(\cdot|s_1)} \left[\widetilde{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t, \tau) \right]$. Given any $\varepsilon > 0$, under Assumptions 2.2, 3.3 and 3.4 it is possible to find a policy $\widehat{\pi}^{(t)}$ that satisfies*

$$\widetilde{V}_t^{\mathsf{sd}}(\pi^{(t)}) - \widetilde{V}_t^{\mathsf{sd}}(\widehat{\pi}^{(t)}) \leq \varepsilon,$$

using at most $\mathsf{poly}\left(|\mathcal{S}|, |\mathcal{A}|, H, d, B, \|\widehat{\mathbf{w}}_t\|_2, \frac{1}{\varepsilon}, \log\left(\frac{N}{\delta}\right)\right)$ time and memory.

Proof The proof shall follow by simply invoking Lemma E.1. Recall from equation (D.5) that

$$\widetilde{\mu}_t^{\mathsf{sd}}(\widehat{\mathbf{w}}_t, \tau) := \min\left\{\mu\left(\mathbf{w}^\top \phi(\tau)\right) + \sqrt{\kappa}\beta_t(\delta)\sum_{h=1}^{H}\|\phi_h(s_h, a_h)\|_{\mathbf{\Sigma}_t^{-1}}, 1\right\} + \sum_{h=1}^{H-1}\xi^{(t)}_{s_h,a_h}.$$

First notice that since $\|\phi(\tau)\|_2 \leq 1$ we have that

$$|\widehat{\mathbf{w}}_t, \phi(\tau)| \leq \|\widehat{\mathbf{w}}_t\|_2\|\phi(\tau)\|_2 \leq \|\widehat{\mathbf{w}}_t\|_2. \tag{67}$$

Next observe that

$$\sqrt{\kappa}\beta_t(\delta)\sum_{h=1}^{H}\|\phi_h(s_h, a_h)\|_{\mathbf{\Sigma}_t^{-1}} \leq \sqrt{\kappa}\beta_t(\delta)\sqrt{\lambda_{\max}(\mathbf{\Sigma}_t^{-1})}\sum_{h=1}^{H}\|\phi_h(s_h, a_h)\|_2 \tag{68}$$

$$\overset{(i)}{\leq} \frac{\sqrt{\kappa}\beta_t(\delta)}{\sqrt{\lambda_{\min}(\mathbf{\Sigma}_t)}}\sqrt{H}\sqrt{\sum_{h=1}^{H}\|\phi_h(s_h, a_h)\|_2^2} \tag{69}$$

$$\overset{(ii)}{\leq} \frac{\sqrt{\kappa}\beta_t(\delta)}{\sqrt{\lambda_{\min}(\mathbf{\Sigma}_t)}}\sqrt{H}\|\phi(\tau)\|_2 \tag{70}$$

$$\overset{(iii)}{\leq} \frac{\sqrt{\kappa}\beta_t(\delta)}{\sqrt{\kappa}}\sqrt{H}\,\|\phi(\tau)\|_2 \tag{71}$$

$$\leq \sqrt{H}\beta_t(\delta) \overset{(iv)}{\leq} \sqrt{H} \times \mathsf{poly}\left(d, B, \log\left(\frac{N}{\delta}\right)\right), \tag{72}$$

where (i) follows since for any $z \in \mathbb{R}^H$, $\|z\|_1 \leq \sqrt{H}\|z\|_2$, (ii) follows since by Assumption 3.4 for any $h \neq h' \in [H]$, the features ϕ_h and $\phi_{h'}$ are orthogonal and by Assumption 3.3 the feature map ϕ is sum-decomposable, (iii) follows since $\mathbf{\Sigma}_t \succeq \kappa\mathbf{I}$, and (iv) follows by the definition of $\beta_t(\delta)$ in equation (4). Finally the definition of $\xi^{(t)}$ in equation (7) we know that

$$\left|\sum_{h=1}^{H-1}\xi^{(t)}_{s_h,a_h}\right| \leq 2H. \tag{73}$$

In light of inequalities (67), (72) and (73) we can conclude that if we invoke Lemma E.1 with a ζ that is a large enough polynomial in $\|\widehat{\mathbf{w}}_t\|_2, d, B, \log(N/\delta), H$ then the claim follows. ■

F Experiments

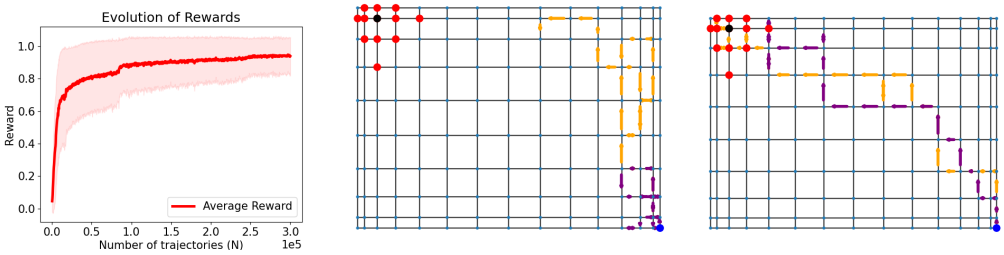

Figure 1. Left: Reward learning curve averaged over 40 independent runs. The shaded region represents a confidence interval which is ±standard deviation. **Middle:** The purple and yellow paths represent two sample paths taken by an initial random policy. **Right:** The purple and yellow paths represent two sample paths taken by a trained policy.

In this section we experimentally show that it is possible to learn a good policy in a simple non-Markovian domain with binary rewards—received once per episode—using a policy gradient

algorithm. We parameterize each policy π_θ by $\theta \in \mathbb{R}^k$. The gradients of the value function can be computed using the REINFORCE [36] algorithm as follows

$$\nabla_\theta V^{\pi_\theta} = \mathbb{E}_{y_\tau, \tau \sim \mathbb{P}^\pi} \left[y_\tau \left(\sum_{h=1}^{H} \nabla_\theta \log \left(\pi_\theta(a_h | s_h) \right) \right) \right].$$

We approximate this expectation empirically by using 30 sample trajectories, and use the Adam optimizer [21] with a default step size of one to update the policy. We studied the behavior of this algorithm on a custom 10×15 grid environment. The agent is initialized at a random location on the grid denoted by the large blue dot. Then the agent is allowed to take one of the actions $\{\mathrm{UP}, \mathrm{DOWN}, \mathrm{LEFT}, \mathrm{RIGHT}\}$, and move to an adjacent node (if permitted). During the last three steps of an episode, with $H = 30$, if the agent stays at either the black dot ('goal') or at any adjacent nodes marked by the red dots, then the agent receives a reward of 1, while if the agent is not at one of these nodes during the last three steps then it receives a reward of 0. The location of the 'goal' node is also randomly chosen at each episode. We parametrize the policy using a fully connected neural network with 10 hidden layers and with width 4. The state representation that is fed to this policy is of the form $(x^{\mathrm{current}}, y^{\mathrm{current}}, x^{\mathrm{goal}}, y^{\mathrm{goal}})$, where $(x^{\mathrm{current}}, y^{\mathrm{current}})$ represents the current coordinates of the agent and $(x^{\mathrm{goal}}, y^{\mathrm{goal}})$ denotes the coordinates of the 'goal' node.