On the Sample Complexity of Learning Infinite-horizon Discounted Linear Kernel MDPs

Yuanzhou Chen [*1] **Jiafan He** [*2] **Quanquan Gu** [2]

Abstract

We study reinforcement learning for infinite-horizon discounted linear kernel MDPs, where the transition probability function is linear in a predefined feature mapping. Existing UCLK (Zhou et al., 2021b) algorithm for this setting only has a regret guarantee, which cannot lead to a tight sample complexity bound. In this paper, we extend uniform-PAC sample complexity from the episodic setting to the infinite-horizon discounted setting, and propose a novel algorithm dubbed UPAC-UCLK that achieves an $\widetilde{O}\big(d^2/((1-\gamma)^4\epsilon^2) + 1/((1-\gamma)^6\epsilon^2)\big)$ uniform-PAC sample complexity, where d is the dimension of the feature mapping, $\gamma \in (0,1)$ is the discount factor of the MDP and ϵ is the accuracy parameter. To the best of our knowledge, this is the first $\widetilde{O}(1/\epsilon^2)$ sample complexity bound for learning infinite-horizon discounted MDPs with linear function approximation (without access to the generative model).

1. Introduction

In reinforcement learning (RL), the central goal is to design an efficient algorithm to learn the optimal (or near-optimal) policy through repeated interactions between the agent and an unknown environment. Markov decision processes (MDPs) are typical models that describe the environment formally. An MDP is described by a tuple of action space, state space, reward function and transition probability function. Based on the planning horizon and the transition probability kernel, MDPs can be categorized into three different types: (1) episodic MDPs (Azar et al., 2017; Dann et al., 2017; Jin et al., 2018; 2020; Ayoub et al., 2020; Zhou et al., 2021a; Zhang et al., 2020a; Zanette et al., 2020; He et al., 2021b), (2) infinite-horizon average reward MDPs (Jaksch et al., 2010; Bartlett & Tewari, 2012; Ouyang et al., 2017; Agrawal & Jia, 2017; Fruit et al., 2018b;a; Talebi & Maillard, 2018; Zhang & Ji, 2019; Fruit et al., 2020; Ortner, 2020; Wei et al., 2021; Wu et al., 2022), and (3) infinite-horizon discounted MDPs (Kakade et al., 2003; Strehl et al., 2006; Dong et al., 2019; Zhang et al., 2020b; Szita & Szepesvári, 2010; Lattimore & Hutter, 2012; Liu & Su, 2020; He et al., 2021a; Zhou et al., 2021a;b). The theoretical studies of infinite-horizon discounted MDPs are still quite limited, compared with the other two types of MDPs. Without the restart process in episodic MDPs, or the bounded diameter condition in infinite-horizon average reward MDPs (Jaksch et al., 2010), the infinite-horizon discounted MDP poses a great challenge for both algorithm design and theoretical analysis.

Existing results for discounted MDPs mainly focus on the tabular setting, where the agent can access each state-action pair (s,a) and estimate the corresponding value functions individually. Without the help of a *generative model* (Kakade et al., 2003) that allows the agent to visit all state-action pairs (s,a), Lattimore & Hutter (2012); Dong et al. (2019); Zhang et al. (2020b) and Liu & Su (2020); He et al. (2021a) showed that it is still possible to obtain a polynomial sample complexity or $O(\sqrt{T})$-regret for discounted tabular MDPs.

However, these tabular RL algorithms are computationally inefficient when the sizes of action and state spaces are large. A common approach to deal with high-dimensional (or even infinite) state space and action space is to use certain function classes, such as linear function class and neural networks, to approximate the transition probability kernel or the value function (Jin et al., 2020; Ayoub et al., 2020). For discounted MDPs, Zhou et al. (2021b) studied a linear kernel MDP, where the transition probability kernel can be represented by a linear function in some known feature mapping, and proposed the UCLK algorithm with an $O(\sqrt{T})$-regret guarantee. Unfortunately, unlike episodic MDPs initialized at the same state for different episodes, the regret guarantee cannot imply a finite sample complexity bound for discounted MDPs. Therefore, UCLK can fail to learn the ϵ-optimal policy, even though its regret is sublinear. Thus, a natural question arises:

[*]Equal contribution [1]School of Mathematical Sciences, Peking University, Beijing, China [2]Computer Science Department, University of California at Los Angeles, Los Angeles, California, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Can we design provably efficient algorithms with polynomial sample complexity for discounted MDPs with linear function approximation?

In this paper, we answer this question affirmatively by presenting a variant of the UCLK algorithm (Zhou et al., 2021b), namely UPAC-UCLK, and prove that this algorithm has a near-optimal sample complexity. In fact, our algorithm enjoys a stronger notion of sample complexity guarantee than the standard PAC sample complexity, which is called uniform-PAC sample complexity Dann et al. (2017). As discussed in Dann et al. (2017), the uniform-PAC guarantee is stronger than both PAC sample complexity and regret, and can further guarantee the convergence to the optimal policy up to an arbitrarily small error.

Our contributions are summarized as follows:

- We adapt the multi-level scheme from He et al. (2021b) to the UCLK algorithm (Zhou et al., 2021b), and propose a novel algorithm UPAC-UCLK for discounted linear kernel MDPs. Compared with the original multi-level scheme, we propose a novel *discounted data inheritance* technique, which adds a discounted portion of data at each level to the subsequent level. With this technique, newly added levels are non-empty at initialization due to data *inheritance*, and high levels contain low data information due to data *discounting*. This technique provides crucial guarantees for our algorithm, allowing us to uniformly bound the sample complexity.
- We show that our algorithm satisfies the uniform-PAC guarantee with sample complexity $\widetilde{O}\big(d^2/((1-\gamma)^4\epsilon^2) + 1/((1-\gamma)^6\epsilon^2)\big)$, where d is the dimension of the feature mapping, $\gamma \in (0,1)$ is the discount factor of the MDP and ϵ is the accuracy parameter. This result immediately implies a high probability regret bound $\widetilde{O}(d\sqrt{T}/(1-\gamma)^2 + \sqrt{T}/(1-\gamma)^3)$, where T is the number of interactions with the environment. The regret bound matches that of UCLK algorithm (Zhou et al., 2021b) up to a logarithmic factor, after ignoring the extra $\widetilde{O}\big(\sqrt{T}/(1-\gamma)^3\big)$ term in the regret. To the best of our knowledge, UPAC-UCLK is the first algorithm for infinite-horizon discounted RL with linear function approximation that enjoys a (uniform) sample complexity guarantee.

The remaining of this paper is organized as follows. Section 2 reviews the mostly related work. Section 3 provides the preliminaries of MDPs and reinforcement learning. Section 4 presents our main algorithms. Section 5 introduces the theoretical guarantees of our algorithm, and discusses its implications. Section 6 provides a proof outline for the main theorem, along with several key technical lemmas. Section 7 concludes this work with discussions on the future work. The detailed proofs are deferred to the appendix.

Notation In this work, we use lower case letters to denote scalars, and use lower and upper case bold face letters to denote vectors in $\mathbb{R}^d$ and $d \times d$ matrices respectively. For a vector $\mathbf{x} \in \mathbb{R}^d$, we denote by $\|\mathbf{x}\|_2$ the Euclidean norm. Furthermore, for a positive-definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, we define the Mahalanobis norm of $\mathbf{x}$ with respect to $\mathbf{\Sigma}$ to be $\|\mathbf{x}\|_{\mathbf{\Sigma}} = \sqrt{\mathbf{x}^\top \mathbf{\Sigma}\mathbf{x}}$. In addition, for a matrix $\mathbf{\Sigma}$ with real eigenvalues, we denote its largest and smallest eigenvalues by $\lambda_{\max}(\mathbf{\Sigma})$ and $\lambda_{\min}(\mathbf{\Sigma})$ respectively. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists an integer subscript n_0 and an absolute constant C, such that for all $n \geq n_0$, $a_n \leq Cb_n$. We use $\widetilde{O}(\cdot)$ to further hide logarithmic factors except for the $\log T$ term in all uniform-PAC guarantees.

2. Related Work

In this section, we will review the mostly related work to our work.

2.1. Infinite-horizon Discounted MDPs

For discounted tabular MDPs, there has been a series of work providing sample complexity guarantees without the help of a generative model (Kakade et al., 2003). For these results, the algorithms have access to every possible state-action pair. They can be mainly grouped into two categories: model-free algorithms and model-based algorithms. For model-free algorithms, the value function for each state-action pair is directly estimated. For instance, Strehl et al. (2006) first proposed the delayed Q-learning algorithm with polynomial sample complexity guarantee. Later, Dong et al. (2019) improved the sample complexity to $\widetilde{O}(SA/((1-\gamma)^7\epsilon^2))$ by proposing an infinite Q-learning with UCB algorithm, where S is the number of states and A is the number of actions. Recently, Zhang et al. (2020b) proposed the UCB-multistage-adv algorithm, which obtains a near-optimal sample complexity $\widetilde{O}(SA/((1-\gamma)^3\epsilon^2))$ that matches the theoretical lower bound. Model-based algorithms, on the other hand, estimate the underlining environment or the transition probability kernel and compute the value function from the estimated environment. For instance, Szita & Szepesvári (2010) introduced the MoRmax algorithm with $\widetilde{O}(SA/((1-\gamma)^6\epsilon^2))$ sample complexity. Later, Lattimore & Hutter (2012) proposed the UCRL-γ algorithm and improved the sample complexity to $\widetilde{O}(SA/((1-\gamma)^3\epsilon^2))$, under a strong assumption on the transition probability kernel. Our work also falls into the category of model-based algorithms, and considers the sample complexity as the performance measure.

Recently, Liu & Su (2020) extended the definition of re-

gret from episodic MDPs to the discounted MDPs, and named it γ-regret. Under this notion of regret, Liu & Su (2020) proposed the Double Q-Learning algorithm with an $\widetilde{O}(\sqrt{SAT}/(1-\gamma)^{2.5})$ regret guarantee for the first T steps. Later, He et al. (2021a) introduced the UCBVI-γ algorithm and improved the regret bound to $\widetilde{O}(\sqrt{SAT}/(1-\gamma)^{1.5})$. Additionally, He et al. (2021a) proved there exists a class of hard-to-learn discounted MDPs such that the regret for any algorithm is lower bounded by $\widetilde{\Omega}(\sqrt{SAT}/(1-\gamma)^{1.5})$. This matches the regret upper bound of UCBVI-γ algorithm up to a logarithmic factor.

In an attempt to design algorithms that can learn efficiently when the state and action spaces are high-dimensional or even infinite, Zhou et al. (2021b) focused on discounted MDPs with linear function approximation, and proposed the UCLK algorithm with a $\widetilde{O}\big(d\sqrt{T}/(1-\gamma)^2\big)$ regret guarantee, where d is the dimension of feature mapping. Later, Zhou et al. (2021a) introduced the Bernstein-type bonus and improved the regret bound to $\widetilde{O}\big(d\sqrt{T}/(1-\gamma)^{1.5}\big)$ in their new UCLK$^+$ algorithm, which matches the regret lower bound proved in Zhou et al. (2021b) up to a logarithmic factor. However, both UCLK and UCLK$^+$ only have regret guarantees, and cannot provide any sample complexity guarantees. To the best of our knowledge, our UPAC-UCLK algorithm gives the first sample complexity guarantee for learning discounted MDPs with linear function approximation.

2.2. Uniform-PAC Guarantees in RL

Traditional analysis on reinforcement learning mainly focuses on the regret in the first T steps or the sample complexity with respect to a specific accuracy parameter ϵ. Unfortunately, both of these performance measures fail to guarantee the convergence to optimal policy. To overcome this problem, Dann et al. (2017) first introduced a stronger performance guarantee named uniform-PAC for episodic MDPs. Different from traditional sample complexity guarantee, uniform-PAC requires bounding the sample complexity at all accuracy parameters ϵ simultaneously, and thus guarantees that the number of steps with suboptimality larger than ϵ is finite. Under this stronger performance measure, Dann et al. (2017) proposed the UBEV algorithm and obtained an $\widetilde{O}(SAH^4/\epsilon^2)$ uniform-PAC guarantee, which implies that the UBEV algorithm converges to the optimal policy. Here, H is the length of each episode. Later, He et al. (2021b) focused on episodic MDPs with linear function approximation and proposed the FLUTE algorithm. By introducing a minimax value function estimator and a multi-level scheme, the FLUTE algorithm obtains an $\widetilde{O}(d^3H^5/\epsilon^2)$ uniform-PAC guarantee for episodic linear MDPs (Jin et al., 2020). Our algorithm UPAC-UCLK also satisfies the stronger uniform-PAC guarantee, under the setting of discounted MDPs with linear function approximation.

3. Preliminaries

In this work, we consider the infinite-horizon discounted Markov decision processes (MDPs), denoted by $M(\mathcal{S}, \mathcal{A}, \gamma, r, \mathbb{P})$. In this tuple, $\mathcal{S}, \mathcal{A}$ are the spaces of states and actions respectively; $\gamma \in (0,1)$ is the discount factor; $r(\cdot,\cdot) : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the reward function, which we assume to be deterministic and known to the agent; $\mathbb{P}(\cdot|\cdot,\cdot) : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}^+$ is the transition probability function that satisfies the equations $\int_{s'\in\mathcal{S}} \mathbb{P}(s'|s,a)\mathrm{d}\mu(s') = 1$ for arbitrary $s \in \mathcal{S}, a \in \mathcal{A}$; here $\mathrm{d}\mu$ is a probability measure on $\mathcal{S}$. We will abbreviate $\mathrm{d}\mu(s)$ as $\mathrm{d}s$ henceforth.

As an agent tries to learn an MDP via interaction with the environment, it repeats the following process: it takes an action $a \in \mathcal{A}$ based on the current state s, then receives the next state $s' \in \mathcal{S}$ with the reward r. We can thus label these states and actions in chronological order with time step subscripts $t = 1, 2, \ldots$. For the agent, its choice of an action a_t is based only on the available observations before the action, namely a tuple $(s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$. Consequently, we can define the agent's (non-stationary) policy π on an MDP $M(\mathcal{S}, \mathcal{A}, \gamma, r, \mathbb{P})$ to be $\pi = \{\pi_t\}_{t=1}^{\infty}$, where $\pi_t : \{\mathcal{S}\times\mathcal{A}\}^{t-1}\times\mathcal{S} \to \mathcal{A}$ maps the tuple $(s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$ to an action a_t. The agent's policy π, combined with the transition probability $\mathbb{P}$, give rise to the infinite-length random process consisting of states and actions $\{s_t, a_t\}_{t=1}^{\infty}$.

Now consider the expected total discounted reward of a given policy π after a certain point in the learning process, a.k.a., value functions. The action-value function and the value function for policy π at time step t are defined as

$$Q_t^\pi(s,a) := \mathbb{E}\left[\sum_{\tau=t}^{\infty}\gamma^{\tau-t}r(s_\tau,a_\tau)\middle| s_1,\ldots,s_t=s,a_t=a\right],$$
$$V_t^\pi(s) := \mathbb{E}\left[\sum_{\tau=t}^{\infty}\gamma^{\tau-t}r(s_\tau,a_\tau)\middle| s_1,\ldots,s_t=s\right],$$

where expectation is conditioned on the states and actions up to time step t. With these in place, we can define the optimal action-value function $Q^*(s,a) := \sup_\pi Q_1^\pi(s,a)$ and optimal value function $V^*(s) := \sup_\pi V_1^\pi(s)$. We refer to the difference between the optimal value function and the value function of policy π at time step t, $\Delta_t = V^*(s_t) - V_t^\pi(s_t)$, as the *suboptimality gap* . Now for any real function defined on the state space $V : \mathcal{S} \to \mathbb{R}$, we define $\mathbb{P}V(s,a) := \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}V(s')$. With this notation, we can now deduce and express the Bellman equations in discounted MDPs as follows:

$$Q_t^\pi(s_t,a_t) = r(s_t,a_t) + \gamma\mathbb{P}V_{t+1}^\pi(s_t,a_t) \tag{3.1}$$
$$Q^*(s_t,a_t) = r(s_t,a_t) + \gamma\mathbb{P}V^*(s_t,a_t). \tag{3.2}$$

We focus on linear kernel MDPs (Zhou et al., 2021b) in this

work, which is also known as linear mixture models (Modi et al., 2019) or linear mixture MDPs (Ayoub et al., 2020).

Definition 3.1 (Linear Kernel MDP (Zhou et al., 2021b))**.** An MDP $M(\mathcal{S}, \mathcal{A}, \gamma, r, \mathbb{P})$ is called a linear kernel MDP if there exists a *known* feature mapping $\boldsymbol{\phi}(\cdot|\cdot,\cdot) : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}^d$ and an *unknown* vector $\boldsymbol{\theta} \in \mathbb{R}^d$, $\|\boldsymbol{\theta}\|_2 \le \sqrt{d}$, such that:

1. For all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, the transition probability function can be expressed as $\mathbb{P}(s'|s, a) = \langle\boldsymbol{\phi}(s'|s, a), \boldsymbol{\theta}\rangle$;
2. For any bounded function $V : \mathcal{S} \to [0, R]$, $\|\boldsymbol{\phi}_V(s, a)\|_2 \le \sqrt{d}R$ holds for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $\boldsymbol{\phi}_V(s, a) := \int_{s'\in\mathcal{S}} \boldsymbol{\phi}(s'|s, a)V(s')\mathrm{d}s'$.

We denote such an MDP by $M(\boldsymbol{\theta}; \boldsymbol{\phi})$ for simplicity.

From Definition 3.1, the following equation holds for any function $V : \mathcal{S} \to [0, R]$:

$$\langle\boldsymbol{\theta}^*, \boldsymbol{\phi}_V(s, a)\rangle = \mathbb{P}V(s, a). \tag{3.3}$$

Now we define the set of all feasible $\boldsymbol{\theta}$'s satisfying the conditions in Definition 3.1 to be

$$\mathcal{B} := \Big\{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \le \sqrt{d} \text{ and } \langle\boldsymbol{\phi}(\cdot|s, a), \boldsymbol{\theta}\rangle \text{ is a probability measure on } \mathcal{S}\Big\}, \tag{3.4}$$

which, as we will see in Lemma A.3 in the appendix, is a convex body when constrained to an affine subspace $\mathcal{L}$ of $\mathbb{R}^d$.

For linear kernel MDPs, Zhou et al. (2021b) proposed the Upper-Confidence Linear Kernel (UCLK) algorithm that achieves near-optimal regret. However, as pointed out by Dann et al. (2017), algorithms with regret bounds do not generally converge to the optimal policy: as long as the regret tends to infinity as $t \to \infty$, high suboptimal gaps may occur infinitely often. To overcome this issue, Dann et al. (2017) proposed a stronger notion of performance measure called uniform-PAC that guarantees the convergence to the optimal policy with high probability. We now extend it to the discounted MDP setting.

Definition 3.2 (Uniform-PAC)**.** For an algorithm $\mathbf{Alg}$ on an MDP, denote its corresponding policy by π. For any $\epsilon > 0$, define $N_\epsilon = \sum_{t=1}^{\infty} \mathbb{1}\{V^*(s_t) - V_t^\pi(s_t) > \epsilon\}$ to be the total number of time steps where the suboptimality gap is greater than ϵ. Algorithm $\mathbf{Alg}$ is said to be uniform-PAC for some $\delta \in (0, 1)$ with sample complexity $\Gamma\big(1/\epsilon, \log(1/\delta)\big)$ if

$$\mathbb{P}\big[\exists\epsilon > 0, N_\epsilon > \Gamma\big(1/\epsilon, \log(1/\delta)\big)\big] \le \delta, \forall\delta \in (0, 1),$$

where $\Gamma(\cdot,\cdot)$ is a polynomial function dependent on the MDP itself.

Remark 3.3. Definition 3.2 is slightly different from that in Dann et al. (2017); He et al. (2021b) due to the different definitions of N_ϵ. Specifically, Dann et al. (2017); He et al. (2021b) focused on episodic MDPs, where the suboptimality gap $\Delta_k = V_1^*(s_1^k) - V_1^\pi(s_1^k)$, and N_ϵ is the number of episodes with suboptimality gap greater than ϵ. Due to the difference between episodic MDPs and discounted MDPs, we define the suboptimality for each time step t as $\Delta_t = V^*(s_t) - V_t^\pi(s_t)$ and define the sample complexity as $N_\epsilon = \sum_{t=1}^{\infty} \mathbb{1}\{V^*(s_t) - V_t^\pi(s_t) > \epsilon\}$. In addition, for linear kernel MDPs, the sample complexity bound should be expressed in the form $\Gamma(1/\epsilon, \log(1/\delta); \gamma, d)$, where γ, d are the parameters of $M(\boldsymbol{\theta}; \boldsymbol{\phi})$, rather than $\Gamma(1/\epsilon, \log(1/\delta); \gamma, |\mathcal{S}|, |\mathcal{A}|)$ for tabular MDPs. One would expect Γ to be polynomial in $1/(1-\gamma)$ and d as well. Though in Dann et al. (2017), uniform-PAC was defined in the finite-horizon setting with finite state and action spaces, the proof of its basic properties does not depend on the specific MDP model. Hence, we still have the conclusions that uniform-PAC sample complexity bound implies both the PAC sample complexity bound and the regret bound, and it guarantees the convergence to the optimal policy with high probability, as stated in Theorem 3 of Dann et al. (2017).

From now on, we assume the true $\boldsymbol{\theta}$ in the target linear kernel MDP $M(\boldsymbol{\theta}; \boldsymbol{\phi})$ to be $\boldsymbol{\theta}^*$.

4. Proposed Algorithms

In this section, we propose a new algorithm called UPAC-UCLK for learning infinite-horizon discounted linear kernel MDPs. Our algorithm UPAC-UCLK is inspired by the UCLK algorithm (Zhou et al., 2021b). To achieve the uniform-PAC guarantee, we incorporate the multi-level partition scheme proposed in He et al. (2021b) and separate all time steps and their corresponding data into different levels. The algorithm is described in Algorithm 1. It is worth noting that the direct application of multi-level scheme to the infinite-horizon discounted MDPs requires maintaining infinite number of levels. However, this is infeasible in practice. In order to overcome this problem, we will maintain L "effective" levels along with extra $3L + C_0$ auxiliary levels in Algorithm 1, which can be seen as a finite-level approximation to the infinite number of levels.

4.1. Planning (Line 4 to 7 of Algorithm 1)

At the beginning of time step t, our algorithm calculates the estimates of $\boldsymbol{\theta}^*$ for all available levels $l = 1, 2, \ldots, 4L+C_0$. By (3.3), we have $\langle\boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_t}(s_t, a_t)\rangle = \mathbb{P}V_t(s_t, a_t) = \mathbb{E}\big[V_t(s_{t+1})\big|s_t, a_t\big]$, so a natural method to estimate $\boldsymbol{\theta}^*$ is using linear regression. In particular, we calculate the estimates by the closed-form solution of linear regression

Algorithm 1 Uniform-PAC UCLK (UPAC-UCLK)

Require: Regularization parameter λ, exploration parameters β^l for $l = 1, 2, \ldots$, number of value iteration rounds U_t for $t = 1, 2, \ldots$.

1: **Initialize** $C_0 \leftarrow 3\lceil \log\big(\sqrt{d}/(1-\gamma)\big)/\log 2\rceil + 6$, $L \leftarrow 0$, $\mathbf{\Sigma}^l \leftarrow \lambda\mathbf{I}$, $\mathbf{b}^l \leftarrow \mathbf{0}$, $l = 1, 2, \ldots$.
2: Receive initial state s_1.
3: **for** time step $t = 1, 2, \ldots$ **do**
4: **for** all level $l \in \{1, 2, \ldots, 4L + C_0\}$ **do**
5: $\widehat{\boldsymbol{\theta}}^l \leftarrow (\mathbf{\Sigma}^l)^{-1}\mathbf{b}^l$, $\mathcal{C}^l \leftarrow \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^l\|_{\mathbf{\Sigma}^l} \le \beta^l\}$.
6: **end for**
7: $\{Q_t^l(\cdot,\cdot)\}_{l=1}^{4L+C_0}, V_t(\cdot) \leftarrow$ ML-EVI$(\{\mathcal{C}^l\}_{l=1}^{4L+C_0}, U_t)$.
8: $a_t \leftarrow \operatorname{argmax}_a \min_{1\le l\le L} Q_t^l(s_t, a)$,
9: receive $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$.
10: $l_t \leftarrow 1$.
11: **while** $l_t \le L$ **and** $\|\boldsymbol{\phi}_{V_t}(s_t, a_t)\|_{(\mathbf{\Sigma}^{l_t})^{-1}} \le 2^{-l_t}\sqrt{d}/(1-\gamma)$ **do**
12: $l_t \leftarrow l_t + 1$
13: **end while**
14: **if** $l_t = L + 1$ **then**
15: **for** level $l = 4L + C_0 + 1, \ldots, 4L + C_0 + 4$ **do**
16: $\mathbf{\Sigma}^l \leftarrow \lambda\mathbf{I} + \frac{1}{2}(\mathbf{\Sigma}^{l-1} - \lambda\mathbf{I})$, $\mathbf{b}^l \leftarrow \frac{1}{2}\mathbf{b}^{l-1}$.
17: **end for**
18: $L \leftarrow L + 1$.
19: **end if**
20: **for** level $l = l_t, \ldots, 4L + C_0$ **do**
21: $\mathbf{\Sigma}^l \leftarrow \mathbf{\Sigma}^l + 2^{l_t - l}\boldsymbol{\phi}_{V_t}(s_t, a_t)\boldsymbol{\phi}_{V_t}(s_t, a_t)^\top$,
22: $\mathbf{b}^l \leftarrow \mathbf{b}^l + 2^{l_t - l}\boldsymbol{\phi}_{V_t}(s_t, a_t)V_t(s_{t+1})$.
23: **end for**
24: **end for**

Algorithm 2 Multi-Level Extended Value Iteration (ML-EVI)

Require: Number of levels L, confidence sets $\mathcal{C}_l$, $l = 1, \ldots, L$, number of value iteration rounds U.

$Q_l^{(0)}(\cdot,\cdot) \leftarrow 1/(1-\gamma)$ for $l = 1, \ldots, L$.
for $u = 1, \ldots, U$ **do**
  $V^{(u-1)}(\cdot) \leftarrow \max_{a\in\mathcal{A}} \min_{1\le l\le L} Q_l^{(u-1)}(\cdot, a)$,
  for $l = 1, \ldots, L$ **do**
    if $\mathcal{B} \cap \mathcal{C}_l \ne \emptyset$ **then**
      $Q_l^{(u)}(\cdot,\cdot) \leftarrow r(\cdot,\cdot) + \gamma \max_{\boldsymbol{\theta}\in\mathcal{B}\cap\mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(u-1)}}(\cdot,\cdot)\rangle$.
    else
      $Q_l^{(u)}(\cdot,\cdot) \leftarrow 1/(1-\gamma)$.
    end if
  end for
end for
$V^{(U)}(\cdot) \leftarrow \max_a \min_{1\le l\le L} Q_l^{(U)}(\cdot, a)$.
return $\{Q_l^{(U)}(\cdot,\cdot)\}_{l=1}^L, V^{(U)}(\cdot)$.

$\widehat{\boldsymbol{\theta}}^l = \big(\mathbf{\Sigma}^l\big)^{-1}\mathbf{b}^l$ for each level, where $\mathbf{\Sigma}^l$ and $\mathbf{b}^l$ are updated in Lines 16, 21 and 22. In Line 5, we then construct the confidence sets $\mathcal{C}^l$ of $\boldsymbol{\theta}^*$ for all levels l, which are centered at $\widehat{\boldsymbol{\theta}}^l$ with radius β^l. The Multi-Level Extended Value Iteration Algorithm (ML-EVI), as described in the next subsection, then calculates the estimates of value functions $\{Q_t^l\}$ and V_t.

4.2. Multi-Level Extended Value Iteration (Algorithm 2)

The ML-EVI algorithm is described in Algorithm 2. It can be seen as an extension of the Extended Value Iteration algorithm (EVI, Algorithm 2 from Zhou et al. (2021b)). By alternating between performing the greedy policy (Line 3) and an approximate Bellman equation (Line 6), it finds an optimistic estimate for the optimal value functions $Q^*(\cdot,\cdot)$ and $V^*(\cdot)$.

Discounted Data Inheritance. In He et al. (2021b), each data point (i.e., a pair of regression predictor and response) is only added to one single level. In contrast, in our setting, the data in one level are inherited by the subsequent levels after being discounted. Under this scheme, our estimate of $\boldsymbol{\theta}^*$ at level l at time step t regresses over the set of data

$$\big\{2^{(l_\tau - l)/2}\big(\boldsymbol{\phi}_{V_\tau}(s_\tau, a_\tau), V_\tau(s_{\tau+1})\big) : \tau \le t, l_\tau \le l\big\},$$

which has a closed-form solution $\widehat{\boldsymbol{\theta}}^l = \big(\mathbf{\Sigma}^l\big)^{-1}\mathbf{b}^l$ with $\mathbf{\Sigma}^l$ and $\mathbf{b}^l$ being defined as follows:

$$\mathbf{\Sigma}^l = \lambda\mathbf{I} + \sum_{\tau=1}^{t} \mathbb{1}\{l_\tau \le l\} 2^{l_\tau - l}\boldsymbol{\phi}_{V_\tau}(s_\tau, a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau, a_\tau)^\top,$$

$$\mathbf{b}^l = \sum_{\tau=1}^{t} \mathbb{1}\{l_\tau \le l\} 2^{l_\tau - l}\boldsymbol{\phi}_{V_\tau}(s_\tau, a_\tau)V_\tau(s_{\tau+1}).$$

We now explain the reasons behind this algorithm design.

As we will see in (6.5) of Section 6, in order to bound the sample complexity, we need to control the term $(V_t - V_{t+1})(s_{t+1})$: the difference on s_{t+1} between two consecutive optimistic value functions output by Algorithm 2. When the total number of levels L is increased by 1 in Line 18 of Algorithm 1, we say a new level is activated. Without data inheritance, newly activated levels hold a very scarce amount of data and have large exploration radii. This implies that the corresponding confidence sets are large. If these levels were selected in the minimization operation of Line 3 or Line 12 in Algorithm 2 at a later iteration u, it may produce an inaccurate estimation of the value function. In contrast, with data inheritance, newly activated levels will have enough data to produce sufficiently small confidence sets and this problem can be alleviated.

In addition, we need to discount the data when inheriting data from lower levels. The main reason is that for those

levels much higher than L, they still have large exploration radii, so adding the undiscounted new data can result in big changes to their confidence sets. With data discounting, the addition of data to levels much higher than L is exponentially small and therefore negligible.

With the help of discounted data inheritance, the term $(V_t - V_{t+1})(s_{t+1})$ can be roughly bounded by $\widetilde{O}\big(\sqrt{\boldsymbol{\phi}_{V_t}(s_t,a_t)^\top(\boldsymbol{\Sigma}^{l_t})^{-1}\boldsymbol{\phi}_{V_t}(s_t,a_t)}\big) = O(2^{-l_t})$ and will diminish as $t \to \infty$.

4.3. Execution (Line 8 to 9 in Algorithm 1)

The algorithm estimates the action-value function Q^* by taking a minimum on the upper bounds Q_t^l, and chooses the next action based on this estimate. The next state is then generated by the environment based on the underlying transition probability function $\mathbb{P}$.

4.4. Level Selection (Line 10 to 23 in Algorithm 1)

Since our new predictor at time step t is $\boldsymbol{\phi}_{V_t}(s_t,a_t)$, its uncertainty at level l is characterized by $\|\boldsymbol{\phi}_{V_t}(s_t,a_t)\|_{(\boldsymbol{\Sigma}^l)^{-1}}$. Based on the uncertainty, our algorithm finds the minimum level l at which the uncertainty exceeds the threshold value at level l, i.e.,

$$l_t = \min_{l \le L} \big\{ \|\boldsymbol{\phi}_{V_t}(s_t,a_t)\|_{(\boldsymbol{\Sigma}^l)^{-1}} > 2^{-l}\sqrt{d}/(1-\gamma) \big\}.$$

If this minimum value does not exist, or in other words the uncertainty is less than the thresholds at all effective levels $l \le L$, l_t is set to $L+1$, $L \leftarrow L+1$, and a new effective level is activated along with 3 other auxiliary levels. Based on the discounted data inheritance scheme, the initialization of these new levels has the following form

$$\boldsymbol{\Sigma}^l \leftarrow \lambda\mathbf{I} + \frac{1}{2}(\boldsymbol{\Sigma}^{l-1} - \lambda\mathbf{I}), \mathbf{b}^l \leftarrow \frac{1}{2}\mathbf{b}^{l-1}.$$

Finally, Algorithm 1 updates the variables $\boldsymbol{\Sigma}, \mathbf{b}$ for all levels $l \ge l_t$ by adding the new data using the discounted data inheritance scheme (Line 20 to 23).

5. Main Results

In this section, we present the theoretical guarantees for our UPAC-UCLK algorithm. Our main result is a theorem which certifies that our algorithm UPAC-UCLK has a uniform-PAC guarantee with an efficient sample complexity bound.

Theorem 5.1. For any linear mixture MDP, if we set $\lambda \ge 1$ and the parameters β^l and U_t in Algorithm 1 as follows:

$$U_t := \lceil \log\big(t(t+1)\big)/(1-\gamma) \rceil,$$
$$\beta^l := \frac{2}{1-\gamma}\big(3\sqrt{d\max\{l,l_0\}} + 2\sqrt{\log(1/\delta)}\big) + 2\sqrt{d\lambda}, \tag{5.1}$$

where $l_0 := \log\big(\sqrt{d}/(1-\gamma)\big)/\log 2$, then with probability at least $1-\delta$, for all accuracy parameter $\epsilon > 0$, the number of time steps with suboptimality gap larger than ϵ in Algorithm 1 is upper bounded by

$$\Gamma\big(1/\epsilon, \log(1/\delta); \gamma, d\big) = \widetilde{O}\bigg(\frac{1}{(1-\gamma)^6\epsilon^2} + \frac{d^2 + d\log(1/\delta)}{(1-\gamma)^4\epsilon^2}\bigg),$$

where d is the dimension of feature mapping and γ is the discount factor.

Remark 5.2. Similar to the main results in Zhou et al. (2021b), our sample complexity bound does not depend on the size of state space $\mathcal{S}$ or that of action space $\mathcal{A}$. This suggests that our algorithm is applicable to MDPs with large state and action spaces.

Corollary 5.3. Under the same conditions in Theorem 5.1, with probability at least $1-\delta$, the regret of Algorithm 1 is bounded as follows

$$\text{Regret}(T) = \widetilde{O}\bigg(\frac{d + \sqrt{d\log(1/\delta)}}{(1-\gamma)^2}\sqrt{T} + \frac{\sqrt{T}}{(1-\gamma)^3}\bigg).$$

Remark 5.4. Corollary 5.3 suggests that UPAC-UCLK enjoys an $\widetilde{O}(d\sqrt{T}/(1-\gamma)^2 + \sqrt{T}/(1-\gamma)^3)$ regret. In comparison, the UCLK algorithm in Zhou et al. (2021b) enjoys an $\widetilde{O}(d\sqrt{T}/(1-\gamma)^2)$ regret. Our algorithm suffers an extra $\widetilde{O}(\sqrt{T}/(1-\gamma)^3)$ term in the regret. This originates from a bound on the difference between ML-EVI outputs for two consecutive time steps, and we leave it as a future work to remove this term.

6. Proof Sketch of the Main Results

In this section, we will present the proof sketch of the main results. The detailed proof can be found in the appendix.

We first present several central lemmas, and then give a short sketch outlining the proof of Theorem 5.1. For the ease of presentation, we introduce new notations for the variables in Algorithm 1: denote the variables L, $\boldsymbol{\Sigma}^l$, $\mathbf{b}^l$, $\widehat{\boldsymbol{\theta}}^l$ and $\mathcal{C}^l$ after Line 6 of time step t by L_t, $\boldsymbol{\Sigma}_t^l$, $\mathbf{b}_t^l$, $\widehat{\boldsymbol{\theta}}_t^l$ and $\mathcal{C}_t^l$.

For simplicity, we define $k_t^l := \sum_{\tau=1}^{t-1} \mathbb{1}\{l_\tau \le l\} 2^{l_\tau - l}$ for $t \ge 1$ and $l \ge 0$, which represents the total amount of data in level l at time step t. With this notation, we have the following lemma upper bounding this value:

Lemma 6.1. Let $l_0 := \log\big(\sqrt{d}/(1-\gamma)\big)/\log 2$. Consider $k^l_{\max} := \lim_{t\to\infty} k^l_t$. For each level $1 \le l \le l_0$, we have $k^l_{\max} \le 20dl_0$, while for $l > l_0$, $k^l_{\max} \le 20dl4^{l-l_0}$.

Lemma 6.1 is parallel to Lemma A.1 from He et al. (2021b). It suggests that the amount of data in level l is finite with an upper bound exponential in l. Note that when l reaches beyond l_0, the threshold value at level l namely $2^{-l_0}\sqrt{d}/(1-\gamma)$ shrinks to below constant 1, which is considered the beginning of "useful" levels in our analysis.

Lemma 6.2. For the factor β^l defined in (5.1), with probability at least $1-\delta/2$, for all steps t and levels l, we have $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}^l_t\|_{\boldsymbol{\Sigma}^l_t} \le \beta^l/2$.

Lemma 6.2 suggests that with the given factor β^l, the true vector $\boldsymbol{\theta}^*$ lies within halved confidence sets in each level and time step. Here the radii of our confidence sets are set to be twice as large as normally defined. This is to ensure that $\boldsymbol{\theta}^*$ stays relatively close to the center of the confidence sets, and will enable our later bounds concerning the difference between confidence sets. For simplicity, we denote the event in Lemma 6.2 as $\mathcal{E}_1$.

Lemma 6.3. Under $\mathcal{E}_1$, for all t, l and $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned} \frac{1}{1-\gamma} &\ge Q^l_t(s,a) \ge Q^*(s,a), \\ \frac{1}{1-\gamma} &\ge V_t(s) \ge V^*(s). \end{aligned} \tag{6.1}$$

Lemma 6.3 suggests that the results of Algorithm 2 are upper bounds of the optimal value functions Q^* and V^*.

Lemma 6.4. Under $\mathcal{E}_1$, for all t and l, there exists a certain $\boldsymbol{\theta}^l_t \in \mathcal{C}^l_t \cap \mathcal{B}$, so that

$$Q^l_t(s_t,a_t) \le r(s_t,a_t) + \gamma\langle \boldsymbol{\theta}^l_t, \boldsymbol{\phi}_{V_t}(s_t,a_t)\rangle + \gamma^{U_t}. \tag{6.2}$$

Lemma 6.4 characterizes the transition error of Algorithm 2. Inequality (6.2) is an approximate version of the Bellman equation (3.2), and suggests that in order to reach an error of $O(1/\epsilon)$, a number of iteration rounds of order $O(\log(1/\epsilon))$ is needed.

We are now ready to begin our proof. Our analysis of sample complexity follows a similar road map as in Lemma 1 of Dong et al. (2019). We take the summation of large suboptimality over all time steps, which is trivially lower bounded, and then upper bound the sum with similar methods as in the proof of Theorem 1 of Zhou et al. (2021b). This yields an inequality centered around the sample complexity, which leads to the eventual conclusion.

Proof sketch of Theorem 5.1. In order to bound the sample complexity $N_\epsilon = \sum_{t=1}^\infty \mathbb{1}\{V^*(s_t) - V^\pi_t(s_t) > \epsilon\}$, consider the set

$$T_\epsilon = \{t : V^*(s_t) - V^\pi_t(s_t) > \epsilon\},$$

we instantly have

$$\epsilon N_\epsilon < \sum_{t\in T_\epsilon} \big(V^*(s_t) - V^\pi_t(s_t)\big). \tag{6.3}$$

We then bound the right hand side of (6.3). By (6.1) from Lemma 6.3, the suboptimality gap $V^*(s_t) - V^\pi_t(s_t)$ is upper bounded by $V_t(s_t) - V^\pi_t(s_t)$. Using (6.2) from Lemma 6.4 and the policy Bellman equation (3.1), we can decompose the suboptimality and eventually obtain

$$\begin{aligned} (V_t - V^\pi_t)(s_t) \le \Gamma(t) + \gamma\Delta(t) + \gamma(V_t - V_{t+1})(s_{t+1}) \\ + \gamma(V_{t+1} - V^\pi_{t+1})(s_{t+1}), \end{aligned} \tag{6.4}$$

where

$$\begin{aligned} \Gamma(t) &= 1/\big(t(t+1)\big) + 3\cdot 2^{l_0-l_t}\gamma\beta^{l_t-1}, \\ \Delta(t) &= \mathbb{P}(V_t - V^\pi_{t+1})(s_t,a_t) - (V_t - V^\pi_{t+1})(s_{t+1}). \end{aligned}$$

By iterating (6.4) for T rounds, we can bound the suboptimality gap by

$$(V^* - V^\pi_t)(s_t) \le \frac{\gamma^T}{1-\gamma} + A_t + B_t + C_t, \tag{6.5}$$

where

$$\begin{aligned} A_t &= \sum_{\tau=t}^{t+T-1} \gamma^{\tau-t}\Gamma(\tau), \\ B_t &= \sum_{\tau=t}^{t+T-1} \gamma^{\tau-t+1}\Delta(\tau), \\ C_t &= \sum_{\tau=t}^{t+T-1} \gamma^{\tau-t+1}(V_\tau - V_{\tau+1})(s_{\tau+1}). \end{aligned}$$

However, when the level l_t is small, the bound above is too large; in fact, it could go much larger than the trivial bound $1/(1-\gamma)$. We thus define the condition $\mathcal{Q}_t = \{l_\tau > l_0, \forall t \le \tau < t+T\}$, and separate the right hand side of (6.3) into two parts accordingly:

$$\begin{aligned} &\sum_{t\in T_\epsilon} \big(V^*(s_t) - V^\pi_t(s_t)\big) \\ &\le \sum_{t\in\mathcal{T}_\epsilon} \mathbb{1}\{\overline{\mathcal{Q}_t}\}\frac{1}{1-\gamma} \\ &\quad + \sum_{t\in\mathcal{T}_\epsilon} \mathbb{1}\{\mathcal{Q}_t\}\left(\frac{\gamma^T}{1-\gamma} + A_t + B_t + C_t\right). \end{aligned}$$

Since $\mathcal{Q}_t$ is false only for a small amount of t, the first term on the right hand side here is easy to bound. We then seek to bound the sums $\sum_{t\in T_\epsilon} \mathbb{1}\{\mathcal{Q}_t\}A_t$, $\sum_{t\in T_\epsilon} \mathbb{1}\{\mathcal{Q}_t\}B_t$ and $\sum_{t\in T_\epsilon} \mathbb{1}\{\mathcal{Q}_t\}C_t$ separately. For the sum of A_t, the individual terms are already explicit expressions. For B_t, we view $\Delta(t)$ as a martingale sequence and use the Azuma-Hoeffding inequality for a bound that holds with probability at least $1-\delta/2$. Bounding the sum of C_t is the hardest and most novel section of our result.

The term C_t comes down to the difference between ML-EVI outputs at two consecutive time steps. Unlike the regret analysis for the UCLK algorithm (Zhou et al., 2021b), where one can trivially bound the sum of the difference by $1/(1-\gamma)$, we introduce a novel method that decomposes the difference and expresses it with terms representing changes in confidence ellipsoids. The effectiveness of this bound relies on the discounted data inheritance scheme we proposed earlier in Section 4.2.

We backtrack the extended value iteration for this bound. Denote the variables $Q_l^{(u)}$ and $V^{(u)}$ in the run of Algorithm 2 at time step t as $Q_{l,t}^{(u)}$ and $V_t^{(u)}$. Next define $D_u := \max_{s\in\mathcal{S}} \big(V_t^{(U_t-u)} - V_{t+1}^{(U_{t+1}-u)}\big)(s)$. By expanding the value functions according to the iteration rules (specified in Lines 3 and 6 of Algorithm 2), we can prove the inequality

$$D_{u-1} \le \gamma D_u + \gamma \max_{\boldsymbol{\phi}_V, l} \big(\max_{\boldsymbol{\theta}_1} \langle \boldsymbol{\theta}_1, \boldsymbol{\phi}_V\rangle - \max_{\boldsymbol{\theta}_2} \langle \boldsymbol{\theta}_2, \boldsymbol{\phi}_V\rangle \big), \tag{6.6}$$

where $\boldsymbol{\theta}_1$ takes the maximum in $\mathcal{C}_t^l \cap \mathcal{B}$, and $\boldsymbol{\theta}_2$ from $\mathcal{C}_{t+1}^l \cap \mathcal{B}$. We can iterate (6.6) with respect to u, and the problem boils down to bounding the difference between confidence sets.

Suppose $\widetilde{\boldsymbol{\theta}}_1 := \operatorname{argmax}_{\boldsymbol{\theta}_1} \langle \boldsymbol{\theta}_1, \boldsymbol{\phi}_V\rangle$. For the maximum concerning $\boldsymbol{\theta}_2$ in (6.6), we need to find a suitable $\widetilde{\boldsymbol{\theta}}_2$ that can be used to lower bound $\max_{\boldsymbol{\theta}_2} \langle \boldsymbol{\theta}_2, \boldsymbol{\phi}_V\rangle$ while being comparable to $\widetilde{\boldsymbol{\theta}}_1$. Starting from $\boldsymbol{\theta}^*$, we go in the direction of $\widetilde{\boldsymbol{\theta}}_1$ until we reach the boundaries of $\mathcal{C}_{t+1}^l \cap \mathcal{B}$ and name our ending point $\widetilde{\boldsymbol{\theta}}_2$. In other words, consider $\mu_{\max} = \sup\{\mu : \boldsymbol{\theta}^* + \mu(\widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*) \in \mathcal{C}_{t+1}^l \cap \mathcal{B}\}$, and let $\widetilde{\boldsymbol{\theta}}_2 = \boldsymbol{\theta}^* + \mu_{\max}(\widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*)$. We have

$$\begin{aligned} &\max_{\boldsymbol{\theta}_1} \langle \boldsymbol{\theta}_1, \boldsymbol{\phi}_V\rangle - \max_{\boldsymbol{\theta}_2} \langle \boldsymbol{\theta}_2, \boldsymbol{\phi}_V\rangle \\ &\le (1-\mu_{\max}) \langle \widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*, \boldsymbol{\phi}_V\rangle. \end{aligned} \tag{6.7}$$

Next, using the fact that $\widetilde{\boldsymbol{\theta}}_2$ belongs to the border of $\mathcal{C}_{t+1}^l \cap \mathcal{B}$, we can obtain an upper bound for $1-\mu_{\max}$. Intuitively, when $\widetilde{\boldsymbol{\theta}}_2 \in \partial\mathcal{C}_{t+1}^l$, we have $\|\widetilde{\boldsymbol{\theta}}_2 - \widehat{\boldsymbol{\theta}}_t^l\|_{\boldsymbol{\Sigma}_t^l} \approx \|\widetilde{\boldsymbol{\theta}}_2 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l} = \beta^l$, and since $\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l\|_{\boldsymbol{\Sigma}_t^l} = \beta^l$, we further have $\widetilde{\boldsymbol{\theta}}_1 \approx \widetilde{\boldsymbol{\theta}}_2$ and $\mu_{\max} \approx 1$. Combining this with (6.7) and (6.6), we obtain an upper bound for $D_0 = \max_{s\in\mathcal{S}}(V_t - V_{t+1})(s)$, and consequently a bound for C_t.

Putting the bounds for A_t, B_t and C_t together, we ultimately acquire an inequality of the form

$$\epsilon N_\epsilon \le C_1 N_\epsilon + C_2\sqrt{N_\epsilon} + C_3,$$

where C_1 can be arbitrarily small with proper parameter choices. Solving this inequality with N_ϵ as the main variable, we can finally obtain an upper bound on N_ϵ that holds with probability at least $1-\delta/2-\delta/2 = 1-\delta$. □

7. Conclusions

We proposed a new algorithm UPAC-UCLK for learning linear kernel MDPs and proved that it is uniform-PAC with sample complexity $\widetilde{O}\big(d^2/((1-\gamma)^4\epsilon^2) + 1/((1-\gamma)^6\epsilon^2)\big)$. This is the first algorithm with uniform PAC guarantee in the discounted MDPs setting. Our result shows that the optimal policy in linear kernel MDPs can be learned efficiently. Our sample complexity bound has a term $\widetilde{O}\big(1/((1-\gamma)^6\epsilon^2)\big)$ that stems from the ML-EVI in our analysis. Achieving minimax optimal sample complexity bound is left as an open question for future work.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. JH and QG are partially supported by the National Science Foundation CAREER Award 1906169. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

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

Agrawal, S. and Jia, R. Posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pp. 1184–1194, 2017.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. F. Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272. JMLR. org, 2017.

Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. *arXiv preprint arXiv:1205.2661*, 2012.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 5713–5723, 2017.

Dong, K., Wang, Y., Chen, X., and Wang, L. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. In *International Conference on Learning Representations*, 2019.

Fruit, R., Pirotta, M., and Lazaric, A. Near optimal exploration-exploitation in non-communicating markov decision processes. *arXiv preprint arXiv:1807.02373*, 2018a.

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

Fruit, R., Pirotta, M., and Lazaric, A. Improved analysis of ucrl2 with empirical bernstein inequality. *arXiv preprint arXiv:2007.05456*, 2020.

He, J., Zhou, D., and Gu, Q. Nearly minimax optimal reinforcement learning for discounted mdps. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021a.

He, J., Zhou, D., and Gu, Q. Uniform-PAC bounds for reinforcement learning with linear function approximation. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021b.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

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

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143, 2020.

Kakade, S. M. et al. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

Lattimore, T. and Hutter, M. Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pp. 320–334. Springer, 2012.

Liu, S. and Su, H. Regret bounds for discounted mdps, 2020.

Modi, A., Jiang, N., Tewari, A., and Singh, S. Sample complexity of reinforcement learning using linearly combined model ensembles. *arXiv preprint arXiv:1910.10597*, 2019.

Ortner, R. Regret bounds for reinforcement learning via markov chain concentration. *Journal of Artificial Intelligence Research*, 67:115–128, 2020.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. Learning unknown markov decision processes: A thompson sampling approach. *arXiv preprint arXiv:1709.04570*, 2017.

Strehl, A. L., Li, L., Wiewiora, E., Langford, J., and Littman, M. L. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 881–888. ACM, 2006.

Szita, I. and Szepesvári, C. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML*, pp. 1031–1038, 2010.

Talebi, M. S. and Maillard, O.-A. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pp. 770–805. PMLR, 2018.

Wei, C.-Y., Jahromi, M. J., Luo, H., and Jain, R. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3007–3015. PMLR, 2021.

Wu, Y., Zhou, D., and Gu, Q. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. *arXiv preprint arXiv:2003.00153*, 2020.

Zhang, Z. and Ji, X. Regret minimization for reinforcement learning by evaluating the optimal bias function. *arXiv preprint arXiv:1906.05110*, 2019.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *arXiv preprint arXiv:2004.10019*, 2020a.

Zhang, Z., Zhou, Y., and Ji, X. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. *arXiv preprint arXiv:2006.03864*, 2020b.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021a.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*. PMLR, 2021b.

A. Proof of the Main Results

We first give a few topological definitions. Under universal set $\mathcal{U} \subset \mathbb{R}^d$, for $x \in \mathbb{R}^d$, define $B_{\mathcal{U}}(x,r) := \{y \in \mathcal{U} : \|y-x\|_2 < r\}$. For $\mathcal{X} \subset \mathcal{U}$, a point $x \in \mathcal{X}$ is its inner point (when constrained to $\mathcal{U}$) if $\exists r > 0, B_{\mathcal{U}}(x,r) \in \mathcal{X}$; a point $x \in \mathcal{U}$ is its boundary point (when constrained to $\mathcal{U}$) if $\forall r > 0, B_{\mathcal{U}}(x,r) \cap \mathcal{X} \neq \varnothing, B_{\mathcal{U}}(x,r) \cap \mathcal{X}^C \neq \varnothing$. Denote the boundary of $\mathcal{X}$ as $\partial_{\mathcal{U}}\mathcal{X}$, which is defined as the set of all boundary points of $\mathcal{X}$ (when constrained to $\mathcal{U}$). If $\mathcal{U} = \mathbb{R}^d$, we will not mention $\mathcal{U}$ when using these terminologies, and we define the corresponding notations in this scenario as $B(x,r) := B_{\mathbb{R}^d}(x,r), \partial := \partial_{\mathbb{R}^d}$.

An affine subspace in $\mathbb{R}^d$ is defined as a subset of $\mathbb{R}^d$ of the form $\{x \in \mathbb{R}^d : L_i(x) = \alpha_i, i = 1\ldots, d_0\}$, where $d_0 \leq d$ is a positive integer, $L_1, \ldots, L_{d_0} : \mathbb{R}^d \to \mathbb{R}$ are mutually independent linear functions of $x \in \mathbb{R}^d$, and $\alpha_1, \ldots, \alpha_{d_0} \in \mathbb{R}$.

Before diving into the main proof, we provide a few more lemmas in aid of our proof.

The first lemma sets up alternative definitions for the variables $\mathbf{\Sigma}_t^l$ and $\mathbf{b}_t^l$, which clarify their utilities and drop the restraint $l \leq 4L_t + C_0$ between l and t.

Lemma A.1. For each time step t and level l, define the alternative and extended expressions for the variables $\mathbf{\Sigma}_t^l := \lambda\mathbf{I} + \sum_{\tau=1}^{t-1} \mathbb{1}\{l_\tau \leq l\} 2^{l_\tau - l} \boldsymbol{\phi}_{V_\tau}(s_\tau.a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau.a_\tau)^\top$ and $\mathbf{b}_t^l := \sum_{\tau=1}^{t-1} \mathbb{1}\{l_\tau \leq l\} 2^{l_\tau - l} \boldsymbol{\phi}_{V_\tau}(s_\tau.a_\tau)V_\tau(s_{\tau+1})$. Then for $l \leq 4L_t + C_0$, the definitions given here yield the same values as the variables given in Algorithm 1.

From now on, for the variables $\mathbf{\Sigma}_t^l$ and $\mathbf{b}_t^l$, we use the definitions given in Lemma A.1. We also define $\widehat{\boldsymbol{\theta}}_t^l := (\mathbf{\Sigma}_t^l)^{-1}\mathbf{b}_t^l$, $\mathcal{C}_t^l := \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t^l\|_{\mathbf{\Sigma}_t^l} \leq \beta^l\}$ for arbitrary t, l.

The next lemma does the same thing for the ML-EVI algorithm. More specifically, it shows that one can extend the total number of levels in Algorithm 2 to infinity and obtain essentially the same results.

Lemma A.2. For each time step t, consider an alternative run of the ML-EVI algorithm with $L = \infty, \mathcal{C}_l = \mathcal{C}_t^l$ for $l = 1, 2, \ldots$ and $U = U_t$. We compare this newly constructed run to the canonical run at time step t.

To do so, we write the variables in this run with an additional suffix $*$, resulting in the notations $Q_{l,*}^{(u)}, V_*^{(u)}$ for iteration round $u = 0, 1, \ldots, U_t$. Then we have

$$
\begin{aligned}
&Q_{l,*}^{(u)} = Q_l^{(u)}, \forall 1 \leq l \leq 4L_t + C_0, \forall 0 \leq u \leq U_t,\\
&Q_{l,*}^{(u)} \geq Q_{4L_t+C_0}^{(u)}, \forall l > 4L_t + C_0, \forall 0 \leq u \leq U_t,\\
&V_*^{(u)} = V^{(u)}, \forall 0 \leq u \leq U_t.
\end{aligned}
$$

We will use this alternative run of ML-EVI instead of the canonical run in future analysis. Thus, we are completely rid of the constraint $l \leq 4L_t + C_0$.

Lemma A.3. The set of $\boldsymbol{\theta} \in \mathbb{R}^d$ that satisfies the conditions for linear kernel MDPs in Definition 3.1, denoted by $\mathcal{B}$, lies within an affine subspace $\mathcal{L} \subset \mathbb{R}^d$; furthermore, it is a convex body (closed and bounded with non-empty interior) when constrained to $\mathcal{L}$.

This lemma describes the general shape of $\mathcal{B}$ and prepares us for future geometrical analysis on the parameter space.

A.1. Proof of Theorem 5.1

We are now ready to prove our main theorem in detail.

Proof of Theorem 5.1. We will work under the event $\mathcal{E}_1$ in this proof, which holds with probability at least $1 - \delta/2$ according to the definition of $\mathcal{E}_1$ and Lemma 6.2. For convenience, our proof is segmented into four parts.

Part I: Sample Complexity Analysis

In the first part, we give an iterable upper bound on the suboptimality gap, and deduce from it an inequality for the sample complexity N_ϵ. Without losing any generality, we assume $\epsilon \leq 1/(1-\gamma)$.

We start by analyzing the suboptimality gap. From Lemma 6.3, we have $V^*(s_t) - V_t^\pi(s_t) \le V_t(s_t) - V_t^\pi(s_t)$. In the following, we seek to deduce an upper bound for $(V_t - V_t^\pi)(s_t)$ that contains the next term $(V_{t+1} - V_{t+1}^\pi)(s_{t+1})$, so we can iterate this bound over t.

As a first step, notice that according to the definitions of value functions for policy π, we have $V_t^\pi(s_t) = Q_t^\pi(s_t, a_t)$; according to Line 12 of Algorithm 2 and Line 8 of Algorithm 1, we have $V_t(s_t) = \max_{a\in\mathcal{A}} \min_{l\ge 1} Q_t^l(s_t, a) = \min_{l\ge 1} Q_t^l(s_t, a_t)$. Combining these equations with Lemma 6.4 and the Bellman equation (3.2), we can deduce

$$\begin{aligned} V_t(s_t) - V_t^\pi(s_t) &= \min_{l\ge 1} Q_t^l(s_t, a_t) - Q_t^\pi(s_t, a_t) \\ &\le Q_t^{l_t - 1}(s_t, a_t) - Q_t^\pi(s_t, a_t) \\ &\le r(s_t, a_t) + \gamma \langle \boldsymbol{\theta}_t^{l_t-1}, \boldsymbol{\phi}_{V_t}(s_t, a_t)\rangle + \gamma^{U_t} - r(s_t, a_t) - \gamma \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{t+1}^\pi}(s_t, a_t)\rangle \\ &= \gamma^{U_t} + \gamma \Big[\big\langle \boldsymbol{\theta}_t^{l_t-1}, \boldsymbol{\phi}_{V_t}(s_t, a_t)\big\rangle - \big\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{t+1}^\pi}(s_t, a_t)\big\rangle \Big]. \end{aligned} \tag{A.1}$$

Now consider the bracketed part of (A.1), we decompose the difference into two parts, namely

$$\big\langle \boldsymbol{\theta}_t^{l_t-1}, \boldsymbol{\phi}_{V_t}(s_t, a_t)\big\rangle - \big\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_{t+1}^\pi}(s_t, a_t)\big\rangle = \underbrace{\big\langle \boldsymbol{\theta}_t^{l_t-1} - \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_t}(s_t, a_t)\big\rangle}_{I_1} + \underbrace{\big\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_t}(s_t, a_t) - \boldsymbol{\phi}_{V_{t+1}^\pi}(s_t, a_t)\big\rangle}_{I_2}. \tag{A.2}$$

For I_1, under the event $\mathcal{E}_1$, we have from Lemma 6.2 that for all l, $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t^l\|_{\boldsymbol{\Sigma}_t^l} \le \beta^l/2$; also, from the condition $\boldsymbol{\theta}_t^l \in \mathcal{C}_t^l \cap \mathcal{B}$ in Lemma 6.4, we have $\|\boldsymbol{\theta}_t^l - \widehat{\boldsymbol{\theta}}_t^l\|_{\boldsymbol{\Sigma}_t^l} \le \beta^l$. On the other hand, from the selection rule of l_t specified in Lines 10 to 13 of Algorithm 1, inequality $\|\boldsymbol{\phi}_{V_t}(s_t, a_t)\|_{(\boldsymbol{\Sigma}_t^{l_t-1})^{-1}} \le 2^{-(l_t-1)}\sqrt{d}/(1-\gamma)$ holds. Together with the Cauchy-Schwartz Inequality $\|x\|_{\boldsymbol{\Sigma}} \cdot \|y\|_{\boldsymbol{\Sigma}^{-1}} \ge \langle x, y\rangle$, we have that

$$\begin{aligned} I_1 &= \big\langle \boldsymbol{\theta}_t^{l_t-1} - \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_t}(s_t, a_t)\big\rangle \\ &\le \|\boldsymbol{\theta}_t^{l_t-1} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t^{l_t-1}} \|\boldsymbol{\phi}_{V_t}(s_t, a_t)\|_{(\boldsymbol{\Sigma}_t^{l_t-1})^{-1}} \\ &\le \big(\|\boldsymbol{\theta}_t^{l_t-1} - \widehat{\boldsymbol{\theta}}_t^{l_t-1}\|_{\boldsymbol{\Sigma}_t^{l_t-1}} + \|\widehat{\boldsymbol{\theta}}_t^{l_t-1} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t^{l_t-1}}\big) \cdot \|\boldsymbol{\phi}_{V_t}(s_t, a_t)\|_{(\boldsymbol{\Sigma}_t^{l_t-1})^{-1}} \\ &\le (\beta^{l_t-1} + \frac{1}{2}\beta^{l_t-1}) \cdot 2^{-(l_t-1)}\sqrt{d}/(1-\gamma) \\ &= 3 \cdot 2^{l_0 - l_t}\beta^{l_t-1}, \end{aligned} \tag{A.3}$$

where in the second inequality we used the triangle inequality, and in the final equation we used the definition of l_0 from Lemma 6.1.

For I_2, recall from (3.3) that $\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_V(s_t, a_t)\rangle = \mathbb{P}V(s_t, a_t)$ for arbitrary V, therefore

$$\begin{aligned} I_2 &= \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_t}(s_t, a_t) - \boldsymbol{\phi}_{V_{t+1}^\pi}(s_t, a_t)\rangle \\ &= \mathbb{P}(V_t - V_{t+1}^\pi)(s_t, a_t). \end{aligned} \tag{A.4}$$

Additionally, from the definition of U_t in (5.1), we have

$$\begin{aligned} \gamma^{U_t} &\le \gamma^{\log[t(t+1)]/(1-\gamma)} \\ &= \Big[\big(1 - (1-\gamma)\big)^{1/(1-\gamma)}\Big]^{\log[t(t+1)]} \\ &\le \big[e^{-1}\big]^{\log[t(t+1)]} \\ &= \frac{1}{t(t+1)}. \end{aligned} \tag{A.5}$$

Now substitute (A.3) and (A.4) into (A.2) and then into (A.1) along with the above inequality, to obtain

$$\begin{aligned} V_t(s_t) - V_t^\pi(s_t) &\le \gamma^{U_t} + \gamma\big[3\cdot 2^{l_0-l_t}\beta^{l_t-1} + \mathbb{P}(V_t - V_{t+1}^\pi)(s_t,a_t)\big] \\ &\le \frac{1}{t(t+1)} + 3\gamma\cdot 2^{l_0-l_t}\beta^{l_t-1} + \gamma\big[\mathbb{P}(V_t - V_{t+1}^\pi)(s_t,a_t) - (V_t - V_{t+1}^\pi)(s_{t+1})\big] + \gamma(V_t - V_{t+1}^\pi)(s_{t+1}) \\ &= \Gamma(t) + \gamma\Delta(t) + \gamma(V_t - V_{t+1})(s_{t+1}) + \gamma(V_{t+1} - V_{t+1}^\pi)(s_{t+1}), \end{aligned} \tag{A.6}$$

where in the final equation we define $\Gamma(t) := 1/\big(t(t+1)\big) + 3\gamma\cdot 2^{l_0-l_t}\beta^{l_t-1}$ and $\Delta(t) := \mathbb{P}(V_t - V_{t+1}^\pi)(s_t,a_t) - (V_t - V_{t+1}^\pi)(s_{t+1})$.

Iterate Inequality (A.6) for T rounds (through $t, t+1, \dots, t+T$), and we get the following suboptimality bound:

$$\begin{aligned} V^*(s_t) - V_t^\pi(s_t) &\le V_t(s_t) - V_t^\pi(s_t) \\ &\le \sum_{\tau=t}^{t+T-1}\gamma^{\tau-t}\big[\Gamma(\tau) + \gamma\Delta(\tau) + \gamma(V_\tau - V_{\tau+1})(s_{\tau+1})\big] + \gamma^T(V_{t+T} - V_{t+T}^\pi)(s_{t+T}) \\ &\le \underbrace{\sum_{\tau=t}^{t+T-1}\gamma^{\tau-t}\Gamma(\tau)}_{A_t} + \underbrace{\sum_{\tau=t}^{t+T-1}\gamma^{\tau-t+1}\Delta(r)}_{B_t} + \underbrace{\sum_{\tau=t}^{t+T-1}\gamma^{\tau-t+1}(V_\tau - V_{\tau+1})(s_{\tau+1})}_{C_t} + \frac{\gamma^T}{1-\gamma}, \end{aligned} \tag{A.7}$$

where the last inequality is due to $0 \le V_{t+T}^\pi(s_{t+T}), V_{t+T}(s_{t+T}) \le 1/(1-\gamma)$.

Now suppose the set of all time steps where the suboptimality gap is greater than ϵ is $\mathcal{T}_\epsilon := \{t : V^*(s_t) - V_t^\pi(s_t) > \epsilon\}$, with a total number of $N_\epsilon = \#\mathcal{T}_\epsilon$ elements. Define the condition $\mathcal{Q}_t = \{l_\tau > l_0, \forall t \le \tau < t+T\}$. Since we also have trivially that $V^*(s_t) - V_t^\pi(s_t) \le 1/(1-\gamma)$, we can obtain by summing (A.7) over all time steps t in set $\mathcal{T}_\epsilon$ that

$$\begin{aligned} \epsilon N_\epsilon &\le \sum_{t\in\mathcal{T}_\epsilon}\big(V^*(s_t) - V_t^\pi(s_t)\big) \\ &\le \sum_{t\in\mathcal{T}_\epsilon}\mathbb{1}\{\mathcal{Q}_t\}\left(\frac{\gamma^T}{1-\gamma} + A_t + B_t + C_t\right) + \sum_{t\in\mathcal{T}_\epsilon}\mathbb{1}\{\overline{\mathcal{Q}_t}\}\frac{1}{1-\gamma} \\ &\le \frac{\gamma^T}{1-\gamma}N_\epsilon + \sum_{t\in\mathcal{T}_\epsilon}\mathbb{1}\{\mathcal{Q}_t\}(A_t + B_t + C_t) + \frac{1}{1-\gamma}\sum_{t=1}^{\infty}\mathbb{1}\{\overline{\mathcal{Q}_t}\}. \end{aligned}$$

Consider now the indicator function for $\overline{\mathcal{Q}_t}$. Only when $l_\tau \le l_0$ holds for some $\tau \in \{t, \dots, t+T-1\}$ is $\mathcal{Q}_t$ false, therefore $\mathbb{1}\{\overline{\mathcal{Q}_t}\} \le \sum_{\tau=t}^{t+T-1}\mathbb{1}\{l_\tau \le l_0\} = \sum_{l=1}^{\lfloor l_0\rfloor}\sum_{\tau=t}^{t+T-1}\mathbb{1}\{l_\tau = l\}$. Combine this with the definition of $k_{\max}^l = \lim_{t\to\infty} k_t^l = \sum_{\tau=1}^{\infty}\mathbb{1}\{l_\tau \le l\}2^{l_\tau - l}$ in Lemma 6.1, the last addend above can be bounded by

$$\begin{aligned} \frac{1}{1-\gamma}\sum_{t=1}^{\infty}\mathbb{1}\{\overline{\mathcal{Q}_t}\} &\le \frac{1}{1-\gamma}\sum_{t=1}^{\infty}\sum_{l=1}^{\lfloor l_0\rfloor}\sum_{\tau=t}^{t+T-1}\mathbb{1}\{l_\tau = l\} \\ &\le \frac{T}{1-\gamma}\sum_{l=1}^{\lfloor l_0\rfloor}\sum_{\tau=1}^{\infty}\mathbb{1}\{l_\tau = l\} \\ &= \frac{T}{1-\gamma}\sum_{l=1}^{\lfloor l_0\rfloor}\left(k_{\max}^l - \frac{1}{2}k_{\max}^{l-1}\right) \\ &= \frac{T}{1-\gamma}\left(k_{\max}^{\lfloor l_0\rfloor} + \frac{1}{2}\sum_{l=1}^{\lfloor l_0\rfloor - 1}k_{\max}^l\right) \\ &\le \frac{T}{1-\gamma}\cdot\frac{l_0+1}{2}\cdot 20dl_0 \\ &= \frac{10dT}{1-\gamma}l_0(l_0+1), \end{aligned}$$

where the second inequality is based on the observation that each τ is visited in the last summation at most T times, and the third inequality is according to the conclusions of Lemma 6.1. Hence we have the inequality

$$\epsilon N_\epsilon \leq \frac{\gamma^T}{1-\gamma} N_\epsilon + \sum_{t \in \mathcal{T}_\epsilon} \mathbb{1}\{\mathcal{Q}_t\} A_t + \sum_{t \in \mathcal{T}_\epsilon} \mathbb{1}\{\mathcal{Q}_t\} B_t + \sum_{t \in \mathcal{T}_\epsilon} \mathbb{1}\{\mathcal{Q}_t\} C_t + \frac{10dT}{1-\gamma} l_0(l_0+1). \tag{A.8}$$

Part II: Upper Bounds for the Summation of A_t and B_t

For this next part, we bound the second and third terms on the right hand side of (A.8). First, for the sum of the term A_t, we expand the expression to the following:

$$\begin{aligned}
\sum_{t \in \mathcal{T}_\epsilon} \mathbb{1}\{\mathcal{Q}_t\} A_t &= \sum_{t \in \mathcal{T}_\epsilon} \sum_{\tau=t}^{t+T-1} \mathbb{1}\{\mathcal{Q}_t\} \gamma^{\tau-t} \Gamma(\tau) \\
&\leq \sum_{t \in \mathcal{T}_\epsilon} \sum_{\tau=t}^{t+T-1} \mathbb{1}\{l_\tau > l_0\} \gamma^{\tau-t} \Gamma(\tau) \\
&= \sum_{t \in \mathcal{T}_\epsilon} \sum_{\tau=t}^{t+T-1} \mathbb{1}\{l_\tau > l_0\} \gamma^{\tau-t} \left[\frac{1}{\tau(\tau+1)} + 3\gamma \cdot 2^{l_0 - l_\tau} \beta^{l_\tau - 1} \right] \\
&\leq \sum_{t \in \mathcal{T}_\epsilon} \sum_{\tau=t}^{t+T-1} \frac{\gamma^{\tau-t}}{\tau(\tau+1)} + 3 \sum_{t \in \mathcal{T}_\epsilon} \sum_{\tau=t}^{t+T-1} \mathbb{1}\{l_\tau > l_0\} \gamma^{\tau-t+1} 2^{l_0 - l_\tau} \beta^{l_\tau - 1}.
\end{aligned} \tag{A.9}$$

Regarding the first summation in (A.9), we shrink the τ's in the denominator to t:

$$\begin{aligned}
\sum_{t \in \mathcal{T}_\epsilon} \sum_{\tau=t}^{t+T-1} \frac{\gamma^{\tau-t}}{\tau(\tau+1)} &\leq \sum_{t \in \mathcal{T}_\epsilon} \sum_{\tau=t}^{t+T-1} \frac{\gamma^{\tau-t}}{t(t+1)} \\
&= \sum_{t \in \mathcal{T}_\epsilon} \frac{1}{t(t+1)} \sum_{\tau=t}^{t+T-1} \gamma^{\tau-t} \\
&\leq \sum_{t \in \mathcal{T}_\epsilon} \frac{1}{t(t+1)} \cdot \frac{1}{1-\gamma} \\
&\leq \frac{1}{1-\gamma}.
\end{aligned} \tag{A.10}$$

We hence obtain by substituting this into (A.9) that

$$\sum_{t \in \mathcal{T}_\epsilon} A_t \leq \frac{1}{1-\gamma} + 3 \sum_{t \in \mathcal{T}_\epsilon} \sum_{\tau=t}^{t+T-1} \mathbb{1}\{l_\tau > l_0\} \gamma^{\tau-t+1} 2^{l_0 - l_\tau} \beta^{l_\tau - 1}. \tag{A.11}$$

We leave the leftover part to be tackled in Part IV.

Second, we discard the indicator term in the sum of B_t and expand it to

$$\begin{aligned}
\sum_{t \in \mathcal{T}_\epsilon} \mathbb{1}\{\mathcal{Q}_t\} B_t &\leq \sum_{t \in \mathcal{T}_\epsilon} \sum_{\tau=t}^{t+T-1} \gamma^{\tau-t+1} \Delta(\tau) \\
&= \sum_{t \in \mathcal{T}_\epsilon} \sum_{n=0}^{T-1} \gamma^{n+1} \Delta(t+n) \\
&= \sum_{n=0}^{T-1} \left[\sum_{t \in \mathcal{T}_\epsilon} \gamma^{n+1} \left[\mathbb{P}(V_{t+n} - V^\pi_{t+n+1})(s_{t+n}, a_{t+n}) - (V_{t+n} - V^\pi_{t+n+1})(s_{t+n+1}) \right] \right],
\end{aligned}$$

where in the first equation we substitute $n = \tau - t$, and in the second we exchange the ordering of summation.

Write $\mathcal{T}_\epsilon = \{t_1, t_2, \ldots, t_{N_\epsilon}\}$, then for $k = 1, \ldots, N_\epsilon$, the random variable t_k is the k-th time step where the suboptimality gap is larger than ϵ and hence a stopping time. Now for each individual $n \geq 0$, let $\mathcal{F}_{k,n}$ be the σ-field generated by all random variables before the generation of s_{t_k+n+1} in Line 9 of Algorithm 1 at time step $t_k + n$. Then

$$\left\{\mathbb{P}\big(V_{t_k+n} - V^\pi_{t_k+n+1}\big)(s_{t_k+n}, a_{t_k+n}) - \big(V_{t_k+n} - V^\pi_{t_k+n+1}\big)(s_{t_k+n+1})\right\}_{k=1}^{N_\epsilon}$$

is a martingale sequence with respect to the filtration $\{\mathcal{F}_{k,n}\}_{k=0}^{N_\epsilon}$. Now notice the term $(V_{t+n} - V^\pi_{t+n+1})(s)$ is bounded within $[-1/(1-\gamma), 1/(1-\gamma)]$, and therefore $\mathbb{P}\big(V_{t+n} - V^\pi_{t+n+1}\big)(s_{t+n}, a_{t+n}) - \big(V_{t+n} - V^\pi_{t+n+1}\big)(s_{t+n+1})$ is bounded within $[-2/(1-\gamma), 2/(1-\gamma)]$. Consequently, from Lemma D.1, with probability at least $1 - \delta/\big(2(n+1)(n+2)\big)$, the following holds:

$$\sum_{t\in\mathcal{T}_\epsilon}\left[\mathbb{P}\big(V_{t+n} - V^\pi_{t+n+1}\big)(s_{t+n}, a_{t+n}) - \big(V_{t+n} - V^\pi_{t+n+1}\big)(s_{t+n+1})\right] \leq \frac{2}{1-\gamma}\sqrt{2N_\epsilon \log\big(2(n+1)(n+2)/\delta\big)}.$$

Taking a union bound, we have that with probability at least $1 - \sum_{n=0}^{\infty}\delta/\big(2(n+1)(n+2)\big) = 1 - \delta/2$, the subsequent upper bound holds:

$$\begin{aligned}
\sum_{t\in\mathcal{T}_\epsilon} B_t &= \sum_{n=0}^{T-1}\left[\gamma^{n+1}\sum_{t\in\mathcal{T}_\epsilon}\left[\mathbb{P}(V_{t+n} - V^\pi_{t+n+1})(s_{t+n}, a_{t+n}) - (V_{t+n} - V^\pi_{t+n+1})(s_{t+n+1})\right]\right] \\
&\leq \sum_{n=0}^{T-1}\gamma^{n+1}\cdot\frac{2}{1-\gamma}\sqrt{2N_\epsilon\log\big(2(n+1)(n+2)/\delta\big)} \\
&\leq \sum_{n=0}^{T-1}\frac{2\gamma^{n+1}}{1-\gamma}\sqrt{2N_\epsilon\log\big(2T(T+1)/\delta\big)} \\
&\leq \frac{2\gamma}{(1-\gamma)^2}\sqrt{2N_\epsilon\log\big(2T(T+1)/\delta\big)}, \qquad \text{(A.12)}
\end{aligned}$$

where in the second inequality, we amplified the $n+1$ under the square roots to T. We denote the event where this inequality holds as $\mathcal{E}_2$, then with probability at least $1 - \delta/2 - \delta/2 = 1 - \delta$, the joint event $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$ holds.

Part III: Upper bound for the Summation of C_t

For the third part, we provide an upper bound for $\sum_{t\in\mathcal{T}_\epsilon} C_t$. Since this process is rather lengthy, we partition it into four steps.

Step 1, we seek to obtain a bound for $\big(V_t - V_{t+1}\big)(s_{t+1})$. According to Line 7 of Algorithm 1, the two value functions V_t, V_{t+1} are results of Algorithm 2 from two consecutive time steps t and $t+1$.

Denote the variables $Q_l^{(u)}, V^{(u)}$ in Algorithm 2 at time step t by $Q_{l,t}^{(u)}, V_t^{(u)}$ for iteration round $u = 1, \ldots, U_t$ and $t = 1, 2, \ldots$. Next for a fixed t, define $D_u := \max_{s\in\mathcal{S}}\big[V_t^{(U_t-u)} - V_{t+1}^{(U_{t+1}-u)}\big](s)$. By the update rules of V and Q in

respectively Line 3 and Line 6 of Algorithm 2, we now give the following derivations:

$$\begin{aligned}
D_{u-1} &= \max_{s\in\mathcal{S}}\left[V_t^{(U_t-u+1)} - V_{t+1}^{(U_{t+1}-u+1)}\right](s)\\
&= \max_{s\in\mathcal{S}}\left[\max_{a\in\mathcal{A}}\min_{l\geq 1} Q_{l,t}^{(U_t-u+1)}(s,a) - \max_{a\in\mathcal{A}}\min_{l\geq 1} Q_{l,t+1}^{(U_{t+1}-u+1)}(s,a)\right]\\
&\leq \max_{s\in\mathcal{S},a\in\mathcal{A}}\left[\min_{l\geq 1} Q_{l,t}^{(U_t-u+1)} - \min_{l\geq 1} Q_{l,t+1}^{(U_{t+1}-u+1)}\right](s,a)\\
&\leq \max_{s\in\mathcal{S},a\in\mathcal{A},l\geq 1}\left[Q_{l,t}^{(U_t-u+1)} - Q_{l,t+1}^{(U_{t+1}-u+1)}\right](s,a)\\
&= \max_{s,a,l}\left[\left(r(s,a)+\gamma\max_{\boldsymbol{\theta}_1\in\mathcal{C}_t^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_1,\boldsymbol{\phi}_{V_t^{(U_t-u)}}(s,a)\right\rangle\right) - \left(r(s,a)+\gamma\max_{\boldsymbol{\theta}_2\in\mathcal{C}_{t+1}^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_2,\boldsymbol{\phi}_{V_{t+1}^{(U_{t+1}-u)}}(s,a)\right\rangle\right)\right]\\
&= \gamma\max_{s,a,l}\left[\max_{\boldsymbol{\theta}_1\in\mathcal{C}_t^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_1,\boldsymbol{\phi}_{V_t^{(U_t-u)}}(s,a)\right\rangle - \max_{\boldsymbol{\theta}_2\in\mathcal{C}_{t+1}^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_2,\boldsymbol{\phi}_{V_{t+1}^{(U_{t+1}-u)}}(s,a)\right\rangle\right],
\end{aligned}\tag{A.13}$$

where we used the trivial facts $\max_{x\in\mathcal{X}} F(x)-\max_{x\in\mathcal{X}} G(x)\leq\max_{x\in\mathcal{X}}(F-G)(x)$ and $\min_{x\in\mathcal{X}} F(x)-\min_{x\in\mathcal{X}} G(x)\leq\max_{x\in\mathcal{X}}(F-G)(x)$ in the inequalities above.

Next we denote for ease of expression that $V_1 = V_t^{(U_t-u)}, V_2 = V_{t+1}^{(U_{t+1}-u)}$, and then separate the difference inside the outer maximum in (A.13) into two parts:

$$\begin{aligned}
&\max_{\boldsymbol{\theta}_1\in\mathcal{C}_t^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_1,\boldsymbol{\phi}_{V_1}(s,a)\right\rangle - \max_{\boldsymbol{\theta}_2\in\mathcal{C}_{t+1}^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_2,\boldsymbol{\phi}_{V_2}(s,a)\right\rangle\\
&\leq \underbrace{\max_{\boldsymbol{\theta}_1\in\mathcal{C}_t^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_1,\boldsymbol{\phi}_{V_1}(s,a)\right\rangle - \max_{\boldsymbol{\theta}_1\in\mathcal{C}_t^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_1,\boldsymbol{\phi}_{V_2}(s,a)\right\rangle}_{J_1} + \underbrace{\max_{\boldsymbol{\theta}_1\in\mathcal{C}_t^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_1,\boldsymbol{\phi}_{V_2}(s,a)\right\rangle - \max_{\boldsymbol{\theta}_2\in\mathcal{C}_{t+1}^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_2,\boldsymbol{\phi}_{V_2}(s,a)\right\rangle}_{J_2}.
\end{aligned}$$

Recall the definition of $\boldsymbol{\phi}_V$ from Definition 3.1, the first part

$$\begin{aligned}
J_1 &\leq \max_{\boldsymbol{\theta}_1\in\mathcal{C}_t^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_1,\left(\boldsymbol{\phi}_{V_1}-\boldsymbol{\phi}_{V_2}\right)(s,a)\right\rangle\\
&= \max_{\boldsymbol{\theta}_1\in\mathcal{C}_t^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_1,\int_{s'\in\mathcal{S}}\boldsymbol{\phi}(s'|s,a)(V_1-V_2)(s')\mathrm{d}s'\right\rangle\\
&= \max_{\boldsymbol{\theta}_1\in\mathcal{C}_t^l\cap\mathcal{B}}\int_{s'\in\mathcal{S}}\left\langle\boldsymbol{\theta}_1,\boldsymbol{\phi}(s'|s,a)\right\rangle(V_1-V_2)(s')\mathrm{d}s'.
\end{aligned}\tag{A.14}$$

By the definition of $\mathcal{B}$ in (3.4), for $\boldsymbol{\theta}\in\mathcal{B}$, $\langle\boldsymbol{\theta},\boldsymbol{\phi}(s'|s,a)\rangle$ is a probability distribution on $\mathcal{S}$, which we now denote by $\widetilde{\mathbb{P}} = \langle\boldsymbol{\theta},\boldsymbol{\phi}\rangle$. We then continue from (A.14):

$$\begin{aligned}
J_1 &\leq \max_{\boldsymbol{\theta}_1\in\mathcal{C}_t^l\cap\mathcal{B},\widetilde{\mathbb{P}}=\langle\boldsymbol{\theta}_1,\boldsymbol{\phi}\rangle}\int_{s'\in\mathcal{S}}\widetilde{\mathbb{P}}(s'|s,a)(V_1-V_2)(s')\mathrm{d}s'\\
&\leq \max_{\boldsymbol{\theta}_1\in\mathcal{B},\widetilde{\mathbb{P}}=\langle\boldsymbol{\theta}_1,\boldsymbol{\phi}\rangle}\mathbb{E}_{s'\sim\widetilde{\mathbb{P}}(\cdot|s,a)}\left[(V_1-V_2)(s')\middle|s,a\right]\\
&\leq \max_{\boldsymbol{\theta}_1\in\mathcal{B},\widetilde{\mathbb{P}}=\langle\boldsymbol{\theta}_1,\boldsymbol{\phi}\rangle}\max_{s'\in\mathcal{S}}(V_1-V_2)(s')\\
&= \max_{s'\in\mathcal{S}}\left[V_t^{(U_t-u)} - V_{t+1}^{(U_{t+1}-u)}\right](s'),
\end{aligned}\tag{A.15}$$

where in the second inequality we relaxed $\boldsymbol{\theta}_1\in\mathcal{C}_t^l\cap\mathcal{B}$ to $\boldsymbol{\theta}_1\in\mathcal{B}$, and the third inequality is due to the fact that the expected value of a random variable is no greater than its maximum possible value. After that, the equation follows because the variable $\boldsymbol{\theta}_1$ is no longer tied to the value of $(V_1-V_2)(s')$.

For J_2, we relax $V_2 = V_{t+1}^{(U_{t+1}-u)}$ to an arbitrary function $V:\mathcal{S}\to[0,1/(1-\gamma)]$:

$$J_2 \leq \max_{V:\mathcal{S}\to[0,1/(1-\gamma)]}\left[\max_{\boldsymbol{\theta}_1\in\mathcal{C}_t^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_1,\boldsymbol{\phi}_V(s,a)\right\rangle - \max_{\boldsymbol{\theta}_2\in\mathcal{C}_{t+1}^l\cap\mathcal{B}}\left\langle\boldsymbol{\theta}_2,\boldsymbol{\phi}_V(s,a)\right\rangle\right].\tag{A.16}$$

Now substitute (A.15) and (A.16) into (A.13):

$$\begin{aligned}
D_{u-1} &\le \gamma \max_{s,a,l}(J_1 + J_2) \\
&\le \gamma \max_{s,a,l} \Bigg[\max_{s' \in \mathcal{S}} \left(V_t^{(U_t - u)} - V_{t+1}^{(U_{t+1} - u)} \right)(s') \\
&\qquad + \max_{V : \mathcal{S} \to [0, 1/(1-\gamma)]} \left(\max_{\boldsymbol{\theta}_1 \in \mathcal{C}_t^l \cap \mathcal{B}} \left\langle \boldsymbol{\theta}_1, \boldsymbol{\phi}_V(s,a) \right\rangle - \max_{\boldsymbol{\theta}_2 \in \mathcal{C}_{t+1}^l \cap \mathcal{B}} \left\langle \boldsymbol{\theta}_2, \boldsymbol{\phi}_V(s,a) \right\rangle \right) \Bigg] \\
&= \gamma \max_{s' \in \mathcal{S}} \left[V_t^{(U_t - u)} - V_{t+1}^{(U_{t+1} - u)} \right](s') + \gamma \max_{s,a,l,V} \left[\max_{\boldsymbol{\theta}_1 \in \mathcal{C}_t^l \cap \mathcal{B}} \left\langle \boldsymbol{\theta}_1, \boldsymbol{\phi}_V(s,a) \right\rangle - \max_{\boldsymbol{\theta}_2 \in \mathcal{C}_{t+1}^l \cap \mathcal{B}} \left\langle \boldsymbol{\theta}_2, \boldsymbol{\phi}_V(s,a) \right\rangle \right] \\
&= \gamma D_u + \gamma \max_{s,a,l,V} \left[\max_{\boldsymbol{\theta}_1 \in \mathcal{C}_t^l \cap \mathcal{B}} \left\langle \boldsymbol{\theta}_1, \boldsymbol{\phi}_V(s,a) \right\rangle - \max_{\boldsymbol{\theta}_2 \in \mathcal{C}_{t+1}^l \cap \mathcal{B}} \left\langle \boldsymbol{\theta}_2, \boldsymbol{\phi}_V(s,a) \right\rangle \right],
\end{aligned}$$

where the first equation holds because the first term in the big bracket is independent of the variables s, a, l, and the final equation holds based on the definition of D_u. Further note that according to Algorithm 1, the confidence sets at levels lower than l_t are not updated from t to $t+1$, and hence these confidence sets are identical for the two runs of ML-EVI. Therefore, we may exclude these smaller levels in the outside maximization above, and deduce

$$D_{u-1} \le \gamma D_u + \gamma \max_{s,a,l \ge l_t, V} \left[\max_{\boldsymbol{\theta}_1 \in \mathcal{C}_t^l \cap \mathcal{B}} \left\langle \boldsymbol{\theta}_1, \boldsymbol{\phi}_V(s,a) \right\rangle - \max_{\boldsymbol{\theta}_2 \in \mathcal{C}_{t+1}^l \cap \mathcal{B}} \left\langle \boldsymbol{\theta}_2, \boldsymbol{\phi}_V(s,a) \right\rangle \right]. \tag{A.17}$$

Step 2, for a given $\boldsymbol{\psi} = \boldsymbol{\phi}_V(s,a)$, we consider the expression in (A.17) inside the outer maximum, namely $\max_{\boldsymbol{\theta}_1 \in \mathcal{C}_t^l \cap \mathcal{B}} \langle \boldsymbol{\theta}_1, \boldsymbol{\psi} \rangle - \max_{\boldsymbol{\theta}_2 \in \mathcal{C}_{t+1}^l \cap \mathcal{B}} \langle \boldsymbol{\theta}_2, \boldsymbol{\psi} \rangle$.

Suppose $\widetilde{\boldsymbol{\theta}}_1 := \operatorname{argmax}_{\boldsymbol{\theta}_1 \in \overline{\mathcal{C}_t^l \cap \mathcal{B}}} \langle \boldsymbol{\theta}_1, \boldsymbol{\psi} \rangle$. If $\widetilde{\boldsymbol{\theta}}_1 \in \mathcal{C}_{t+1}^l$, we immediately have $\max_{\boldsymbol{\theta}_1 \in \mathcal{C}_t^l \cap \mathcal{B}} \langle \boldsymbol{\theta}_1, \boldsymbol{\psi} \rangle = \langle \widetilde{\boldsymbol{\theta}}_1, \boldsymbol{\psi} \rangle \le \max_{\boldsymbol{\theta}_2 \in \mathcal{C}_{t+1}^l \cap \mathcal{B}} \langle \boldsymbol{\theta}_2, \boldsymbol{\psi} \rangle$ and the expression of interest is upper bounded by 0. Otherwise when $\widetilde{\boldsymbol{\theta}}_1 \notin \mathcal{C}_{t+1}^l$, consider $\mu_{\max} = \sup\{\mu : \boldsymbol{\theta}^* + \mu(\widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*) \in \mathcal{C}_{t+1}^l \cap \mathcal{B}\}$.

According to Lemma A.3, $\mathcal{B}$ is a convex body with non-empty interior when constrained to the affine subspace $\mathcal{L} \subset \mathbb{R}^d$. Combine this with $\boldsymbol{\theta}^*, \widetilde{\boldsymbol{\theta}}_1 \in \mathcal{B}$, we see the segment $\mathcal{J} \subset \mathbb{R}^d$ with $\boldsymbol{\theta}^*$ and $\widetilde{\boldsymbol{\theta}}_1$ as endpoints is contained by $\mathcal{B}$. Since $\mathcal{C}_{t+1}^l$ is a closed ellipsoid and hence a convex body in $\mathbb{R}^d$, its intersection with $\mathcal{B}$ must also be a convex body when constrained to $\mathcal{L}$. Furthermore, by the conclusions of Lemma 6.2, $\boldsymbol{\theta}^*$ is an inner point in $\mathcal{C}_{t+1}^l$ under $\mathcal{E}_1$, so the intersection of $\mathcal{C}_{t+1}^l$ and $\mathcal{J}$ must be a nontrivial segment $\mathcal{J}'$. Since the set $\{\boldsymbol{\theta} = \boldsymbol{\theta}^* + \mu(\widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*) \in \mathcal{C}_{t+1}^l \cap \mathcal{B}\}$, as the intersection of a line and a convex set, contains $\mathcal{J}'$, the set $\{\mu : \boldsymbol{\theta}^* + \mu(\widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*) \in \mathcal{C}_{t+1}^l \cap \mathcal{B}\}$ must be a non-trivial closed interval $\mathcal{I}$ on $\mathbb{R}$.

Since $\boldsymbol{\theta}^* \in \mathcal{C}_{t+1}^l \cap \mathcal{B}$ under $\mathcal{E}_1$ and $\widetilde{\boldsymbol{\theta}}_1 \notin \mathcal{C}_{t+1}^l$ by our earlier assumption, we have $0 \in \mathcal{I}$ and $1 \notin \mathcal{I}$, and therefore $0 \le \mu_{\max} < 1$. Now define $\widetilde{\boldsymbol{\theta}}_2 := \boldsymbol{\theta}^* + \mu_{\max}(\widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*) \in \mathcal{C}_{t+1}^l \cap \mathcal{B}$, and obtain:

$$\begin{aligned}
\max_{\boldsymbol{\theta}_1 \in \mathcal{C}_t^l \cap \mathcal{B}} \langle \boldsymbol{\theta}_1, \boldsymbol{\psi} \rangle - \max_{\boldsymbol{\theta}_2 \in \mathcal{C}_{t+1}^l \cap \mathcal{B}} \langle \boldsymbol{\theta}_2, \boldsymbol{\psi} \rangle &= \langle \widetilde{\boldsymbol{\theta}}_1, \boldsymbol{\psi} \rangle - \max_{\boldsymbol{\theta}_2 \in \mathcal{C}_{t+1}^l \cap \mathcal{B}} \langle \boldsymbol{\theta}_2, \boldsymbol{\psi} \rangle \\
&\le \langle \widetilde{\boldsymbol{\theta}}_1, \boldsymbol{\psi} \rangle - \langle \boldsymbol{\theta}^* + \mu_{\max}(\widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*), \boldsymbol{\psi} \rangle \\
&= (1 - \mu_{\max}) \langle \widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*, \boldsymbol{\psi} \rangle.
\end{aligned} \tag{A.18}$$

It is worth noting that the second multiplier $\langle \widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*, \boldsymbol{\psi} \rangle$ is non-negative based on the definition of $\widetilde{\boldsymbol{\theta}}_1$.

Next, we start with giving a bound for the second multiplier. Since $\widetilde{\boldsymbol{\theta}}_1 \in \mathcal{B}$, we have $\widetilde{\mathbb{P}}(\cdot|s,a) = \langle \widetilde{\boldsymbol{\theta}}_1, \boldsymbol{\phi}(\cdot|s,a) \rangle$ is a probability function for arbitrary $s \in \mathcal{S}, a \in \mathcal{A}$, and therefore

$$\begin{aligned}
\langle \widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*, \boldsymbol{\psi} \rangle &= \langle \widetilde{\boldsymbol{\theta}}_1, \boldsymbol{\psi} \rangle - \langle \boldsymbol{\theta}^*, \boldsymbol{\psi} \rangle \\
&= \widetilde{\mathbb{P}} V(s,a) - \mathbb{P} V(s,a) \\
&\le \frac{1}{1-\gamma},
\end{aligned} \tag{A.19}$$

where we used $0 \le V(s) \le 1/(1-\gamma)$ for arbitrary $s \in \mathcal{S}$.

Now we focus on $\mu_{\max}$. By its definition, we have $\widetilde{\boldsymbol{\theta}}_2 = \boldsymbol{\theta}^* + \mu_{\max}(\widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*) \in \partial_{\mathcal{L}}(\mathcal{C}_{t+1}^l \cap \mathcal{B})$. If $\widetilde{\boldsymbol{\theta}}_2 \in \partial_{\mathcal{L}}\mathcal{B}$, since for $0 \le \mu < \mu_{\max}$, $\boldsymbol{\theta}^* + \mu(\widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*) \in \mathcal{B}$, we can deduce that for any $\mu > \mu_{\max}$ that $\boldsymbol{\theta}^* + \mu(\widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*) \notin \mathcal{B}$. Because $\mu_{\max} < 1$, this leads to $\widetilde{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}^* + 1 \cdot (\widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*) \notin \mathcal{B}$, a contradiction. Otherwise $\widetilde{\boldsymbol{\theta}}_2 \in \partial_{\mathcal{L}}\mathcal{C}_{t+1}^l \subset \partial\mathcal{C}_{t+1}^l$, which implies

$$
\begin{aligned}
\beta^l &= \|\widetilde{\boldsymbol{\theta}}_2 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l} \\
&= \|\mu_{\max}\widetilde{\boldsymbol{\theta}}_1 + (1-\mu_{\max})\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l} \\
&= \|\mu_{\max}(\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l) + (1-\mu_{\max})(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{t+1}^l)\|_{\boldsymbol{\Sigma}_{t+1}^l} \\
&\le \mu_{\max}\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l} + (1-\mu_{\max})\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l} \\
&\le \mu_{\max}\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l} + (1-\mu_{\max}) \cdot \frac{1}{2}\beta^l,
\end{aligned} \tag{A.20}
$$

where we used the triangle inequality in the first inequality, and the conclusion of Lemma 6.2 in the second inequality.

Moving the second addend on the right hand side of (A.20) to the left hand side, we get

$$
\begin{aligned}
\frac{1}{2}(1+\mu_{\max})\beta^l &\le \mu_{\max}\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l} \\
&\le \frac{1}{2}(1+\mu_{\max})\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l}, \qquad \text{(since } \mu_{\max} \le 1\text{)}
\end{aligned}
$$

which leads to the conclusion $\beta^l \le \|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l}$. We thus detract from $\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l}$ both sides of (A.20) and obtain

$$
\begin{aligned}
\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l} - \beta^l &\ge (1-\mu_{\max}) \times \left(\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l} - 0.5\beta^l\right) \\
&\ge (1-\mu_{\max}) \cdot 0.5\beta^l,
\end{aligned}
$$

which yields

$$
\begin{aligned}
1 - \mu_{\max} &\le \frac{\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l} - \beta^l}{0.5\beta^l} \\
&= 2\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l}/\beta^l - 2.
\end{aligned} \tag{A.21}
$$

Step 3, we now set out to bound $\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l}$, aiming to prove it is close to β^l. In preparation, we first list the following relations between the variables based on the expressions given in Lemma A.1: $\boldsymbol{\Sigma}_{t+1}^l = \boldsymbol{\Sigma}_t^l + 2^{l_t - l}\boldsymbol{\phi}_0\boldsymbol{\phi}_0^\top$, $\mathbf{b}_{t+1}^l = \mathbf{b}_t^l + 2^{l_t-l}\boldsymbol{\phi}_0 V_0$, where $\boldsymbol{\phi}_0 := \boldsymbol{\phi}_{V_t}(s_t, a_t)$, $V_0 := V_t(s_{t+1})$. We also have $\widehat{\boldsymbol{\theta}}_\tau^l = (\boldsymbol{\Sigma}_\tau^l)^{-1}\mathbf{b}_\tau^l$, $\tau = t, t+1$. Now we break up the value using the triangle inequality:

$$
\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l} \le \|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l\|_{\boldsymbol{\Sigma}_{t+1}^l} + \|\widehat{\boldsymbol{\theta}}_t^l - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\boldsymbol{\Sigma}_{t+1}^l}. \tag{A.22}
$$

We expand the first norm in order to use the relation $\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l\|_{\boldsymbol{\Sigma}_t^l} \le \beta^l$ for an upper bound, specifically:

$$
\begin{aligned}
\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l\|_{\boldsymbol{\Sigma}_{t+1}^l}^2 &= (\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l)^\top \boldsymbol{\Sigma}_{t+1}^l (\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l) \\
&= (\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l)^\top (\boldsymbol{\Sigma}_t^l + 2^{l_t-l}\boldsymbol{\phi}_0\boldsymbol{\phi}_0^\top)(\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l) \\
&= (\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l)^\top \boldsymbol{\Sigma}_t^l (\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l) + 2^{l_t-l}\left[\boldsymbol{\phi}_0^\top(\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l)\right]^2 \\
&\le \|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l\|_{\boldsymbol{\Sigma}_t^l}^2 + 2^{l_t-l}\left[\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_t^l\|_{\boldsymbol{\Sigma}_t^l} \cdot \|\boldsymbol{\phi}_0\|_{(\boldsymbol{\Sigma}_t^l)^{-1}}\right]^2 \qquad \text{(Cauchy-Schwartz)} \\
&\le (\beta^l)^2 + 2^{l_t-l}(\beta^l)^2\|\boldsymbol{\phi}_0\|_{(\boldsymbol{\Sigma}_t^l)^{-1}}^2.
\end{aligned} \tag{A.23}
$$

For the second norm, the vector is the difference between two confidence set centers. We tackle this difference by extracting $(\mathbf{\Sigma}_{t+1}^l)^{-1}$ from both centers:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_t^l - \widehat{\boldsymbol{\theta}}_{t+1}^l &= (\mathbf{\Sigma}_t^l)^{-1}\mathbf{b}_t^l - (\mathbf{\Sigma}_{t+1}^l)^{-1}\mathbf{b}_{t+1}^l \\
&= (\mathbf{\Sigma}_{t+1}^l)^{-1}\big[\mathbf{\Sigma}_{t+1}^l(\mathbf{\Sigma}_t^l)^{-1}\mathbf{b}_t^l - \mathbf{b}_{t+1}^l\big] \\
&= (\mathbf{\Sigma}_{t+1}^l)^{-1}\big[(\mathbf{\Sigma}_t^l + 2^{l_t-l}\boldsymbol{\phi}_0\boldsymbol{\phi}_0^\top)(\mathbf{\Sigma}_t^l)^{-1}\mathbf{b}_t^l - (\mathbf{b}_t^l + 2^{l_t-l}\boldsymbol{\phi}_0 V_0)\big] \\
&= (\mathbf{\Sigma}_{t+1}^l)^{-1}\big[\mathbf{b}_t^l + 2^{l_t-l}\big(\boldsymbol{\phi}_0^\top(\mathbf{\Sigma}_t^l)^{-1}\mathbf{b}_t^l\big)\boldsymbol{\phi}_0 - \mathbf{b}_t^l - 2^{l_t-l}V_0\boldsymbol{\phi}_0\big] \\
&= 2^{l_t-l}\big(\boldsymbol{\phi}_0^\top(\mathbf{\Sigma}_t^l)^{-1}\mathbf{b}_t^l - V_0\big)(\mathbf{\Sigma}_{t+1}^l)^{-1}\boldsymbol{\phi}_0. \qquad \text{(A.24)}
\end{aligned}$$

Substitute the results (A.23) and (A.24) back into (A.22):

$$\begin{aligned}
\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_{t+1}^l\|_{\mathbf{\Sigma}_{t+1}^l} &\le \sqrt{(\beta^l)^2 + 2^{l_t-l}(\beta^l)^2\|\boldsymbol{\phi}_0\|^2_{(\mathbf{\Sigma}_t^l)^{-1}}} + \|2^{l_t-l}\big(\boldsymbol{\phi}_0^\top(\mathbf{\Sigma}_t^l)^{-1}\mathbf{b}_t^l - V_0\big)(\mathbf{\Sigma}_{t+1}^l)^{-1}\boldsymbol{\phi}_0\|_{\mathbf{\Sigma}_{t+1}^l} \\
&= \beta^l\sqrt{1 + 2^{l_t-l}\|\boldsymbol{\phi}_0\|^2_{(\mathbf{\Sigma}_t^l)^{-1}}} + 2^{l_t-l}\big|\boldsymbol{\phi}_0^\top(\mathbf{\Sigma}_t^l)^{-1}\mathbf{b}_t^l - V_0\big|\|(\mathbf{\Sigma}_{t+1}^l)^{-1}\boldsymbol{\phi}_0\|_{\mathbf{\Sigma}_{t+1}^l} \\
&\le \beta^l\big(1 + 2^{l_t-l-1}\|\boldsymbol{\phi}_0\|^2_{(\mathbf{\Sigma}_t^l)^{-1}}\big) + 2^{l_t-l}\big|\boldsymbol{\phi}_0^\top\widehat{\boldsymbol{\theta}}_t^l - V_0\big|\|\boldsymbol{\phi}_0\|_{(\mathbf{\Sigma}_{t+1}^l)^{-1}}, \qquad \text{(A.25)}
\end{aligned}$$

where we used the trivial inequality $\sqrt{1+x} \le 1 + x/2$ for any $x \ge 0$ in the last relation.

To tackle the right hand side of (A.25), we need to provide an upper bound for $\|\boldsymbol{\phi}_0\|_{(\mathbf{\Sigma}_\tau^l)^{-1}}$, where $\tau = t, t+1$. First we have trivially that $\mathbf{\Sigma}_{t+1}^l \succeq \mathbf{\Sigma}_t^l$. Furthermore, by the alternative definition of $\mathbf{\Sigma}_t^l$ given in Lemma A.1, we can see based on $l \ge l_t$ that when $l_t > 1$,

$$\begin{aligned}
\mathbf{\Sigma}_t^l &= \lambda\mathbf{I} + \sum_{\tau=1}^{t-1}\mathbb{1}\{l_\tau \le l\}2^{l_\tau - l}\boldsymbol{\phi}_{V_\tau}(s_\tau.a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau.a_\tau)^\top \\
&\succeq \lambda\mathbf{I} + \sum_{\tau=1}^{t-1}\mathbb{1}\{l_\tau \le l_t - 1\}2^{l_\tau - l}\boldsymbol{\phi}_{V_\tau}(s_\tau.a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau.a_\tau)^\top \\
&= \lambda\mathbf{I} + 2^{l_t-l-1}\sum_{\tau=1}^{t-1}\mathbb{1}\{l_\tau \le l_t - 1\}2^{l_\tau - l_t + 1}\boldsymbol{\phi}_{V_\tau}(s_\tau.a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau.a_\tau)^\top \\
&\succeq 2^{l_t-l-1}\mathbf{\Sigma}_t^{l_t-1}.
\end{aligned}$$

The above relations lead to $\|\boldsymbol{\phi}_0\|_{(\mathbf{\Sigma}_{t+1}^l)^{-1}} \le \|\boldsymbol{\phi}_0\|_{(\mathbf{\Sigma}_t^l)^{-1}} \le \|\boldsymbol{\phi}_0\|_{(2^{l_t-l-1}\mathbf{\Sigma}_t^{l_t-1})^{-1}} = 2^{(l-l_t+1)/2}\|\boldsymbol{\phi}_0\|_{(\mathbf{\Sigma}_t^{l_t-1})^{-1}} \le 2^{(l-l_t+1)/2}\cdot 2^{-l_t+1}\sqrt{d}/(1-\gamma) = 2^{(l-3l_t+3)/2}\sqrt{d}/(1-\gamma)$, where the l_t selection rule in Lines 10 to 13 of Algorithm 1 yields the last inequality. Recall our definition of l_0 from Lemma 6.1, we have shown that when $l_t > 1$,

$$\|\boldsymbol{\phi}_0\|_{(\mathbf{\Sigma}_{t+1}^l)^{-1}} \le 2^{(l+2l_0-3l_t+3)/2}. \qquad \text{(A.26)}$$

Note that in the special case where $l_t = 1$, we have trivially $\|\boldsymbol{\phi}_0\|_{(\mathbf{\Sigma}_{t+1}^l)^{-1}} \le \|\boldsymbol{\phi}_0\|_{\lambda^{-1}\mathbf{I}} = \lambda^{-1/2}\|\boldsymbol{\phi}_0\|_2 \le \sqrt{d}/(1-\gamma) = 2^{l_0}$, and (A.26) still holds.

Apart from this, we also need to bound $\big|\boldsymbol{\phi}_0^\top\widehat{\boldsymbol{\theta}}_t^l - V_0\big|$ in (A.25). Since $\mathbb{P}V_t(s_t,a_t) = \langle\boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_t}(s_t,a_t)\rangle$ from (3.3), we have

$$\begin{aligned}
\big|\boldsymbol{\phi}_0^\top\widehat{\boldsymbol{\theta}}_t^l - V_0\big| &= \big|\langle\widehat{\boldsymbol{\theta}}_t^l, \boldsymbol{\phi}_0\rangle - V_t(s_{t+1})\big| \\
&= \big|\langle\widehat{\boldsymbol{\theta}}_t^l, \boldsymbol{\phi}_0\rangle - \langle\boldsymbol{\theta}^*, \boldsymbol{\phi}_0\rangle + \mathbb{P}V_t(s_t,a_t) - V_t(s_{t+1})\big| \\
&\le |\langle\widehat{\boldsymbol{\theta}}_t^l - \boldsymbol{\theta}^*, \boldsymbol{\phi}_0\rangle| + \big|\mathbb{P}V_t(s_t,a_t) - V_t(s_{t+1})\big| && \text{(triangle inequality)} \\
&\le \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t^l\|_{\mathbf{\Sigma}_t^l}\cdot\|\boldsymbol{\phi}_0\|_{(\mathbf{\Sigma}_t^l)^{-1}} + \big|\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}V_t(s') - V_t(s_{t+1})\big| && \text{(Cauchy-Schwartz)} \\
&\le \frac{1}{2}\beta^l\|\boldsymbol{\phi}_0\|_{(\mathbf{\Sigma}_t^l)^{-1}} + \frac{1}{1-\gamma}, \qquad \text{(A.27)}
\end{aligned}$$

where in the last inequality we used the conclusion of Lemma 6.2 and $0 \le V_t(s) \le 1/(1-\gamma)$.

Now we turn back to (A.25). Implementing the inequalities (A.26) and (A.27) in the right hand side, we obtain

$$\begin{aligned}
\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}^l_{t+1}\|_{\boldsymbol{\Sigma}^l_{t+1}} &\le \beta^l\big(1 + 2^{l_t-l-1}\|\boldsymbol{\phi}_0\|^2_{(\boldsymbol{\Sigma}^l_t)^{-1}}\big) + 2^{l_t-l}\bigg(\frac{1}{2}\beta^l\|\boldsymbol{\phi}_0\|_{(\boldsymbol{\Sigma}^l_t)^{-1}} + \frac{1}{1-\gamma}\bigg)\|\boldsymbol{\phi}_0\|_{(\boldsymbol{\Sigma}^l_t)^{-1}} \\
&= \beta^l + 2^{l_t-l}\bigg(\frac{1}{2}\beta^l\|\boldsymbol{\phi}_0\|^2_{(\boldsymbol{\Sigma}^l_t)^{-1}} + \frac{1}{2}\beta^l\|\boldsymbol{\phi}_0\|^2_{(\boldsymbol{\Sigma}^l_t)^{-1}} + \frac{1}{1-\gamma}\|\boldsymbol{\phi}_0\|_{(\boldsymbol{\Sigma}^l_t)^{-1}}\bigg) \\
&\le \beta^l + 2^{l_t-l}\bigg(\beta^l 2^{l+2l_0-3l_t+3} + \frac{1}{1-\gamma}2^{(l+2l_0-3l_t+3)/2}\bigg) \\
&= \beta^l\bigg(1 + 2^{2l_0-2l_t+3} + \big(1/(1-\gamma)\beta^l\big)\cdot 2^{l_0-(l+l_t-3)/2}\bigg) \\
&\le \beta^l\bigg(1 + 2^{2l_0-2l_t+3} + (1/6\sqrt{d})\cdot 2^{l_0-l_t+3/2}\sqrt{d}\bigg) \\
&\le \beta^l\bigg(1 + 2^{2l_0-2l_t+3} + 2^{l_0-l_t}/2\sqrt{d}\bigg),
\end{aligned} \tag{A.28}$$

where the third inequality is due to $l \ge l_t$ and the definition of β^l from (5.1), which leads to $(1-\gamma)\beta^l \ge 6\sqrt{dl} \ge 6\sqrt{d}$.

Step 4, we are putting everything together. Substituting (A.28) into (A.21), and then into (A.18) together with (A.19), we have

$$\begin{aligned}
\max_{\boldsymbol{\theta}_1\in\mathcal{C}^l_t\cap\mathcal{B}}\langle\boldsymbol{\theta}_1,\boldsymbol{\psi}\rangle - \max_{\boldsymbol{\theta}_2\in\mathcal{C}^l_{t+1}\cap\mathcal{B}}\langle\boldsymbol{\theta}_2,\boldsymbol{\psi}\rangle &\le (1-\mu_{\max})\langle\widetilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*,\boldsymbol{\psi}\rangle \\
&\le \big(2\|\widetilde{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}^l_{t+1}\|_{\boldsymbol{\Sigma}^l_{t+1}}/\beta^l - 2\big)\cdot\frac{1}{1-\gamma} \\
&\le \big(2^{2l_0-2l_t+4} + 2^{l_0-l_t+1.5}/3\sqrt{d}\big)\cdot\frac{1}{1-\gamma}.
\end{aligned}$$

Now substitute the above inequality into (A.17) to get $D_{u-1} \le \gamma D_u + \big(\gamma/(1-\gamma)\big)\cdot\big[2^{2l_0-2l_t+4} + 2^{l_0-l_t+1.5}/3\sqrt{d}\big]$. Recalling from (5.1) that $U_{t+1} \ge U_t$, and from (A.5) that $\gamma^{U_t} \le 1/\big(t(t+1)\big)$, along with the trivial bound for $D_{U_t} \le 1/(1-\gamma)$, we iterate this inequality for D_u over $u = 1,\dots,U_t$ to obtain:

$$\begin{aligned}
D_0 &\le \gamma^{U_t}D_{U_t} + \sum_{u=1}^{U_t}\frac{\gamma^u}{1-\gamma}\cdot\big[2^{2l_0-2l_t+4} + 2^{l_0-l_t+1.5}/3\sqrt{d}\big] \\
&\le \frac{1}{t(t+1)}\cdot\frac{1}{1-\gamma} + \frac{1}{1-\gamma}\cdot\big[2^{2l_0-2l_t+4} + 2^{l_0-l_t+1.5}/3\sqrt{d}\big]\sum_{u=1}^{U_t}\gamma^u \\
&\le \frac{1}{(1-\gamma)t(t+1)} + \frac{\gamma}{(1-\gamma)^2}\big[2^{2l_0-2l_t+4} + 2^{l_0-l_t+1.5}/3\sqrt{d}\big].
\end{aligned}$$

Combining this with $V_t(s_{t+1}) - V_{t+1}(s_{t+1}) = V_t^{(U_t)}(s_t) - V_{t+1}^{(U_{t+1})}(s_t) \le D_0$, we can now bound the sum of C_t by

$$\begin{aligned}
\sum_{t\in\mathcal{T}_\epsilon}\mathbb{1}\{\mathcal{Q}_t\}C_t &= \sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\mathbb{1}\{\mathcal{Q}_t\}\gamma^{\tau-t+1}(V_\tau - V_{\tau+1})(s_{\tau+1}) \\
&\le \sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\mathbb{1}\{\mathcal{Q}_t\}\gamma^{\tau-t+1}\bigg[\frac{1}{(1-\gamma)\tau(\tau+1)} + \frac{\gamma}{(1-\gamma)^2}\big[2^{2l_0-2l_\tau+4} + 2^{l_0-l_\tau+1.5}/3\sqrt{d}\big]\bigg] \\
&\le \frac{\gamma}{1-\gamma}\sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\frac{\gamma^{\tau-t}}{\tau(\tau+1)} + \sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\mathbb{1}\{l_\tau > l_0\}\frac{\gamma^{\tau-t+2}}{(1-\gamma)^2}\big[2^{2l_0-2l_t+4} + 2^{l_0-l_t+1.5}/3\sqrt{d}\big] \\
&\le \frac{\gamma}{(1-\gamma)^2} + \sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\mathbb{1}\{l_\tau > l_0\}\frac{\gamma^{\tau-t+2}}{(1-\gamma)^2}\big[2^{2l_0-2l_t+4} + 2^{l_0-l_t+1.5}/3\sqrt{d}\big],
\end{aligned} \tag{A.29}$$

where in the third inequality we relaxed the indicator function $\mathbb{1}\{\mathcal{Q}_t\}$ to 1 and $\mathbb{1}\{l_\tau > l_0\}$ for the two summations respectively. The final inequality is due to (A.10).

Part IV: Final Arrangements

We now combine the three bounds for the sums of A_t, B_t, C_t, namely (A.11), (A.12) and (A.29), and substitute them into (A.8) to obtain

$$\begin{aligned}
\epsilon N_\epsilon &\le \frac{\gamma^T}{1-\gamma} N_\epsilon + \left[\frac{1}{1-\gamma} + 3\sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1} \mathbb{1}\{l_\tau > l_0\}\gamma^{\tau-t+1} 2^{l_0-l_\tau}\beta^{l_\tau-1}\right] + \frac{2\gamma}{(1-\gamma)^2}\sqrt{2N_\epsilon \log\left(2T(T+1)/\delta\right)} \\
&\quad + \left[\frac{\gamma}{(1-\gamma)^2} + \sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1} \mathbb{1}\{l_\tau > l_0\}\frac{\gamma^{\tau-t+2}}{(1-\gamma)^2}\left[2^{2l_0-2l_t+4} + 2^{l_0-l_t+1.5}/3\sqrt{d}\right]\right] + \frac{10dT}{1-\gamma} l_0(l_0+1) \\
&= \frac{\gamma^T}{1-\gamma} N_\epsilon + \frac{2\gamma}{(1-\gamma)^2}\sqrt{2N_\epsilon \log\left(2T(T+1)/\delta\right)} + \frac{1}{(1-\gamma)^2} + \frac{10dT}{1-\gamma} l_0(l_0+1) \\
&\quad + \sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1} \mathbb{1}\{l_\tau > l_0\}\gamma^{\tau-t}\left[3\gamma 2^{l_0-l_\tau}\beta^{l_\tau-1} + \frac{16\gamma^3}{(1-\gamma)^2}\cdot 2^{2l_0-2l_t} + \frac{2\sqrt{2}\gamma^2}{3(1-\gamma)^2\sqrt{d}}\cdot 2^{l_0-l_t}\right]. \qquad \text{(A.30)}
\end{aligned}$$

We will now focus on the second half of (A.30). More generally, we present an upper bound for the term $\sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\mathbb{1}\{l_\tau > l_0\}\gamma^{\tau-t}\eta^{l_0-l_\tau} f(l_\tau - 1)$ in the following, where $f:\mathbb{R}_+\to\mathbb{R}$ is of the form $f(x) = c_1\sqrt{x} + c_2$, $c_1, c_2 \ge 0$, and $2 \le \eta \le 4$. We do this by using indicator functions to break up the summation, namely:

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1} \mathbb{1}\{l_\tau > l_0\}\gamma^{\tau-t}\eta^{l_0-l_\tau} f(l_\tau - 1) \\
&= \sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\sum_{l=\lfloor l_0\rfloor+1}^{\infty} \gamma^{\tau-t}\eta^{l_0-l} f(l-1)\,\mathbb{1}\{l = l_\tau\} \\
&= \sum_{n=0}^{T-1}\sum_{\tau=1+n}^{\infty}\sum_{l=\lfloor l_0\rfloor+1}^{\infty} \gamma^{n}\eta^{l_0-l} f(l-1)\,\mathbb{1}\{l = l_\tau\}\,\mathbb{1}\{\tau - n \in \mathcal{T}_\epsilon\} \\
&= \sum_{n=0}^{T-1}\gamma^n \sum_{l=\lfloor l_0\rfloor+1}^{\infty} \eta^{l_0-l} f(l-1) \sum_{\tau=1+n}^{\infty} \mathbb{1}\{l = l_\tau\}\,\mathbb{1}\{\tau - n \in \mathcal{T}_\epsilon\} \\
&\le \sum_{n=0}^{T-1}\gamma^n \sum_{l=\lfloor l_0\rfloor+1}^{\infty} \eta^{l_0-l} f(l-1)\min\left\{\sum_{\tau=1+n}^{\infty}\mathbb{1}\{l = l_\tau\}, \sum_{\tau=1+n}^{\infty}\mathbb{1}\{\tau - n \in \mathcal{T}_\epsilon\}\right\} \\
&\le \sum_{n=0}^{T-1}\gamma^n \sum_{l=\lfloor l_0\rfloor+1}^{\infty} \eta^{l_0-l} f(l-1)\min\{k^l_{\max}, N_\epsilon\} \\
&\le \frac{1}{1-\gamma}\sum_{l=\lfloor l_0\rfloor+1}^{\infty} \eta^{l_0-l} f(l-1)\min\{20dl4^{l-l_0}, N_\epsilon\}, \qquad \text{(A.31)}
\end{aligned}$$

where: in the first equation we expanded $\mathbb{1}\{l_\tau > l_0\}$ into $\sum_{l>l_0}\mathbb{1}\{l = l_\tau\}$; in the second equation we used the variable $n = \tau - t$ to replace t, which adds a multiplier $\mathbb{1}\{t := \tau - n \in \mathcal{T}_\epsilon\}$ to the expression under summation; in the second to last inequality we used the definition of $k^l_{\max}$ from Lemma 6.1 and the definition $N_\epsilon = \#\mathcal{T}_\epsilon$; the last inequality is due to Lemma 6.1.

Ideally, we can determine an exact threshold integer value $L_0' = \max\{l > l_0 : 20dl4^{l-l_0} \le N_\epsilon\}$, and take the correct minimization in (A.31) for all $l \in \mathbb{Z}$, yet this will intensely complicate our deductions. Instead, we will set a undetermined threshold value L_0, under which the minimization takes the upper bound $20dl4^{l-l_0}$, and above which the minimization takes

the upper bound N_ϵ:

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\mathbb{1}\{l_\tau > l_0\}\gamma^{\tau-t}\eta^{l_0-l_\tau}f(l_\tau-1)\\
&\le \frac{1}{1-\gamma}\sum_{l=\lfloor l_0\rfloor+1}^{L_0}\eta^{l_0-l}f(l-1)\cdot 20dl4^{l-l_0}+\frac{1}{1-\gamma}\sum_{l=L_0+1}^{\infty}\eta^{l_0-l}f(l-1)N_\epsilon\\
&\le \frac{1}{1-\gamma}\cdot f(L_0-1)\cdot 20dL_0\sum_{l=\lfloor l_0\rfloor+1}^{L_0}\eta^{l_0-l}4^{l-l_0}+\frac{N_\epsilon}{1-\gamma}\sum_{l=L_0+1}^{\infty}\eta^{l_0-l}f(l-1)\\
&= \frac{20d}{1-\gamma}L_0f(L_0-1)\sum_{l=\lfloor l_0\rfloor+1}^{L_0}(4/\eta)^{l-l_0}+\frac{N_\epsilon}{1-\gamma}\sum_{l=L_0+1}^{\infty}\eta^{l_0-l}\big(c_1\sqrt{l-1}+c_2\big),
\end{aligned}\tag{A.32}$$

where the second inequality holds because $f(l-1)\cdot 20dl$ is increasing in l, and in the equation we plugged in $f(x)=c_1\sqrt{x}+c_2$.

To tackle the second half of (A.32), gather from the first conclusion of Lemma D.3 that $\sum_{l=L}^{\infty}\eta^{l_0-l}\sqrt{l}\le\sqrt{L+1}\eta^{l_0-L+1}/(\eta-1)$, which leads to the upper bound

$$\begin{aligned}
\sum_{l=L_0+1}^{\infty}\eta^{l_0-l}\big(c_1\sqrt{l-1}+c_2\big) &= c_2\sum_{l=L_0+1}^{\infty}\eta^{l_0-l}+c_1\sum_{l=L_0}^{\infty}\eta^{l_0-l-1}\sqrt{l}\\
&\le \big(c_2+c_1\sqrt{L_0+1}\big)\eta^{l_0-L_0}/(\eta-1)\\
&= f(L_0+1)\eta^{l_0-L_0}/(\eta-1).
\end{aligned}\tag{A.33}$$

Substituting (A.33) into (A.32), we have

$$\begin{aligned}
\sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\gamma^{\tau-t}\eta^{l_0-l_\tau}f(l_\tau-1) &\le \frac{20d}{1-\gamma}L_0f(L_0-1)\sum_{l=\lfloor l_0\rfloor+1}^{L_0}(4/\eta)^{l-l_0}+\frac{N_\epsilon}{1-\gamma}f(L_0+1)\eta^{l_0-L_0}/(\eta-1)\\
&\le \frac{f(L_0+1)}{1-\gamma}\bigg[20dL_0\sum_{l=\lfloor l_0\rfloor+1}^{L_0}(4/\eta)^{l-l_0}+N_\epsilon\eta^{l_0-L_0}/(\eta-1)\bigg],
\end{aligned}\tag{A.34}$$

where for the second inequality, we used the fact $f(L_0-1)\le f(L_0+1)$ due to $c_1\ge 0$.

Now we go back to consider the second half of (A.30). With $\beta^l=2\big(3\sqrt{dl}+2\sqrt{\log(1/\delta)}\big)/(1-\gamma)+2\sqrt{d\lambda}$, we can take $\eta=2,4$ in (A.34) to get

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\mathbb{1}\{l_\tau > l_0\}\gamma^{\tau-t}\bigg[3\gamma 2^{l_0-l_\tau}\beta^{l_\tau-1}+\frac{16\gamma^3}{(1-\gamma)^2}\cdot 2^{2l_0-2l_t}+\frac{2\sqrt{2}\gamma^2}{3(1-\gamma)^2\sqrt{d}}\cdot 2^{l_0-l_t}\bigg]\\
&=\sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\mathbb{1}\{l_\tau > l_0\}\gamma^{\tau-t}2^{l_0-l_\tau}\bigg[3\gamma\beta^{l_\tau-1}+\frac{2\sqrt{2}\gamma^2}{3(1-\gamma)^2\sqrt{d}}\bigg]+\sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\mathbb{1}\{l_\tau > l_0\}\gamma^{\tau-t}4^{l_0-l_\tau}\cdot\frac{16\gamma^3}{(1-\gamma)^2}\\
&\le \frac{1}{1-\gamma}\bigg(3\gamma\beta^{L_0+1}+\frac{2\sqrt{2}\gamma^2}{3(1-\gamma)^2\sqrt{d}}\bigg)\bigg(20dL_0\sum_{l=\lfloor l_0\rfloor+1}^{L_0}2^{l-l_0}+N_\epsilon 2^{l_0-L_0}\bigg)\\
&\quad+\frac{16\gamma^3}{(1-\gamma)^3}\bigg(20dL_0\sum_{l=\lfloor l_0\rfloor+1}^{L_0}1+N_\epsilon 4^{l_0-L_0}/3\bigg)\\
&\le \bigg(\frac{3\beta^{L_0+1}}{1-\gamma}2^{l_0-L_0}+\frac{2\sqrt{2}}{3(1-\gamma)^3\sqrt{d}}2^{l_0-L_0}+\frac{16}{3(1-\gamma)^3}4^{l_0-L_0}\bigg)N_\epsilon\\
&\quad+\bigg(\frac{3\beta^{L_0+1}}{1-\gamma}2^{L_0-l_0+1}+\frac{2\sqrt{2}}{3(1-\gamma)^3\sqrt{d}}2^{L_0-l_0+1}+\frac{16}{(1-\gamma)^3}L_0\bigg)\cdot 20dL_0,
\end{aligned}\tag{A.35}$$

where in the last inequality we we used $\gamma \leq 1$ to get rid of all the excess γ's.

Now take

$$T := \left\lceil \log \frac{3}{(1-\gamma)\epsilon} \Big/ (1-\gamma) \right\rceil,$$

and

$$L_0 := \left\lceil l_0 + \log\left(M + K\sqrt{2\log(M+K)/\log 2 + 2(l_0+1)} \right) \Big/ \log 2 \right\rceil,$$

where

$$\begin{aligned} M &= \frac{3}{\epsilon} \cdot \left[\frac{1}{(1-\gamma)^3\sqrt{d}} + \frac{12\sqrt{\log(1/\delta)}}{(1-\gamma)^2} + \frac{6\sqrt{d\lambda}}{1-\gamma} \right], \\ K &= \frac{54\sqrt{d}}{(1-\gamma)^2\epsilon}. \end{aligned}$$

It is not hard to check that

$$L_0 = O\left(\log \frac{d\log(1/\delta)}{(1-\gamma)} \right) + O\left(\log \frac{1}{\epsilon} \right),$$

where we separated the term $\log(1/\epsilon)$ since the maximum possible value of $\epsilon \leq 1-\gamma$ is dependent on the parameter γ. In the third conclusion of Lemma D.3, by taking $a = 2^{l_0+1}M$ and $b = 2^{l_0+1}K$, we have $L_0 + 1 = \lceil \log(a + b\sqrt{2\log(a+b)/\log 2})/\log 2 \rceil$, and hence

$$\begin{aligned} 2^{L_0-l_0} &\geq 2^{-l_0-1}(a + b\sqrt{L_0+1}) \\ &= M + K\sqrt{L_0+1} \\ &= \frac{3}{\epsilon} \cdot \left[\frac{1}{(1-\gamma)^3\sqrt{d}} + \frac{12\sqrt{\log(1/\delta)}}{(1-\gamma)^2} + \frac{6\sqrt{d\lambda}}{1-\gamma} + \frac{18\sqrt{d}}{(1-\gamma)^2}\sqrt{L_0+1} \right] \\ &= \frac{3}{\epsilon} \cdot \left[\frac{1}{(1-\gamma)^3\sqrt{d}} + \frac{3\beta^{L_0+1}}{1-\gamma} \right], \end{aligned}$$

where the final equation is due to the definition of β^l in (5.1). As a side effect of the second to last expression above, we have the relation $2^{L_0-l_0} \geq 3(6+18\sqrt{2})\sqrt{d}$ due to $\epsilon \leq 1/(1-\gamma)$, $\lambda \geq 1$ and $L_0 \geq 1$. On the other hand, we have an upper bound of similar composition:

$$\begin{aligned} 2^{L_0-l_0} &\leq 2\left\lceil M + K\sqrt{2\log(M+K)/\log 2 + 2l_0 + 2} \right\rceil \\ &\leq 2\left\lceil M + K\sqrt{2L_0+2} \right\rceil \\ &= \frac{6}{\epsilon} \cdot \left[\frac{1}{(1-\gamma)^3\sqrt{d}} + \frac{12\sqrt{\log(1/\delta)}}{(1-\gamma)^2} + \frac{6\sqrt{d\lambda}}{1-\gamma} + \frac{18\sqrt{2d}}{(1-\gamma)^2}\sqrt{L_0+1} \right], \end{aligned}$$

where in the second inequality we used the relation $L_0 \geq \log(M+K)/\log 2$.

We now substitute the two inequalities above into (A.35):

$$
\begin{aligned}
&\sum_{t\in\mathcal{T}_\epsilon}\sum_{\tau=t}^{t+T-1}\mathbb{1}\{l_\tau>l_0\}\gamma^{\tau-t}\left[3\gamma 2^{l_0-l_\tau}\beta^{l_\tau-1}+\frac{16\gamma^3}{(1-\gamma)^2}\cdot 2^{2l_0-2l_t}+\frac{2\sqrt{2}\gamma^2}{3(1-\gamma)^2\sqrt{d}}\cdot 2^{l_0-l_t}\right]\\
&\le\left(\frac{3\beta^{L_0+1}}{1-\gamma}+\frac{2\sqrt{2}}{3(1-\gamma)^3\sqrt{d}}+\frac{16}{3(1-\gamma)^3}\cdot\frac{1}{3(6+18\sqrt{2})\sqrt{d}}\right)2^{l_0-L_0}N_\epsilon\\
&\quad+\left(\frac{3\beta^{L_0+1}}{1-\gamma}+\frac{2}{3(1-\gamma)^3\sqrt{d}}\right)\cdot 20dL_0 2^{L_0-l_0}+\frac{320dL_0^2}{(1-\gamma)^3}\\
&\le\frac{\epsilon}{3}N_\epsilon+20dL_0\cdot\frac{6}{\epsilon}\left[\frac{1}{(1-\gamma)^3\sqrt{d}}+\frac{12\sqrt{\log(1/\delta)}}{(1-\gamma)^2}+\frac{6\sqrt{d\lambda}}{1-\gamma}+\frac{18\sqrt{2d}}{(1-\gamma)^2}\sqrt{L_0+1}\right]\\
&\quad\times\left(\frac{2\sqrt{2}}{3(1-\gamma)^3\sqrt{d}}+\frac{12\sqrt{\log(1/\delta)}}{(1-\gamma)^2}+\frac{6\sqrt{d\lambda}}{1-\gamma}+\frac{18\sqrt{d}}{(1-\gamma)^2}\sqrt{L_0+1}\right)+\frac{320dL_0^2}{(1-\gamma)^3}\\
&\le\frac{\epsilon}{3}N_\epsilon+120dL_0/\epsilon\cdot\sqrt{1^2+12^2+6^2+(18\sqrt{2})^2}\cdot\sqrt{(2\sqrt{2}/3)^2+12^2+6^2+18^2}\\
&\quad\times\left[\frac{1}{(1-\gamma)^6 d}+\frac{\log(1/\delta)}{(1-\gamma)^4}+\frac{dL_0}{(1-\gamma)^4}+\frac{d\lambda}{(1-\gamma)^2}\right]+\frac{320dL_0^2}{(1-\gamma)^3}\\
&\le\frac{\epsilon}{3}N_\epsilon+CdL_0/\epsilon\cdot\left[\frac{1}{(1-\gamma)^6 d}+\frac{\log(1/\delta)}{(1-\gamma)^4}+\frac{dL_0}{(1-\gamma)^4}+\frac{d\lambda}{(1-\gamma)^2}\right],
\end{aligned}
\tag{A.36}
$$

where: the second inequality is partly due to $2\sqrt{2}/3+16/9(6+18\sqrt{2})\le 1$; we used the standard Cauchy-Schwartz inequality in the third inequality; for the last expression, C is an absolute constant.

Besides, by the definition of T,

$$
\begin{aligned}
\gamma^T&\le\left(\gamma^{1/(1-\gamma)}\right)^{\log[3/(1-\gamma)\epsilon]}\\
&\le\exp\left[-\log\left(3/(1-\gamma)\epsilon\right)\right]\\
&\le(1-\gamma)\epsilon/3.
\end{aligned}
$$

We then bring this and (A.36) back to (A.30):

$$
\begin{aligned}
\epsilon N_\epsilon\le\frac{\epsilon}{3}N_\epsilon+\frac{2\gamma}{(1-\gamma)^2}\sqrt{2N_\epsilon\log\left(2T(T+1)/\delta\right)}+\frac{1}{(1-\gamma)^2}+\frac{10dT}{1-\gamma}l_0(l_0+1)\\
+\frac{\epsilon}{3}N_\epsilon+\frac{CdL_0}{\epsilon}\left[\frac{1}{(1-\gamma)^6 d}+\frac{\log(1/\delta)}{(1-\gamma)^4}+\frac{dL_0}{(1-\gamma)^4}+\frac{d\lambda}{(1-\gamma)^2}\right].
\end{aligned}
$$

We go on to move the two terms $(\epsilon/3)N_\epsilon$ to the left hand side, then multiply both sides by the factor $3/\epsilon$, thus transforming the above inequality into (after erasing yet another γ from the first term):

$$
\begin{aligned}
N_\epsilon\le\frac{6}{(1-\gamma)^2\epsilon}\sqrt{2\log\left(2T(T+1)/\delta\right)}\sqrt{N_\epsilon}+\frac{3}{(1-\gamma)^2\epsilon}+\frac{30dT}{(1-\gamma)\epsilon}l_0(l_0+1)\\
+\frac{3C}{\epsilon^2}\cdot\left[\frac{L_0}{(1-\gamma)^6}+\frac{d\log(1/\delta)L_0+d^2L_0^2}{(1-\gamma)^4}+\frac{d^2\lambda L_0}{(1-\gamma)^2}\right].
\end{aligned}
$$

Using the second conclusion in Lemma D.3, which states that $x-a\sqrt{x}\le b\Rightarrow x\le a^2+2b$ for positive real numbers a,b,

we see the above inequality implies

$$\begin{aligned}
N_\epsilon &\le \left[\frac{6}{(1-\gamma)^2\epsilon}\sqrt{2\log\big(2T(T+1)/\delta\big)}\right]^2 + \frac{6}{(1-\gamma)^2\epsilon} + \frac{60dT}{(1-\gamma)\epsilon}l_0(l_0+1) \\
&\quad + \frac{6C}{\epsilon^2}\cdot\left[\frac{L_0}{(1-\gamma)^6} + \frac{d\log(1/\delta)L_0 + d^2L_0^2}{(1-\gamma)^4} + \frac{d^2\lambda L_0}{(1-\gamma)^2}\right] \\
&= O\bigg(\frac{L_0}{(1-\gamma)^6\epsilon^2} + \frac{d\log(1/\delta)L_0 + d^2L_0^2 + \log T}{(1-\gamma)^4\epsilon^2} + \frac{d^2\lambda L_0}{(1-\gamma)^2\epsilon^2} + \frac{dl_0^2\log\big(1/(1-\gamma)\epsilon\big)}{(1-\gamma)^2\epsilon}\bigg) \\
&= \widetilde{O}\bigg(\frac{1}{(1-\gamma)^6\epsilon^2} + \frac{d^2 + d\log(1/\delta)}{(1-\gamma)^4\epsilon^2}\bigg),
\end{aligned}$$

which concludes the proof. □

A.2. Proof of Corollary 5.3

Proof. Theorem 3 of Dann et al. (2017) states that if an algorithm is uniform-PAC with sample complexity $\Gamma(\epsilon,\delta) = \widetilde{O}(C_1/\epsilon + C_2/\epsilon^2)$ for some $\delta > 0$, then with probability at least $1-\delta$, its regret is bounded by $\widetilde{O}(\sqrt{C_2 T} + \max\{C_1, C_2\})$. Taking $C_1 = 0$, $C_2 = 1/(1-\gamma)^3 + \big(d + \sqrt{d\log(1/\delta)}\big)/(1-\gamma)^2$, and discarding the non-dominating terms without $\sqrt{T}$, this suggests that our algorithm has a high probability regret bound of

$$\text{Regret}(T) = \widetilde{O}\bigg(\frac{d + \sqrt{d\log(1/\delta)}}{(1-\gamma)^2}\sqrt{T} + \frac{\sqrt{T}}{(1-\gamma)^3}\bigg). \tag{A.37}$$

□

B. Proof of Lemmas in Section 6

In this section, we focus on the lemmas presented in Section 6 and give detailed proofs for each of them.

B.1. Proof of Lemma 6.1

Proof of Lemma 6.1. By Definition 3.1, we have $\|\boldsymbol{\phi}_{V_t}(s_t,a_t)\|_2 \le \sqrt{d}\cdot 1/(1-\gamma)$. Take $\{X_t\} = \{2^{(l_\tau - l)/2}\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau) : \tau \ge 1, l_\tau \le l\}$, $V = \lambda\mathbf{I}$ and $L = \sqrt{d}/(1-\gamma)$ which, considering Lemma A.1, means that

$$\begin{aligned}
\overline{V}_{t'} &= \lambda\mathbf{I} + \sum_{\tau=1}^{t}\mathbb{1}\{l_\tau \le l\}2^{l_\tau - l}\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)^\top \\
&= \boldsymbol{\Sigma}_{t+1}^l,
\end{aligned}$$

where $t' = \sum_{\tau=1}^{t}\mathbb{1}\{l_\tau \le l\}$. We can thus deduce from Lemma 11 of Abbasi-Yadkori et al. (2011) the following:

$$\sum_{\tau=1}^{t-1}\mathbb{1}\{l_\tau \le l\}\min\big\{2^{l_\tau - l}\|\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)\|^2_{(\boldsymbol{\Sigma}_\tau^l)^{-1}}, 1\big\} \le 2\log\frac{\det(\boldsymbol{\Sigma}_t^l)}{\det(\lambda\mathbf{I})}. \tag{B.1}$$

On one hand, from Lines 10 to 13 of algorithm 1, we have when $l_t = l$ and $l_t \ne L_t + 1$, $\|\boldsymbol{\phi}_{V_t}(s_t,a_t)\|_{(\boldsymbol{\Sigma}_t^l)^{-1}} \ge 2^{-l}\sqrt{d}/(1-\gamma)$. Combine this with the definition $k_t^l = \sum_{\tau=1}^{t-1}\mathbb{1}\{l_\tau \le l\}2^{l_\tau - l}$ and the observation that $l_t = l = L_t + 1$ happens exactly once for each level l, we have that when t is large enough, the left hand side of inequality (B.1) can be lower bounded by

$$\begin{aligned}
\sum_{\tau=1}^{t-1}\mathbb{1}\{l_\tau \le l\}\min\big\{2^{l_\tau - l}\|\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)\|^2_{(\boldsymbol{\Sigma}_\tau^l)^{-1}}, 1\big\} &\ge \sum_{\tau=1}^{t-1}\mathbb{1}\{l_\tau = l, L_\tau + 1 \ne l\}\min\big\{2^{l_\tau - l}\cdot 2^{-2l}d/(1-\gamma)^2, 1\big\} \\
&= \bigg[\sum_{\tau=1}^{t-1}\mathbb{1}\{l_\tau = l, L_\tau + 1 \ne l\}\bigg]\min\big\{4^{-l}d/(1-\gamma)^2, 1\big\} \\
&= \big(k_t^l - \frac{1}{2}k_t^{l-1} - 1\big)\min\big\{4^{l_0 - l}, 1\big\}.
\end{aligned}$$

On the other hand, for the upper bound of the right hand side of (B.1), notice that

$$\begin{aligned}\det(\mathbf{\Sigma}_t^l) &\le \left[\frac{\operatorname{tr}(\mathbf{\Sigma}_t^l)}{d}\right]^d \\ &= \left[\frac{d\lambda + \sum_{\tau=1}^{t-1} \mathbb{1}\{l_\tau \le l\} 2^{l_\tau - l} \operatorname{tr}\left[\boldsymbol{\phi}_{V_\tau}(s_\tau, a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau, a_\tau)^\top\right]}{d}\right]^d \\ &\le \left[\frac{d\lambda + k_t^l \cdot d/(1-\gamma)^2}{d}\right]^d \\ &= \left(\lambda + k_t^l/(1-\gamma)^2\right)^d, \end{aligned} \tag{B.2}$$

where the second inequality holds because $\operatorname{tr}\left[\boldsymbol{\phi}_{V_t}(s_t, a_t)\boldsymbol{\phi}_{V_t}(s_t, a_t)^\top\right] = \|\boldsymbol{\phi}_{V_t}(s_t, a_t)\|_2^2 \le d/(1-\gamma)^2$. Substituting these bounds into (B.1), we have

$$\begin{aligned}\left(k_t^l - \frac{1}{2}k_t^{l-1} - 1\right)\min\left\{4^{l_0 - l}, 1\right\} &\le 2\log\frac{(\lambda + k_t^l/(1-\gamma)^2)^d}{\lambda^d} \\ &\le 2d\log\left(1 + k_t^l/(1-\gamma)^2\right), \end{aligned} \tag{B.3}$$

where in the second inequality we used $\lambda \ge 1$.

The above inequality immediately implies that k_t^l is finite for arbitrary t, l. Take $t \to \infty$, we can replace the suffixes t in (B.3) by ∞. Next we consider two separate cases of l.

First consider the case $l \le l_0$, where (B.3) becomes

$$k_{\max}^l - \frac{1}{2}k_{\max}^{l-1} - 1 \le 2d\log\left(1 + k_{\max}^l/(1-\gamma)^2\right).$$

We will prove through induction on l that

$$k_{\max}^l \le 20dl_0. \tag{B.4}$$

Note that (B.4) holds trivially for $l = 0$. Now suppose (B.4) already holds for $l - 1$. Then

$$k_{\max}^l - 2d\log\left(1 + k_{\max}^l/(1-\gamma)^2\right) \le \frac{1}{2}\cdot 20dl_0 + 1.$$

By viewing $k_{\max}^l$ as a variable and differentiating on it, we see the left hand side is increasing for $k_{\max}^l > 2d$. It is evident from (B.4) that the desired upper bound is indeed greater than $2d$, hence we only need to prove that the above inequality does not hold when $k_{\max}^l$ is above this upper bound, in other words

$$20dl_0 - 2d\log\left(1 + 20dl_0/(1-\gamma)^2\right) \ge 10dl_0 + 1.$$

Substitute $4^{l_0} = d/(1-\gamma)^2$ and move the first term on the right to the left, this is equivalent to

$$10dl_0 - 2d\log\left(1 + 20l_0 \cdot 4^{l_0}\right) \ge 1.$$

The left hand side

$$\begin{aligned} d\left[10l_0 - 2\log\left(1 + 20l_0 \cdot 4^{l_0}\right)\right] &\ge 10l_0 - 2\log(1/4 + 20) - 2\log\left(l_0 \cdot 4^{l_0}\right) \\ &= 10l_0 - 2\log(81/4) - 2\left(2\log 2 \cdot l_0 + \log l_0\right) \\ &= (10 - 4\log 2)l_0 - 2\log l_0 - 2\log(81/4) \\ &\ge 10 - 4\log 2 - 2\log(81/4) \\ &> 1, \end{aligned}$$

where in the second inequality we shrank l_0 to 1 based on monotony. This concludes the induction, and hence we have proven (B.4).

Now for $l > l_0$, the bound (B.3) becomes

$$\big(k_{\max}^l - \frac{1}{2}k_{\max}^{l-1} - 1\big) \cdot 4^{l_0 - l} \le 2d \log\big(1 + k_{\max}^l/(1-\gamma)^2\big). \tag{B.5}$$

Write $k_{\max}^l = \alpha_l dl 4^{l-l_0}$. Notice that from (B.4), $k_{\max}^{\lfloor l_0 \rfloor} \le 20d(\lfloor l_0 \rfloor + 1) \cdot 4^{\lfloor l_0 \rfloor + 1 - l_0}$. We further define $\alpha_{\lfloor l_0 \rfloor} = 80$, then plug these into (B.5), to obtain

$$\alpha_l dl - \frac{1}{8}\alpha_{l-1} dl - 4^{l_0 - l} \le 2d \log(1 + \alpha_l 4^l l).$$

This inequality holds for arbitrary $l > l_0$: for $l = \lfloor l_0 \rfloor + 1$, this is the exact result of substituting the bound for $k_{\max}^{\lfloor l_0 \rfloor}$ into (B.5); for $l > l_0 + 1$, we further expanded $l - 1$ on the left hand side to l to obtain this inequality. We then go on to divide both sides by dl:

$$\alpha_l - \frac{1}{8}\alpha_{l-1} - \frac{1}{dl} \le \frac{2}{l}\log(1 + \alpha_l 4^l l). \tag{B.6}$$

From here we only need to verify through induction that $\alpha_l \le 20$ for $l > l_0$, which yields $k_t^l \le 20dl4^{l-l_0}$ as required. For the induction basis, by shrinking d and l to 1, we have $\alpha_{\lfloor l_0 \rfloor + 1} \le 11 + 2\log(1 + 4\alpha_{\lfloor l_0 \rfloor + 1})$. Because the function $x - 2\log(1 + 4x)$ is increasing in x when $x \ge 2$, and $20 > 11 + 2\log(1 + 4 \times 20)$, we can deduce from here $\alpha_{\lfloor l_0 \rfloor + 1} \le 20$.

Suppose we already have $\alpha_{l-1} \le 20$ in (B.6). We can then obtain $\alpha_l \le 5/2 + 1/l + (2/l)\log(1 + \alpha_l 4^l l)$. Combine with the trivial lower bound $\alpha_l 4^l l \ge k_{\max}^l \ge 1$, we have

$$\begin{aligned}
\alpha_l &\le \frac{5}{2} + \frac{1}{l} + \frac{2}{l}\log\big(2\alpha_l 4^l l\big) \\
&\le \frac{7}{2} + 2\log 2 + 2\log\alpha_l + 2\log 4 + \frac{2}{l}\log l \\
&\le \left[\frac{7}{2} + \log 64 + \frac{2}{3}\log 3\right] + 2\log\alpha_l \\
&< 9 + 2\log\alpha_l.
\end{aligned}$$

Therefore, in light of the fact that $x - 2\log x$ is increasing for $x \ge 2$, and that $20 > 9 + 2\log 20$, we have $\alpha_l \le 20$, which concludes the proof. □

B.2. Proof of Lemma 6.2

Proof of Lemma 6.2. By Lemma A.1, we have the following expression for $\widehat{\boldsymbol{\theta}}_t^l$:

$$\widehat{\boldsymbol{\theta}}_t^l = \left(\lambda\mathbf{I} + \sum_{\tau=1}^{t-1} \mathbb{1}\{l_\tau \le l\} 2^{l_\tau - l} \boldsymbol{\phi}_{V_\tau}(s_\tau.a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau.a_\tau)^\top\right)^{-1} \left(\sum_{\tau=1}^{t-1} \mathbb{1}\{l_\tau \le l\} 2^{l_\tau - l} \boldsymbol{\phi}_{V_\tau}(s_\tau.a_\tau) V_\tau(s_{\tau+1})\right).$$

Next, we have from (3.3) that $\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_t}(s_t, a_t)\rangle = \mathbb{P}V_t(s_t, a_t) = \mathbb{E}_{s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)}[V_t(s_{t+1})]$, which means $\{V_t(s_{t+1}) - \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{V_t}(s_t, a_t)\rangle\}$ forms a martingale difference sequence. Furthermore, since $V_t(s)$ is bounded by $1/(1-\gamma)$, this sequence is $1/(1-\gamma)$-subgaussian. Combine this with the fact $\|\boldsymbol{\theta}^*\| \le \sqrt{d}$ from Definition 3.1, we may deduce from Theorem 2 of Abbasi-Yadkori et al. (2011) that with probability at least $\delta/\big(2l(l+1)\big)$, the following holds for all $t \ge 0$:

$$\begin{aligned}
\|\widehat{\boldsymbol{\theta}}_t^l - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t^l} &\le \frac{1}{1-\gamma}\sqrt{2\log\left[\frac{\det(\boldsymbol{\Sigma}_t^l)^{1/2}\det(\lambda\mathbf{I})^{-1/2}}{\delta/\big(2l(l+1)\big)}\right]} + \sqrt{\lambda}\cdot\sqrt{d} \\
&\le \frac{1}{1-\gamma}\sqrt{2\log\left[\frac{[\lambda + k_t^l/(1-\gamma)^2]^{d/2}}{\lambda^{d/2}}\cdot\frac{2l(l+1)}{\delta}\right]} + \sqrt{\lambda d} \\
&= \frac{1}{1-\gamma}\sqrt{d\log\big(1 + k_t^l/(1-\gamma)^2\lambda\big) + 2\log\big(2l(l+1)/\delta\big)} + \sqrt{\lambda d} \\
&\le \frac{1}{1-\gamma}\sqrt{d\log(1 + 20\cdot 4^{l'}l') + 2\log\big(2l'(l'+1)/\delta\big)} + \sqrt{\lambda d},
\end{aligned} \tag{B.7}$$

where in the second inequality we borrowed (B.2) from the proof of Lemma 6.1, and in the final inequality we used the conclusion of Lemma 6.1 and substituted $l' = \max\{l, l_0\}$.

We now seek to prove the right hand side of (B.7) is no greater than half of β^l. First, for the expression inside the big square root, we have

$$\begin{aligned}
&d\log(1+20\cdot 4^l l) + 2\log\big(2l(l+1)/\delta\big) \\
&\le d\big[\log(1+20\cdot 4^l l) + 2\log\big(2l(l+1)\big)\big] + 2\log(1/\delta) \\
&\le d\big[\log\big((1+1/80)\cdot 20\cdot 4^l l\big) + \log 4 + 2\log l + 2\log(l+1)\big] + 2\log(1/\delta) \\
&= d\big[\log 81 + (\log 4)\cdot l + 3\log l + 2\log\big((l+1)/2\big) + \log 4\big] + 2\log(1/\delta) \\
&\le d\big[\log(81\cdot 4) + (\log 4)\cdot l + 3(l-1) + 2\big((l+1)/2 - 1\big)\big] + 2\log(1/\delta) \\
&= d\big[(4+\log 4)l + (\log 324 - 4)\big] + 2\log(1/\delta) \\
&< 9dl + 2\log(1/\delta),
\end{aligned}$$

where we used the relation $\log x \le x - 1$ for real number $x > 0$ twice in the third inequality.

Second, we take the square root in the above to obtain as a continuation of (B.7) that

$$\begin{aligned}
\|\widehat{\boldsymbol{\theta}}_t^l - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t^l} &\le \frac{1}{1-\gamma}\sqrt{9dl' + 2\log(1/\delta)} + \sqrt{d\lambda} \\
&\le \frac{1}{1-\gamma}\big(3\sqrt{dl'} + 2\sqrt{\log(1/\delta)}\big) + \sqrt{d\lambda} \\
&= \frac{1}{2}\beta^l.
\end{aligned}$$

Now take a uniform bound, with probability at least $1 - \sum_{l=1}^{\infty}\delta/\big(2l(l+1)\big) = 1 - \delta/2$, $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_t^l\|_{\boldsymbol{\Sigma}_t^l} \le \beta^l/2$ holds for all $l \ge 1$, which finishes the proof. □

B.3. Proof of Lemma 6.3

Proof of Lemma 6.3. Focusing on one specific run of the ML-EVI algorithm, we use induction on u to prove that $1/(1-\gamma) \ge Q_l^{(u)}(s,a) \ge Q^*(s,a)$. When $u = 0$, $1/(1-\gamma) = Q_l^{(0)}(s,a) \ge Q^*(s,a)$, $\forall s, a$.

Suppose the conclusion holds for $u-1$. First recall that under the event $\mathcal{E}_1$ in Lemma 6.2, $\boldsymbol{\theta}^* \in \mathcal{B}$ and $\mathcal{B} \cap \mathcal{C}_t^l \neq \varnothing$ for all t, l, so the special update rule in Line 8 is never called. Based on the update rule for V in Line 3 of Algorithm 2, $V^{(u-1)}(s) = \max_a \min_l Q_l^{(u-1)}(s,a)$, which right hand side is clearly upper bounded by $1/(1-\gamma)$ and lower bounded by $\max_a Q^*(s,a) = V^*(s)$. Now combine this with the update rule of Q in Line 6, we see that

$$\begin{aligned}
Q_l^{(u)}(s,a) &= r(s,a) + \gamma \max_{\boldsymbol{\theta}\in\mathcal{B}\cap\mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(u-1)}}(s,a)\rangle \\
&= r(s,a) + \gamma \max_{\boldsymbol{\theta}\in\mathcal{B}\cap\mathcal{C}_l} \int_{\mathcal{S}} V^{(u-1)}(s')\langle \boldsymbol{\theta}, \boldsymbol{\phi}(s'|s,a)\rangle \mathrm{d}s' \\
&\ge r(s,a) + \gamma \int_{\mathcal{S}} V^{(u-1)}(s')\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}(s'|s,a)\rangle \mathrm{d}s' \\
&\ge r(s,a) + \gamma \int_{\mathcal{S}} V^{*}(s')\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}(s'|s,a)\rangle \mathrm{d}s' \\
&= Q^*(s,a),
\end{aligned}$$

where the second equation is based on the expression of $\boldsymbol{\phi}_V$ given in Definition 3.1, and the last equation is the Bellman equation for optimal value functions in (3.2), combined with $\mathbb{P}(s'|s,a) = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}(s'|s,a)\rangle$ from Definition 3.1.

On the other hand, for the lower bound, the definition of $\mathcal{B}$ in (3.4) tells us $\langle \boldsymbol{\theta}, \boldsymbol{\phi}(\cdot|s,a)\rangle$ is a probability measure on $\mathcal{S}$ for

arbitrary $\boldsymbol{\theta} \in \mathcal{B}$, so their exists such a probability measure $\widetilde{\mathbb{P}}$ such that

$$\begin{aligned} Q_l^{(u)}(s,a) &= r(s,a) + \gamma \max_{\boldsymbol{\theta} \in \mathcal{B} \cap \mathcal{C}_l} \int_{\mathcal{S}} V^{(u-1)}(s') \langle \boldsymbol{\theta}, \boldsymbol{\phi}(s'|s,a) \rangle \mathrm{d}s' \\ &= r(s,a) + \gamma \int_{\mathcal{S}} \widetilde{\mathbb{P}}(s'|s,a) V^{(u-1)}(s') \mathrm{d}s' \\ &= r(s,a) + \gamma \mathbb{E}_{s' \sim \widetilde{\mathbb{P}}(s'|s,a)} V^{(u-1)}(s') \\ &\leq 1 + \gamma \cdot \frac{1}{1-\gamma} \\ &= \frac{1}{1-\gamma}, \end{aligned} \tag{B.8}$$

where in the inequality we used $r(s,a) \leq 1$ and the proven conclusion $V^{(u-1)} \leq 1/(1-\gamma)$ from the induction hypothesis.

Since Q_t^l, V_t are the returns of Algorithm 2 at time step t, we have that $Q_t^l(s,a)$ is the value function from the final iteration, and so $1/(1-\gamma) \geq Q_t^l(s,a) \geq Q^*(s,a)$; furthermore, from Line 12 of Algorithm 2, we have that $1/(1-\gamma) \geq V_t(s) = \max_a \min_{1 \leq l \leq L} Q_t^l(s,a) \geq \max_a Q^*(s,a) = V^*(s)$. These are the desired results, and our proof is completed. □

B.4. Proof of Lemma 6.4

Proof of Lemma 6.4. Again focusing on one single run of ML-EVI, we first prove the following inequality:

$$\max_{s,a,l} \left| Q_l^{(u)}(s,a) - Q_l^{(u-1)}(s,a) \right| \leq \gamma^{u-1}, \forall u \in \{1, 2, \ldots, U\}, \tag{B.9}$$

where in the maximization, the variables are taken from the following sets: $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $l \in \{1, 2, \ldots, L\}$.

We use induction on u for the proof. For $u = 1$, since $Q_l^{(0)}(s,a) = 1/(1-\gamma)$, we have $V^{(0)}(s) = \max_{a \in \mathcal{A}} \min_{1 \leq l \leq L} 1/(1-\gamma) = 1/(1-\gamma)$, and hence from (B.8) we see

$$\begin{aligned} Q_l^{(1)}(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim \widetilde{\mathbb{P}}(s'|s,a)} V^{(0)}(s') \\ &= r(s,a) + \frac{\gamma}{1-\gamma} \\ &\in \left[\frac{\gamma}{1-\gamma}, \frac{1}{1-\gamma} \right], \end{aligned}$$

which leads to $\left| Q_l^{(0)}(s,a) - Q_l^{(1)}(s,a) \right| \leq \left| 1/(1-\gamma) - \gamma/(1-\gamma) \right| = 1$ for all s, a, l.

Now assume inequality (B.9) holds for $u-1$. From the update rule of Q in Line 6, we have

$$\begin{aligned} Q_l^u(s,a) &= r(s,a) + \gamma \max_{\boldsymbol{\theta} \in \mathcal{B} \cap \mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(u-1)}}(s,a) \rangle, \\ Q_l^{u-1}(s,a) &= r(s,a) + \gamma \max_{\boldsymbol{\theta} \in \mathcal{B} \cap \mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(u-2)}}(s,a) \rangle. \end{aligned}$$

We therefore obtain by subtracting the two equations that

$$\begin{aligned} \left| Q_l^u(s,a) - Q_l^{u-1}(s,a) \right| &= \gamma \left| \max_{\boldsymbol{\theta} \in \mathcal{B} \cap \mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(u-1)}}(s,a) \rangle - \max_{\boldsymbol{\theta} \in \mathcal{B} \cap \mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(u-2)}}(s,a) \rangle \right| \\ &\leq \gamma \max_{\boldsymbol{\theta} \in \mathcal{B} \cap \mathcal{C}_l} \left| \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(u-1)}}(s,a) - \boldsymbol{\phi}_{V^{(u-2)}}(s,a) \rangle \right| \\ &= \gamma \left| \widetilde{\mathbb{P}} \big(V^{(u-1)} - V^{(u-2)} \big)(s,a) \right| \\ &\leq \gamma \max_s \left| \big(V^{(u-1)} - V^{(u-2)} \big)(s) \right|, \end{aligned} \tag{B.10}$$

where as before, we denote $\widetilde{\mathbb{P}} = \langle \widetilde{\boldsymbol{\theta}}, \boldsymbol{\phi}_V \rangle$ as the probability measure corresponding to $\widetilde{\boldsymbol{\theta}}$, which attains the maximum in the second line above.

Next, the update rule for V in Line 3 of Algorithm 2 suggests that

$$V^{(u-1)}(s) = \max_a \min_{1\le l\le L} Q_l^{(u-1)}(s,a),$$
$$V^{(u-2)}(s) = \max_a \min_{1\le l\le L} Q_l^{(u-2)}(s,a).$$

Plugging these into (B.10), we get

$$\begin{aligned}
\left|Q_l^{(u)}(s,a) - Q_l^{(u-1)}(s,a)\right| &\le \gamma \max_s \left| \max_a \min_{1\le l\le L} Q_l^{(u-1)}(s,a) - \max_a \min_{1\le l\le L} Q_l^{(u-2)}(s,a)\right| \\
&\le \gamma \max_{s,a,l} \left|Q_l^{(u-1)}(s,a) - Q_l^{(u)}(s,a)\right| \\
&\le \gamma \cdot \gamma^{u-2} \\
&= \gamma^{u-1},
\end{aligned}$$

where we used the induction hypothesis in the third inequality.

Now consider the result of Algorithm 2 at time step t, namely $Q_t^l = Q_l^{(U)}$ and $V_t = V^{(U)}$. There again exists a probability measure $\widetilde{\mathbb{P}} = \langle \widetilde{\boldsymbol{\theta}}, \boldsymbol{\phi}\rangle$ to support the following:

$$\begin{aligned}
Q_t^l(s_t,a_t) &= Q_l^{(U)}(s_t,a_t) \\
&= r(s_t,a_t) + \gamma \max_{\boldsymbol{\theta}\in\mathcal{B}\cap\mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(U-1)}}(s_t,a_t)\rangle \\
&= r(s_t,a_t) + \gamma \widetilde{\mathbb{P}} V^{(U-1)}(s_t,a_t) \\
&= r(s_t,a_t) + \gamma\left[\widetilde{\mathbb{P}} V_t(s_t,a_t) + \widetilde{\mathbb{P}}(V^{(U-1)} - V^{(U)})(s_t,a_t)\right] \\
&\le r(s_t,a_t) + \gamma \widetilde{\mathbb{P}} V_t(s_t,a_t) + \gamma \max_{s\in\mathcal{S}} \left|V^{(U-1)} - V^{(U)}\right|(s) \\
&\le r(s_t,a_t) + \gamma \max_{\boldsymbol{\theta}\in\mathcal{B}\cap\mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V_t}(s_t,a_t)\rangle + \gamma \max_{s\in\mathcal{S}} \left| \max_a \min_l Q^{(U-1)}(s,a) - \max_a \min_l Q^{(U)}(s,a)\right| \\
&\le r(s_t,a_t) + \gamma \max_{\boldsymbol{\theta}\in\mathcal{B}\cap\mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V_t}(s_t,a_t)\rangle + \gamma \max_{s,a,l} \left|Q^{(U-1)}(s,a) - Q^{(U)}(s,a)\right| \\
&\le r(s_t,a_t) + \gamma \max_{\boldsymbol{\theta}\in\mathcal{B}\cap\mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V_t}(s_t,a_t)\rangle + \gamma^U,
\end{aligned}$$

where in the last inequality we implemented (B.9). This conludes our proof of this lemma. □

C. Proof of Lemmas in Appendix A

C.1. Proof of Lemma A.1

Proof of Lemma A.1. We use induction on time step t. When $t = 1$, according to the initial values $\boldsymbol{\Sigma}_t^l = \lambda\mathbf{I}$ and $\mathbf{b}_t^l = \mathbf{0}$, the alternative expressions hold.

Suppose the conclusion holds for t. We compare the variables at time steps t and $t+1$ of Algorithm 1, respectively after the execution of Line 7. According to the update rules in Lines 14 to 19, if $l_t = L_t + 1$ and thus $L_{t+1} = L_t + 1$, new dormant levels for $l = 4L_t + C_0 + 1, \ldots, 4L_t + C_0 + 4$ are created. By the rule specified in Line 16, we use the induction hypothesis on (t,l) and $(t, 4L_t + C_0)$, to see that for $l > 4L_t + C_0$, the additional variables $\boldsymbol{\Sigma}$ are

$$\begin{aligned}
\boldsymbol{\Sigma}^l - \lambda\mathbf{I} &= \frac{1}{2}\left(\boldsymbol{\Sigma}^{l-1} - \lambda\mathbf{I}\right) \\
&= (1/2)^{l-(4L_t+C_0)}\left(\boldsymbol{\Sigma}_t^{4L_t+C_0} - \lambda\mathbf{I}\right) \\
&= 2^{(4L_t+C_0)-l}\sum_{\tau=1}^{t-1} \mathbb{1}(l_\tau \le 4L_t + C_0) 2^{l_\tau - (4L_t+C_0)} \boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)^\top \\
&= \sum_{\tau=1}^{t-1} \mathbb{1}(l_\tau \le l) 2^{l_\tau - l} \boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)^\top \\
&= \boldsymbol{\Sigma}_t^l - \lambda\mathbf{I},
\end{aligned}$$

where in the second to last equation we used the fact $l_\tau \le L_t + 1 < 4L_t + C_0$ for arbitrary $\tau < t$. Similarly, for the new variables $\mathbf{b}$,

$$\begin{aligned}
\mathbf{b}^l &= \frac{1}{2}\mathbf{b}^{l-1} \\
&= (1/2)^{l-(4L_t+C_0)}\mathbf{b}_t^{4L_t+C_0} \\
&= 2^{(4L_t+C_0)-l}\sum_{\tau=1}^{t-1}\mathbb{1}(l_\tau \le 4L_t + C_0)2^{l_\tau-(4L_t+C_0)}\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)V_\tau(s_{\tau+1}) \\
&= \sum_{\tau=1}^{t-1}\mathbb{1}(l_\tau \le l)2^{l_\tau-l}\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)V_\tau(s_{\tau+1}) \\
&= \mathbf{b}_t^l.
\end{aligned}$$

In other words, these variables for the newly created dormant levels are the same as the respective extended variables at time step t, and so this update rule can be ignored since we are using the alternative expressions for $\mathbf{\Sigma}_t^l$ and $\mathbf{b}_t^l$ for all t, l.

Now we go on to the update rules in Lines 20 to 23. For levels $l < l_t$, the two variables of interest do not change, hence

$$\begin{aligned}
\mathbf{\Sigma}_{t+1}^l &= \mathbf{\Sigma}_t^l \\
&= \lambda\mathbf{I} + \sum_{\tau=1}^{t-1}\mathbb{1}(l_\tau \le l)2^{l_\tau-l}\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)^\top \\
&= \lambda\mathbf{I} + \sum_{\tau=1}^{t}\mathbb{1}(l_\tau \le l)2^{l_\tau-l}\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)^\top,
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{b}_{t+1}^l &= \mathbf{b}_t^l \\
&= \lambda\mathbf{I} + \sum_{\tau=1}^{t-1}\mathbb{1}(l_\tau \le l)2^{l_\tau-l}\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)V_\tau(s_{\tau+1}) \\
&= \lambda\mathbf{I} + \sum_{\tau=1}^{t}\mathbb{1}(l_\tau \le l)2^{l_\tau-l}\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)V_\tau(s_{\tau+1}).
\end{aligned}$$

For levels $l \ge l_t$, a single term is added to the two variables respectively. Namely,

$$\begin{aligned}
\mathbf{\Sigma}_{t+1}^l &= \mathbf{\Sigma}_t^l + 2^{l_t-l}\boldsymbol{\phi}_{V_t}(s_t,a_t)\boldsymbol{\phi}_{V_t}(s_t,a_t)^\top \\
&= \lambda\mathbf{I} + \sum_{\tau=1}^{t-1}\mathbb{1}(l_\tau \le l)2^{l_\tau-l}\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)^\top + 2^{l_t-l}\boldsymbol{\phi}_{V_t}(s_t,a_t)\boldsymbol{\phi}_{V_t}(s_t,a_t)^\top \\
&= \lambda\mathbf{I} + \sum_{\tau=1}^{t}\mathbb{1}(l_\tau \le l)2^{l_\tau-l}\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)^\top,
\end{aligned}$$

and that

$$\begin{aligned}
\mathbf{b}_{t+1}^l &= \mathbf{b}_t^l + 2^{l_t-l}\boldsymbol{\phi}_{V_t}(s_t,a_t)V_t(s_{t+1}) \\
&= \sum_{\tau=1}^{t-1}\mathbb{1}(l_\tau \le l)2^{l_\tau-l}\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)V_\tau(s_{\tau+1}) + 2^{l_t-l}\boldsymbol{\phi}_{V_t}(s_t,a_t)V_\tau(s_{\tau+1}) \\
&= \sum_{\tau=1}^{t}\mathbb{1}(l_\tau \le l)2^{l_\tau-l}\boldsymbol{\phi}_{V_\tau}(s_\tau,a_\tau)V_\tau(s_{\tau+1}).
\end{aligned}$$

These expressions suggest the conclusion holds for arbitrary $l \le 4L_t + C_0$ at time step t. □

C.2. Proof of Lemma A.2

Proof of Lemma A.2. We use induction on u for this proof. For $u = 0$, by the initialization rules $Q^{(0)}_{l_1,*} = Q^{(0)}_{l_2} = 1/(1-\gamma)$, $\forall l_1 \in \{1, 2, \ldots\}, l_2 \in \{1, \ldots, 4L_t + C_0\}$, and hence $V^{(0)}_* = V^{(0)} = 1/(1-\gamma)$.

Suppose the conclusions already hold for $u-1$. Consider the conclusions of Lemma D.2, which states the confidence sets of levels above $4L_t + C_0$ trivially contain the entirety of $\mathcal{B}$, or $\mathcal{B} \cap \mathcal{C}_l = \mathcal{B} \supset \mathcal{B} \cap \mathcal{C}_{4L_t+C_0}$. Then for any level $l > 4L_t + C_0$ and state-action pair (s, a), by the update rule for Q in Line 6 of Algorithm 2:

$$\begin{aligned}
Q^{(u)}_{l,*}(s,a) &= r(s,a) + \gamma \max_{\boldsymbol{\theta} \in \mathcal{B} \cap \mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(u-1)}_*}(s,a) \rangle \\
&= r(s,a) + \gamma \max_{\boldsymbol{\theta} \in \mathcal{B} \cap \mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(u-1)}}(s,a) \rangle \\
&\geq r(s,a) + \gamma \max_{\boldsymbol{\theta} \in \mathcal{B} \cap \mathcal{C}_{4L_t+C_0}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(u-1)}}(s,a) \rangle \\
&= Q^{(u)}_{4L_t+C_0}(s,a).
\end{aligned}$$

Moreover, for any $l \leq 4L_t + C_0$, we have

$$\begin{aligned}
Q^{(u)}_{l,*}(s,a) &= r(s,a) + \gamma \max_{\boldsymbol{\theta} \in \mathcal{B} \cap \mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(u-1)}_*}(s,a) \rangle \\
&= r(s,a) + \gamma \max_{\boldsymbol{\theta} \in \mathcal{B} \cap \mathcal{C}_l} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V^{(u-1)}}(s,a) \rangle \\
&= Q^{(u)}_l(s,a)
\end{aligned}$$

We now give the relation for the value functions V. For any state s, by the update rule in Line 3 of Algorithm 2,

$$\begin{aligned}
V^{(u)}_*(s) &= \max_{a \in \mathcal{A}} \min_{l \geq 1} Q^{(u)}_{l,*}(s,a) \\
&= \max_{a \in \mathcal{A}} \min \left\{ \min_{1 \leq l \leq 4L_t + C_0} Q^{(u)}_{l,*}(s,a), \min_{l > 4L_t + C_0} Q^{(u)}_{l,*}(s,a) \right\} \\
&= \max_{a \in \mathcal{A}} \min \left\{ \min_{1 \leq l \leq 4L_t + C_0} Q^{(u)}_{l}(s,a), \min_{l > 4L_t + C_0} Q^{(u)}_{l,*}(s,a) \right\} \\
&= \max_{a \in \mathcal{A}} \min_{1 \leq l \leq 4L_t + C_0} Q^{(u)}_l(s,a) \\
&= V^{(u)}(s),
\end{aligned}$$

where we used the relations above in the third and fourth equations. This concludes the induction and our proof. □

C.3. Proof of Lemma A.3

Proof of Lemma A.3. Consider the definition in (3.4):

$$\mathcal{B} := \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \leq \sqrt{d} \text{ and } \langle \boldsymbol{\phi}(\cdot|s,a), \boldsymbol{\theta} \rangle \text{ is a probability measure on } \mathcal{S} \right\}.$$

We separate $\mathcal{B}$ into $\mathcal{B}_1 \cap \mathcal{B}_2$, where $\mathcal{B}_1 := \left\{ \boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \leq \sqrt{d} \right\}$ and $\mathcal{B}_2 := \left\{ \boldsymbol{\theta} : \langle \boldsymbol{\phi}(\cdot|s,a), \boldsymbol{\theta} \rangle \text{ is a probability measure on } \mathcal{S} \right\}$. Obviously $\mathcal{B}_1$ is a convex body in $\mathbb{R}^d$. For $\mathcal{B}_2$, consider the various constraints. For an arbitrary state-action pair (s, a), $\langle \boldsymbol{\phi}(\cdot|s,a), \boldsymbol{\theta} \rangle$ being a probability measure is equivalent to the following conditions:

$$\begin{aligned}
&\langle \boldsymbol{\phi}(s'|s,a), \boldsymbol{\theta} \rangle \geq 0, \forall s' \in \mathcal{S}, \\
&\int_{s' \in \mathcal{S}} \langle \boldsymbol{\phi}(s'|s,a), \boldsymbol{\theta} \rangle \mathrm{d}s' = 1.
\end{aligned}$$

The first constraint defines a closed half-space $\mathcal{H}_{s'|s,a} := \{\boldsymbol{\theta} : \langle \boldsymbol{\phi}(s'|s,a), \boldsymbol{\theta} \rangle \geq 0\}$ in $\mathbb{R}^d$ for all states $s' \in \mathcal{S}$, while the second defines a hyperplane $\mathcal{P}_{s,a} := \{\boldsymbol{\theta} : \langle \int_{s' \in \mathcal{S}} \boldsymbol{\phi}(s'|s,a) \mathrm{d}s', \boldsymbol{\theta} \rangle = 1\}$ in $\mathbb{R}^d$. With these notations, we can now write

$$\mathcal{B}_2 = \left(\bigcap_{s',s,a} \mathcal{H}_{s'|s,a} \right) \bigcap \left(\bigcap_{s,a} \mathcal{P}_{s,a} \right).$$

Since the sets $\mathcal{H}_{s'|s,a}$ and $\mathcal{P}_{s,a}$ are all closed convex sets, $\mathcal{B}_2$ being the intersection of these sets must also be closed and convex, and finally so must the intersection of all the sets $\mathcal{B}$. Since $\boldsymbol{\theta}^* \in \mathcal{B}$, this intersection is non-empty.

Now consider the affine subspace $\mathcal{L} := \{\boldsymbol{\theta}^* + \lambda(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) : \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{B}, \lambda \in \mathbb{R}\}$ of $\mathbb{R}^d$ with dimension k, which is basically the subspace passing through $\boldsymbol{\theta}^*$ and spanned by all the vectors insider $\mathcal{B}$. To see that $\mathcal{L}$ is indeed an affine subspace of $\mathbb{R}^d$, we take $\kappa_1, \kappa_2 \in \mathcal{L}$, where $\kappa_1 = \boldsymbol{\theta}^* + \lambda_1(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)$ and $\kappa_2 = \boldsymbol{\theta}^* + \lambda_2(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)$, $\boldsymbol{\theta}_i \in \mathcal{B}$, $i = 1, 2, 3, 4$. Then for any $\mu \in \mathbb{R}$,

$$\begin{aligned}
\mu\kappa_1 + (1-\mu)\kappa_2 &= \boldsymbol{\theta}^* + \big(\lambda_1\mu\boldsymbol{\theta}_1 - \lambda_1\mu\boldsymbol{\theta}_2 + \lambda_2(1-\mu)\boldsymbol{\theta}_3 - \lambda_2(1-\mu)\boldsymbol{\theta}_4\big) \\
&= \boldsymbol{\theta}^* + \big(|\lambda_1\mu| + |\lambda_2(1-\mu)|\big) \cdot \big[\big(\mu'\boldsymbol{\theta}'_1 + (1-\mu')\boldsymbol{\theta}'_3\big) - \big(\mu'\boldsymbol{\theta}'_2 + (1-\mu')\boldsymbol{\theta}'_4\big)\big] \\
&\in \mathcal{L},
\end{aligned}$$

where $\mu' = |\lambda_1\mu| / \big(|\lambda_1\mu| + |\lambda_2(1-\mu)|\big)$, and we used the fact $\mathcal{B}$ is convex in the final relation.

By taking $\lambda = 1$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}^*$ in the definition of $\mathcal{B}$, it is clear that $\mathcal{B} \subset \mathcal{L}$. On the other hand, this subspace can be generated by the vectors $\{\boldsymbol{\theta} - \boldsymbol{\theta}^* : \boldsymbol{\theta} \in \mathcal{B}\}$ starting from $\boldsymbol{\theta}^*$. From this vector set, we select a basis for the subspace, namely $\{\boldsymbol{\theta}_i^* - \boldsymbol{\theta}^* : i = 1, \dots, k\}$, and denote $\boldsymbol{\theta}_0^* = \boldsymbol{\theta}^*$.

Now consider the simplex $\{\sum_{i=0}^k \lambda_i \boldsymbol{\theta}_i^* : \lambda_i \geq 0, \forall i, \sum_{i=0}^k \lambda_i = 1\}$, which, since it has $k+1$ linearly independent vertices, is k-dimensional with inner points when constrained to $\mathcal{L}$. Furthermore, because $\mathcal{B}$ is convex and contains all its vertices, this simplex must be a subset of $\mathcal{B}$, so $\mathcal{B}$ itself has inner points when constrained to $\mathcal{L}$. Combine this with the fact that $\mathcal{B}_1$ is bounded, and so must $\mathcal{B}$ be, we have the conclusion that $\mathcal{B}$ is a convex body when constrained to $\mathcal{L}$. □

D. Auxiliary Lemmas

We will give a few auxiliary lemmas in this section. The first is the well-known Azuma-Hoeffding inequality.

Lemma D.1. (Azuma-Hoeffding Inequality (Cesa-Bianchi & Lugosi, 2006)). Let $\{x_i\}_{i=1}^n$ be a real-valued martingale difference sequence with respect to the filtration $\{\mathcal{F}_i\}_{i=1}^n$, which suggests $\mathbb{E}\big[x_i \big| \mathcal{F}_i\big] = 0$ and x_i is $\mathcal{F}_{i+1}$-measurable. Further assume $|x_i| \leq K$ for some positive constant K. Then with probability at least $1 - \delta$, we have

$$\Big|\sum_{i=1}^n x_i\Big| \leq K\sqrt{2n \log \frac{1}{\delta}}.$$

This next lemma proves that the fictional confidence sets with levels greater than $4L_t + C_0$ are trivial.

Lemma D.2. For time step t and level $l > 4L_t + C_0$, we have $\mathcal{C}_t^l \supset \mathcal{B}$.

Proof of Lemma D.2. We first give an upper bound for the term k_t^l when $l > 4L_t + C_0$. The definition of k_t^l tells us $k_t^l = \sum_{\tau=1}^{t-1} \mathbb{1}\{l_\tau \leq l\} 2^{l_\tau - l} = 2^{L_t - l} \sum_{\tau=1}^{t-1} \mathbb{1}\{l_\tau \leq l\} 2^{l_\tau - L_t} = 2^{L_t - l} k_t^{L_t}$. Recall from Algorithm 1 that $C_0 = 4\lceil l_0 \rceil + 6$, we go on to apply the conclusion of Lemma 6.1:

$$\begin{aligned}
k_t^l &\leq 2^{L_t - l} \cdot 20d \max\{L_t 2^{2L_t - 2l_0}, l_0\} \\
&= 2^{L_t - l} \cdot 20(1-\gamma)^2 \max\{L_t 2^{2L_t}, l_0 2^{2l_0}\} \\
&\leq 20(1-\gamma)^2 \max\{L_t 2^{-L_t - C_0}, l_0 2^{-3L_t + 2l_0 - C_0}\} \\
&\leq 20/64 \cdot 2^{-l_0} \max\{L_t 2^{-L_t}, l_0 2^{-l_0}\} \\
&\leq 5/16 \cdot (1-\gamma)/\sqrt{d} \\
&< \log 1.5 \cdot (1-\gamma)/\sqrt{d},
\end{aligned}$$

where we used the definition for l_0 in Lemma 6.1 twice above, and used the trivial relation $2^x \geq x$ for arbitrary real number x in the third inequality.

Now we target the maximum eigenvalue of $\boldsymbol{\Sigma}_t^l$ when $l > 4L_t + C_0$. Borrowing from the determinant upper bound (B.2)

from the proof of Lemma 6.1, we have

$$\begin{aligned}\lambda_{\max}(\mathbf{\Sigma}_t^l) &\le \frac{\det(\mathbf{\Sigma}_t^l)}{\lambda^{d-1}} \\ &\le \frac{(\lambda + \lambda k_t^l/d)^d}{\lambda^{d-1}} \\ &= \lambda(1 + k_t^l/d)^d \\ &\le \lambda \exp k_t^l \\ &\le 1.5\lambda,\end{aligned}$$

where in the first inequality we used the fact all eigenvalues of $\mathbf{\Sigma}_t^l$ are no less than λ, and in the third inequality we used the elementary inequality $1 + x \le \exp(x)$.

Next we upper bound the 2-norm of $\mathbf{b}_t^l$ when $l > 4L_t + C_0$. By the alternative expression in Lemma A.1:

$$\begin{aligned}\|\mathbf{b}_t^l\|_2 &\le \sum_{\tau=1}^{t-1} \mathbb{1}(l \ge l_\tau) 2^{l_\tau - l} \|\boldsymbol{\phi}_{V_\tau}(s_\tau, a_\tau)\|_2 \cdot \left|V_\tau(s_{\tau+1})\right| \\ &\le k_t^l \cdot \frac{\sqrt{d}}{1-\gamma} \cdot \frac{1}{1-\gamma} \\ &\le \frac{\log 1.5}{1-\gamma},\end{aligned}$$

where in the second inequality we used $\|\boldsymbol{\phi}_{V_\tau}(s_\tau, a_\tau)\|_2 \le \sqrt{d}/(1-\gamma)$ from Definition 3.1 and $V_\tau(s_{\tau+1}) \le 1/(1-\gamma)$ from Lemma 6.3.

Now for any $\boldsymbol{\theta} \in \mathcal{B}$, we substitute these two bounds in the following deductions:

$$\begin{aligned}\|\boldsymbol{\theta} - (\mathbf{\Sigma}_t^l)^{-1}\mathbf{b}_t^l\|_{\mathbf{\Sigma}_t^l} &\le \|\boldsymbol{\theta}\|_{\mathbf{\Sigma}_t^l} + \|(\mathbf{\Sigma}_t^l)^{-1}\mathbf{b}_t^l\|_{\mathbf{\Sigma}_t^l} \\ &= \|\boldsymbol{\theta}\|_{\mathbf{\Sigma}_t^l} + \|\mathbf{b}_t^l\|_{(\mathbf{\Sigma}_t^l)^{-1}} \\ &\le \|\boldsymbol{\theta}\|_2 \cdot \sqrt{\lambda_{\max}(\mathbf{\Sigma}_t^l)} + \|\mathbf{b}_t^l\|_2 \cdot \frac{1}{\sqrt{\lambda_{\min}(\mathbf{\Sigma}_t^l)}} \\ &\le \sqrt{d} \cdot \sqrt{1.5\lambda} + \frac{\log 1.5}{1-\gamma} \cdot \frac{1}{\sqrt{\lambda}} \\ &= \sqrt{1.5d\lambda} + \frac{\log 1.5}{1-\gamma} \\ &< \beta^l,\end{aligned}$$

where in the second inequality we used the relation $\|x\|_{\mathbf{\Sigma}}^2 = x^\top \mathbf{\Sigma} x \le x^\top [\lambda_{\max}(\mathbf{\Sigma})\mathbf{I}] x = \|x\|_2^2 \cdot \lambda_{\max}(\mathbf{\Sigma})$, and in the third inequality we used $\|\boldsymbol{\theta}\|_2 \le \sqrt{d}$ for any $\boldsymbol{\theta} \in \mathcal{B}$. Thus when $l > 4L_t + C_0$, $\boldsymbol{\theta} \in \mathcal{C}_t^l$ for any $\boldsymbol{\theta} \in \mathcal{B}$, and hence $\mathcal{B} \subset \mathcal{C}_t^l$. □

Finally, we list a few useful elementary inequalities and relations that were implemented in our proof of the main theory, and gather them into one single lemma below.

Lemma D.3. The following conclusions hold:

1. For $\eta \ge 2$ and integer $L > 0$, we have $\sum_{l=L}^{\infty} \eta^{-l}\sqrt{l} \le \sqrt{L+1}\eta^{-L+1}/(\eta - 1)$.
2. For any positive real numbers a, b and x, the inequality $x - a\sqrt{x} \le b$ implies the inequality $x \le a^2 + 2b$.
3. For $a, b \ge 1$, any real number $x \ge \log\left(a + b\sqrt{2\log(a+b)/\log 2}\right)/\log 2$ satisfies the equation $2^x \ge a + b\sqrt{x}$.

Proof of Lemma D.3. For the first inequality:

$$\begin{aligned}
\sum_{l=L}^{\infty}\eta^{-l}\sqrt{l} &= \sum_{l=L}^{\infty}\eta^{-l}\Big(\sqrt{L}+\sum_{l'=L+1}^{l}\big(\sqrt{l'}-\sqrt{l'-1}\big)\Big)\\
&= \sqrt{L}\sum_{l=L}^{\infty}\eta^{-l}+\sum_{l'=L+1}^{\infty}\big(\sqrt{l'}-\sqrt{l'-1}\big)\sum_{l=l'}^{\infty}\eta^{-l}\\
&= \sqrt{L}\frac{\eta^{-L+1}}{\eta-1}+\sum_{l'=L+1}^{\infty}\big(\sqrt{l'}-\sqrt{l'-1}\big)\frac{\eta^{-l'+1}}{\eta-1}\\
&\le \sqrt{L}\frac{\eta^{-L+1}}{\eta-1}+\sum_{l'=L+1}^{\infty}\big(\sqrt{L+1}-\sqrt{L}\big)\frac{\eta^{-l'+1}}{\eta-1}\\
&= \sqrt{L}\frac{\eta^{-L+1}}{\eta-1}+\big(\sqrt{L+1}-\sqrt{L}\big)\frac{\eta^{-L+1}}{(\eta-1)^2}\\
&\le \sqrt{L+1}\eta^{-L+1}/(\eta-1),
\end{aligned}$$

where we swapped the order of summation in the second equation, used the fact that $\sqrt{l}-\sqrt{l-1}$ is decreasing in l in the first inequality, and $\eta-1\ge 1$ in the last inequality.

For the second relation, suppose positive real numbers a, b, x satisfy $x - a\sqrt{x} \le b$. Formulating the left hand side to a squared expression, we get $(\sqrt{x}-a/2)^2 \le a^2/4+b$, and further $\sqrt{x}\le a/2+\sqrt{a^2/4+b}$. This leads to

$$\begin{aligned}
x &\le \big(a/2+\sqrt{a^2/4+b}\big)^2\\
&\le 2\cdot\left(\frac{a^2}{4}+\frac{a^2}{4}+b\right)\\
&= a^2+2b,
\end{aligned}$$

where we used the inequality $(x+y)^2\le 2(x^2+y^2)$ in the inequality.

For the third relation, we first substitute $x_0 := \log\big(a+b\sqrt{2\log(a+b)/\log 2}\big)/\log 2$ into $2^x - a - b\sqrt{x}$, and obtain

$$\begin{aligned}
2^{x_0}-a-b\sqrt{x_0} &= a+b\sqrt{2\log(a+b)/\log 2}-a-b\sqrt{x_0}\\
&= b\frac{\log(a+b)^2/\log 2 - x_0}{\sqrt{\log(a+b)^2/\log 2}+\sqrt{x_0}}\\
&= G(a,b)\log\frac{(a+b)^2}{a+b\sqrt{2\log(a+b)/\log 2}}/\log 2,
\end{aligned}$$

where we gathered everything besides the numerator in the second line into $G(a,b)$ for the final equation. $G(a,b)$ as a function in a, b is evidently always positive. After this we can see $a+b\sqrt{2\log(a+b)/\log 2}\le a+b\cdot\sqrt{2(a+b)}\le(a+b)^2$ since $a, b\ge 1$, so $2^{x_0}-a-b\sqrt{x_0}\ge 0$. Now we only need to prove for $x\ge x_0$, $\mathrm{d}(2^x-a-b\sqrt{x})/\mathrm{d}x\ge 0$, which is guaranteed since

$$\begin{aligned}
\frac{\mathrm{d}(2^x-a-b\sqrt{x})}{\mathrm{d}x} &= \log 2\cdot 2^x-\frac{b}{2\sqrt{x}}\\
&> \log 2\cdot(2^{x_0}-b\sqrt{x_0})\\
&> 0,
\end{aligned}$$

where we used the fact $x\ge x_0\ge\log(1+1)/\log 2 = 1$ and $2\log 2>1$ in the first inequality. This concludes our proof. □