Provably Efficient Infinite-Horizon Average-Reward Reinforcement Learning with Linear Function Approximation

Woojin Chae[1] WOOJEENY02@KAIST.AC.KR

Dabeen Lee[2,†] DABEENL@KAIST.AC.KR

[1] *Department of Mathematical Sciences, KAIST, Daejeon 34141, South Korea*

[2] *Department of Industrial and Systems Engineering, KAIST, Daejeon 34141, South Korea*

† *Corresponding author*

Abstract

This paper proposes a computationally tractable algorithm for learning infinite-horizon average-reward linear Markov decision processes (MDPs) and linear mixture MDPs under the Bellman optimality condition. While guaranteeing computational efficiency, our algorithm for linear MDPs achieves the best-known regret upper bound of $\widetilde{\mathcal{O}}(d^{3/2}\mathrm{sp}(v^*)\sqrt{T})$ over T time steps where $\mathrm{sp}(v^*)$ is the span of the optimal bias function v^* and d is the dimension of the feature mapping. For linear mixture MDPs, our algorithm attains a regret bound of $\widetilde{\mathcal{O}}(d \cdot \mathrm{sp}(v^*)\sqrt{T})$. The algorithm applies novel techniques to control the covering number of the value function class and the span of optimistic estimators of the value function, which is of independent interest.

1 Introduction

Reinforcement learning (RL) with function approximation schemes has achieved remarkable success in a wide range of areas, including video games (Mnih et al., 2015), Go (Silver et al., 2017), robotics (Kober et al., 2013), and autonomous driving (Yurtsever et al., 2020). Such empirical progress has stimulated endeavors to expand our theoretical understanding of RL with function approximation.

As a first step toward establishing theoretical foundations, linear function approximation frameworks have received significant attention. Among them, there has been a plethora of activities in *linear Markov decision processes (MDPs)* (Yang and Wang, 2019; Jin et al., 2020) and *linear mixture MDPs* (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b) where the underlying transition kernel and the reward function are assumed to be parameterized as a linear function of some given feature mappings over state-action pairs or state-action-state triplets. In particular, for the finite-horizon setting, nearly minimax optimal algorithms have already been developed for linear MDPs (He et al., 2023; Agarwal et al., 2023; Hu et al., 2022) and for linear mixture MDPs (Zhou et al., 2021a).

However, there still remains a gap in our understanding of *infinite-horizon average-reward MDPs* with linear function approximation. The best-known regret lower bound is $\Omega(d\sqrt{DT})$ for both linear MDPs and linear mixture MDPs, achieved by a communicating MDP instance with a d-dimensional feature mapping and diameter D over T time steps (Wu et al., 2022). On the other hand, the best-known regret upper bound for linear MDPs

is $\widetilde{\mathcal{O}}(d^{3/2}\mathrm{sp}(v^*)\sqrt{T})$ where $\mathrm{sp}(v^*)$ is the span of the optimal bias function v^* for the underlying MDP of bounded span (Wei et al., 2021). Here, if the underlying MDP is communicating, the span is bounded above by the diameter, which means that there is a gap of $\widetilde{\mathcal{O}}(\sqrt{d \cdot \mathrm{sp}(v^*)})$ between the lower and upper bounds. Let alone the gap, the algorithm is computationally inefficient, as it requires solving a fixed-point equation at each iteration. For linear mixture MDPs, there is an algorithm that achieves a regret upper bound of $\widetilde{\mathcal{O}}(d\sqrt{DT})$ (Wu et al., 2022), but it is limited to the communicating case. The current status of progress on the infinite-horizon average-reward setting motivates the following question.

Does there exist a computationally efficient algorithm with a tight regret upper bound for learning infinite-horizon average-reward linear and linear mixture MDPs of bounded span?

This paper answers the question affirmatively. Let us summarize our contributions as follows.

- We propose a computationally efficient algorithm, least-squares value iteration with discounting and clipping (LSVI-DC, Algorithm 1) that works for both linear and linear mixture MDPs. We show that for the linear MDP setting, LSVI-DC achieves the best-known regret upper bound of $\widetilde{\mathcal{O}}(d^{3/2}\mathrm{sp}(v^*)\sqrt{T})$ (Theorem 1). It is the first correct provably efficient algorithm that attains the best-known regret upper bound for learning infinite-horizon average-reward linear MDPs. At the core of the algorithm is the least squares-based estimation method for the unknown transition coefficient vector due to Jin et al. (2020).

- By the reduction from a tabular MDP to a linear MDP, the regret bound for linear MDPs translates to $\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)\sqrt{S^3A^3T})$ for the model-free tabular setting, which improves upon the best-known regret bound of $\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)S^5A^2\sqrt{T})$ by Zhang and Xie (2023) (Appendix G).

- For linear mixture MDPs, LSVI-DC guarantees a regret upper bound of $\widetilde{\mathcal{O}}(d\cdot\mathrm{sp}(v^*)\sqrt{T})$ (Theorem 2). The regret upper bound is close to the best-known regret lower bound with a gap of $\widetilde{\mathcal{O}}(\sqrt{\mathrm{sp}(v^*)})$. The algorithm applies the linear regression-based method for the unknown transition coefficient vector due to Wu et al. (2022).

- LSVI-DC consists of many novel components in its design, which is of independent interest for infinite-horizon average-reward reinforcement learning. First, we apply the *clipping* operation to control the span of intermediate value functions, which is crucial to provide a bounded regret for the weakly communicating case. The clipping operation is much simpler to implement than constrained optimization-based frameworks to control the span. Second, we approximate a given average-reward MDP by a discounted-reward MDP on which we run a discounted optimistic value iteration. Third, to avoid the issue of blowing up the *covering number* for the case of linear MDPs, we take the *max-pooling* step instead of making the value function estimator monotonically decreasing.

The idea of approximating an average-reward MDP by a discounted-reward MDP has been adopted for the tabular case (Wei et al., 2020; Zhang and Xie, 2023) and used for learning linear MDPs (Hong et al., 2024). Clipping an optimistic value function estimator to control

Table 1: Summary of our results and comparison of algorithms for linear and linear mixture MDPs

Algorithm	Regret ($\widetilde{\mathcal{O}}(\cdot)$)	Assumption	Computation	Structure
FOPO (Wei et al., 2021)	$d^{3/2}\mathrm{sp}(v^*)\sqrt{T}$	Bellman optimality (finite span)	Inefficient	Linear MDP
OLSVI.FH (Wei et al., 2021)	$d^{3/4}\mathrm{sp}(v^*)T^{3/4}$	Bellman optimality (finite span)	Efficient	
MDP-EXP2 (Wei et al., 2021)	$d \cdot \tau_{\mathrm{mix}}^{3/2}\sqrt{T}$	Uniform mixing	Efficient	
LSVI-DC (Theorem 1)	$d^{3/2}\mathrm{sp}(v^*)\sqrt{T}$	Bellman optimality (finite span)	Efficient	
UCRL2-VTR (Wu et al., 2022)	$d\sqrt{DT}$	Communicating (finite diameter)	Efficient	Linear mixture MDP
LSVI-DC (Theorem 2)	$d \cdot \mathrm{sp}(v^*)\sqrt{T}$	Bellman optimality (finite span)	Efficient	
Lower Bound (Wu et al., 2022)	$\Omega(d\sqrt{DT})$			

its size is already a common practice when designing an algorithm for finite-horizon and infinite-horizon discounted-reward MDPs. However, the clipping operation in our algorithm sets the threshold in a different way to control the span of value functions, not their sizes, and it was first introduced by Hong et al. (2024). Our max-pooling step is also inspired by Hong et al. (2024), but their max-pooling step has an issue, which makes their algorithm incomplete (See Appendix A).

In this paper, we provide a correct provably efficient algorithm based on these components. We remark that not only our algorithm design but also our regret analysis are different from Hong et al. (2024). We summarize our results and compare the most relevant works to ours in Table 1.

2 Related Work

Reinforcement Learning with Linear Function Approximation Recently, there has been remarkable progress in reinforcement learning frameworks with linear function approximation (Jiang et al., 2017; Yang and Wang, 2019, 2020; Jin et al., 2020; Wang et al., 2021; Modi et al., 2020; Dann et al., 2018; Du et al., 2021; Sun et al., 2019; Zanette et al., 2020a,b; Cai et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Weisz et al., 2021; Zhou et al., 2021b,a; He et al., 2021; Zhou and Gu, 2022; Hu et al., 2022; He et al., 2023; Agarwal et al., 2023). These works develop frameworks for MDP classes with certain linear structures. Among them, the most relevant to this paper are linear and linear mixture MDPs. Linear MDPs assume that the transition probability and the reward function are linear in some fea-

ture mapping $\varphi(s,a)$ for each state-action pair (s,a). In linear mixture MDPs, the reward function is assumed to be linear in such $\varphi(s,a)$ while the transition probability is linear in some feature mapping $\phi(s,a,s')$ for each state-action-state triplet (s,a,s'). Although the two classes are closely related, one cannot be covered by the other (Zhou et al., 2021b). For the finite-horizon setting, we have minimax optimal algorithms for linear MDPs (He et al., 2023; Agarwal et al., 2023; Hu et al., 2022) and for linear mixture MDPs (Zhou et al., 2021a). For learning infinite-horizon average-reward linear MDPs, Wei et al. (2021) developed several algorithms, as summarized in Table 1. FOPO achieves the best-known regret upper bound, but it needs to solve a fixed-point equation at each iteration, making the algorithm intractable. OLSVI.FH and MDP-EXP2 are tractable algorithms, but OLSVI.FH leads to a suboptimal regret and MDP-EXP2 works under a restrictive assumption. For learning infinite-horizon average-reward linear mixture MDPs, Wu et al. (2022) developed an algorithm that is shown to be minimax optimal for the communicating case. He et al. (2024) proposed an algorithm for RL with general function approximation, LOOP, incorporating linear and linear mixture MDPs as subclasses. LOOP attains $\widetilde{\mathcal{O}}(d^{3/2}\mathrm{sp}(v^*)^{3/2}\sqrt{T})$ regret for linear MDPs, which is worse than FOPO and our algorithm. Moreover, it also achieves $\widetilde{\mathcal{O}}(d^{3/2}\mathrm{sp}(v^*)^{3/2}\sqrt{T})$ regret for linear mixture MDPs, but a slightly different set of assumptions from ours and that of Wu et al. (2022) is imposed. Nevertheless, LOOP is hardly practical as it relies on solving a complex constrained optimization problem.

Infinite-Horizon Average-Reward Reinforcement Learning The seminal work by Auer et al. (2008) pioneered algorithmic frameworks for model-based online learning of MDPs. They proved a regret lower bound of $\Omega(\sqrt{DSAT})$ where S is the number of states and A is the number of actions. Then they provided UCRL2 which is based on extended value iteration over some optimistic sets for estimating the transition probability and guarantees a regret bound of $\widetilde{\mathcal{O}}(DS\sqrt{AT})$. Bartlett and Tewari (2009) considered the class of weakly communicating MDPs and MDPs with bounded span, for which they proposed an algorithm that achieves a regret upper bound of $\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)S\sqrt{AT})$. Since then, there has been a long line of work toward closing the gap between regret upper and lower bounds (Filippi et al., 2010; Talebi and Maillard, 2018; Fruit et al., 2018, 2020; Bourel et al., 2020; Zhang and Ji, 2019; Agrawal and Jia, 2017; Ouyang et al., 2017; Abbasi-Yadkori et al., 2019; Wei et al., 2021; Zhang and Xie, 2023; Boone and Zhang, 2024). In particular, Fruit et al. (2018) and Zhang and Ji (2019) refined the regret lower bound to $\Omega(\sqrt{\mathrm{sp}(v^*)SAT})$. The first result with a regret upper bound matching the lower bound is due to Zhang and Ji (2019), but their algorithm is not tractable. Recently, Boone and Zhang (2024) developed a tractable algorithm that guarantees a regret bound of $\mathcal{O}(\sqrt{\mathrm{sp}(v^*)SAT})$. For model-free schemes, Wei et al. (2020) introduced a Q-learning-based algorithm that attains a suboptimal regret bound of $\widetilde{\mathcal{O}}(T^{2/3})$. Recently, Zhang and Xie (2023) developed UCB-AVG that guarantees $\mathcal{O}(\mathrm{sp}(v^*)S^5A^2\sqrt{T})$.

3 Preliminaries

Notations Given a vector $x \in \mathbb{R}^d$ and a positive semidefinite matrix $A \in \mathbb{R}^{d\times d}$, $\|x\|_2$ denotes the ℓ_2-norm of x, $\|x\|_A = \sqrt{x^\top Ax}$, $\|A\|_2$ is the spectral norm of A, and $\|A\|_F$ is the Frobenius norm of A. For any positive integers m, n with $m < n$, $[n]$ and $[m:n]$ denote $\{1,\dots,n\}$ and $\{m,\dots,n\}$, respectively.

Infinite-Horizon Average-Reward MDP We consider an MDP given by $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathbb{P}(s' \mid s, a)$ specifies the probability of transitioning to state s' from state s after taking action a, and $r(s, a) \in [0, 1]$ is the reward from action a at state s. A (stochastic) stationary policy is given as a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ where $\Delta(\mathcal{A})$ is the set of probability measures on $\mathcal{A}$, and we use notation $\pi(a \mid s)$ for the probability of taking action a at state s under policy π. When π is a deterministic policy, we write that $a = \pi(s)$ with abuse of notation where a is the action with $\pi(a \mid s) = 1$. At each time step t, an algorithm takes action a_t at given state s_t, after which it observes the next state s_{t+1} drawn from distribution $\mathbb{P}(\cdot \mid s_t, a_t)$. Then the cumulative reward over T steps is $\sum_{t=1}^{T} r(s_t, a_t)$. Then the (long-term) average reward is given by $\liminf_{T\to\infty} \mathbb{E}[\sum_{t=1}^{T} r(s_t, a_t)]/T$, which can be maximized by a deterministic stationary policy (See Puterman, 2014). We denote by $J^\pi(s) = \liminf_{T\to\infty} \mathbb{E}[\sum_{t=1}^{T} r(s_t, a_t) \mid s_1 = s]/T$ the average reward of a stationary policy π starting from initial state s.

In this paper, we focus on the class of MDPs satisfying the following form of Bellman optimality condition. There exist $J^* \in \mathbb{R}$, $v^* : \mathcal{S} \to \mathbb{R}$, and $q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$J^* + q^*(s, a) = r(s, a) + \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[v^*(s') \right] \quad \text{and} \quad v^*(s) = \max_{a \in \mathcal{A}} q^*(s, a). \tag{1}$$

Under the Bellman optimality condition, the optimal average reward $J^*(s) := \max_\pi J^\pi(s)$ is invariant with the initial state s, and $J^*(s) = J^*$ for any $s \in \mathcal{S}$ (Bartlett and Tewari, 2009). Moreover, the class of weakly communicating MDPs satisfies the condition (See Puterman, 2014). There indeed exist other general classes of MDPs with which the condition holds (Hernandez-Lerma, 2012, Section 3.3). For any function $h : \mathcal{S} \to \mathbb{R}$, we define its span as $\mathrm{sp}(h) := \max_{s \in \mathcal{S}} h(s) - \min_{s \in \mathcal{S}} h(s)$. Then we consider the following notion of regret to analyze the performance of an algorithm.

$$\mathrm{Regret}(T) = T \cdot J^* - \sum_{t=1}^{T} r(s_t, a_t).$$

Infinite-Horizon Discounted-Reward MDP We also consider the discounted cumulative reward of a stationary policy π given by $V^\pi(s) = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \mid s_1 = s]$ where s is the initial state and $\gamma \in (0, 1)$ is a discount factor. Similarly, we consider $Q^\pi(s, a) = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \mid (s_1, a_1) = (s, a)]$. Then we define the optimal value function V^* and the optimal action-value function Q^* as $V^*(s) = \max_\pi V^\pi(s)$ and $Q^*(s, a) = \max_\pi Q^\pi(s, a)$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$. It is known that there exists a deterministic stationary policy that gives rise to V^* and Q^* (See Puterman, 2014; Agarwal et al., 2021). Moreover, V^* and Q^* satisfy the following Bellman optimality equation.

$$Q^*(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[V^*(s') \right] \quad \text{and} \quad V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a). \tag{2}$$

Our approach is to approximate an average-reward MDP by a discounted-reward MDP. In fact, as the discount factor gets close to 1, the discounted cumulative reward converges to the average reward for a stationary policy (See Puterman, 2014).

Linear and Linear Mixture MDPs In this work, we focus on linear MDPs and linear mixture MDPs, which are formally defined as follows.

Assumption 1 (Linear MDP, Jin et al., 2020; Wei et al., 2021) *MDP $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r)$ is a linear MDP with a feature map $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

$$r(s, a) = \langle \varphi(s, a), \theta \rangle, \quad \mathbb{P}(s' \mid s, a) = \langle \varphi(s, a), \mu(s') \rangle$$

where $\mu = (\mu_1, \ldots, \mu_d)$ is an unknown measure over $\mathcal{S}$ and $\theta \in \mathbb{R}^d$ is an unknown vector. We assume that $\|\varphi(s, a)\|_2 \leq 1$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\|\theta\|_2 \leq \sqrt{d}$, and $\|\mu(\mathcal{S})\|_2 \leq \sqrt{d}$.

Assumption 2 (Linear Mixture MDP, Zhou et al., 2021a; Wu et al., 2022) *MDP $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r)$ is a linear mixture MDP with feature maps $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

$$r(s, a) = \langle \varphi(s, a), \theta^* \rangle, \quad \mathbb{P}(s' \mid s, a) = \langle \phi(s, a, s'), \theta^* \rangle$$

where $\theta^* \in \mathbb{R}^d$ is an unknown vector with $\|\theta^*\|_2 \leq B_\theta$ for some $B_\theta \in \mathbb{R}$. We further assume that for any bounded function $F : \mathcal{S} \to [0, H]$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\|\phi_F(s, a)\|_2 \leq H B_\phi$ for some $B_\phi \in \mathbb{R}$ where $\phi_F(s, a) = \int_{s'} \phi(s, a, s') F(s') ds'$. Lastly, $\|\varphi(s, a)\|_2 \leq 1$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$.

4 The Proposed Algorithm

In this section, we present our algorithm, LSVI-DC, described in Algorithm 1. LSVI-DC is designed to work for both linear and linear mixture MDPs, while some steps are different for the two settings. As common in algorithms for learning infinite-horizon average-reward MDPs such as UCRL2 (Auer et al., 2008) and UCRL2-VTR (Wu et al., 2022), LSVI-DC also proceeds with episodes. Following UCRL2-VTR, when to start the next episode is determined based on the Gram matrix (line 10). Each episode of LSVI-DC consists of two phases, the planning phase (lines 4-9) and the execution phase (lines 10-16). During the planning phase, we run optimistic value iteration for a discounted MDP with estimated parameters. Then, based on optimistic value functions deduced from the planning phase, we take and execute a greedy deterministic (non-stationary) policy for the execution phase. What follows provides a more detailed discussion of the important components of LSVI-DC.

Least Squares-Based Parameter Estimation For a linear MDP under Assumption 1, we may argue that $r(s, a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}[V(s')]$ for any bounded value function $V : \mathcal{S} \to \mathbb{R}$ can be written as $\langle \varphi(s, a), w \rangle$ for some $w \in \mathbb{R}^d$ with a bounded norm (Jin et al., 2020) (See also Lemma 5.1). In line 6, we compute an estimator $w^k_{(n)}$ for w that corresponds to $V = V^k_{(n)} - \min_{s' \in \mathcal{S}} V^k_{(n)}(s')$ by solving the following regularized least-squares problem.

$$w^k_{(n)} \leftarrow \operatorname*{argmin}_{w \in \mathbb{R}^d} \lambda \|w\|_2^2 + \sum_{\tau=1}^{t_k - 1} \left(r(s_\tau, a_\tau) + \gamma \left(V^k_{(n)}(s_{\tau+1}) - \min_{s' \in \mathcal{S}} V^k_{(n)}(s') \right) - \langle \varphi(s, a), w \rangle \right)^2.$$

The expression in line 6 for $w^k_{(n)}$ is precisely the optimal solution to this least-squares problem. For a linear mixture MDP under Assumption 2, $r(s, a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}[V(s')]$ equals

Algorithm 1 Least-Squares Value Iteration with Discounting and Clipping (LSVI-DC)

1: **Input:** discount factor $\gamma \in (0,1)$, regularization parameter $\lambda > 0$, upper bound H of $2 \cdot \mathrm{sp}(v^*)$, bonus factor β, radius values $\beta_1, \ldots, \beta_T$

2: **Initialize:**initial state s_1, $t \leftarrow 1$, $\Sigma_1 \leftarrow \lambda I$, $b_1 = 0$, $\theta_1 = 0$

3: **for** episode $k = 1, 2, \cdots$ **do**

4: $\quad t_k \leftarrow t$, $V^k_{(1)} \leftarrow \frac{1}{1-\gamma}$, $\widetilde{Q}^k_{(1)} \leftarrow \frac{1}{1-\gamma}$, $N_k \leftarrow T - t_k + 1$

5: $\quad$ **for** round $n = 1, \cdots, N_k$ **do**

6: $\quad\quad$ **(Linear MDP)**:

$$w^k_{(n)} \leftarrow \Sigma_{t_k}^{-1} \sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau)(r(s_\tau, a_\tau) + \gamma(V^k_{(n)}(s_{\tau+1}) - \min_{s' \in \mathcal{S}} V^k_{(n)}(s')))$$

$$\widetilde{Q}^k_{(n+1)}(\cdot,\cdot) \leftarrow \min\left\{\left\langle \varphi(\cdot,\cdot), w^k_{(n)} \right\rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V^k_{(n)}(s') + \beta \|\varphi(\cdot,\cdot)\|_{\Sigma_{t_k}^{-1}}, \frac{1}{1-\gamma}\right\}$$

$\quad\quad$ **(Linear mixture MDP)**:

$$\varphi^k_{(n)}(\cdot,\cdot) \leftarrow \varphi(\cdot,\cdot) + \gamma \cdot \phi_{\bar{V}^k_{(n)}}(\cdot,\cdot) \text{ where } \bar{V}^k_{(n)}(\cdot) = V^k_{(n)}(\cdot) - \min_{s' \in \mathcal{S}} V^k_{(n)}(s')$$

$$\widetilde{Q}^k_{(n+1)}(\cdot,\cdot) \leftarrow \min\left\{\left\langle \varphi^k_{(n)}(\cdot,\cdot), \theta_{t_k} \right\rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V^k_{(n)}(s') + \beta_{t_k} \|\varphi^k_{(n)}(\cdot,\cdot)\|_{\Sigma_{t_k}^{-1}}, \frac{1}{1-\gamma}\right\}$$

7: $\quad\quad \widetilde{V}^k_{(n+1)}(\cdot) \leftarrow \max_{a \in \mathcal{A}} \widetilde{Q}^k_{(n+1)}(\cdot, a)$

8: $\quad\quad V^k_{(n+1)}(\cdot) \leftarrow \min\left\{\widetilde{V}^k_{(n+1)}(\cdot),\ \min_{s' \in \mathcal{S}} \widetilde{V}^k_{(n+1)}(s') + H\right\}$

9: $\quad$ **end for**

10: $\quad$ **while** $\det(\Sigma_t) \leq 2\det(\Sigma_{t_k})$ **do**

11: $\quad\quad$ Take $\xi_t(a) \in \operatorname{argmax}_{n \in [T-t+1:N_k]} \widetilde{Q}^k_{(n)}(s_t, a)$ for all $a \in \mathcal{A}$

12: $\quad\quad$ Set $Q_t(s_t, a) \leftarrow \widetilde{Q}^k_{(\xi_t(a))}(s_t, a)$ for all $a \in \mathcal{A}$

13: $\quad\quad$ Take $a_t \in \operatorname{argmax}_{a \in \mathcal{A}} Q_t(s_t, a)$, receive $r(s_t, a_t)$, and obtain $s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$

14: $\quad\quad$ **(Linear MDP)**:

$$\Sigma_{t+1} \leftarrow \Sigma_t + \varphi(s_t, a_t)\varphi(s_t, a_t)^\top$$

$\quad\quad$ **(Linear mixture MDP)** :

$$\text{Take } W_t(\cdot) \leftarrow V^k_{(\xi_t(a_t)-1)}(s) - \min_{s' \in \mathcal{S}} V^k_{(\xi_t(a_t)-1)}(s')$$

$$\text{Update } \Sigma_{t+1} \leftarrow \Sigma_t + (\varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t))(\varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t))^\top$$

$$\text{Update } b_{t+1} \leftarrow b_t + (\varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t))(r(s_t, a_t) + \gamma \cdot W_t(s_{t+1}))$$

$$\text{Set } \theta_{t+1} \leftarrow \Sigma_{t+1}^{-1} b_{t+1}$$

15: $\quad\quad t \leftarrow t + 1$

16: $\quad$ **end while**

17: **end for**

$\langle \varphi(s,a) + \gamma \cdot \phi_V(s,a), \theta \rangle$. As described in line 14, we compute θ_t as an estimator for θ by solving the following value-targeted regression problem motivated by Ayoub et al. (2020); Wu et al. (2022).

$$\theta_t \leftarrow \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \lambda \|\theta\|_2^2 + \sum_{\tau=1}^{t-1} \left(r(s_\tau, a_\tau) + \gamma \cdot W_\tau(s_{\tau+1}) - \langle \varphi(s_\tau, a_\tau) + \gamma \cdot \phi_{W_\tau}(s_\tau, a_\tau), \theta \rangle\right)^2.$$

Note that θ_t computed in line 14 is the optimal solution to the regression problem.

Optimistic Value Iteration with Discounting For a linear MDP, we prove that the action-value function given by line 6 is an optimistic estimator for a proper value of β (Lemma 5.7). Note that the function is clipped by $(1-\gamma)^{-1}$, because we know that the infinite-horizon discounted cumulative reward is bounded above by $(1-\gamma)^{-1}$. Similarly, for the linear mixture MDP setting, the action-value function from line 6 with an appropriate β_{t_k} is an optimistic estimator (Lemma 5.7).

Clipping Operation In each round of value iteration, LSVI-DC applies the clipping operation given in line 8. Note that the value function $\widetilde{V}^k_{(n+1)}$ from line 7 does not necessarily have a bounded span. After the operation, it is clear that the span of $V^k_{(n+1)}$ becomes bounded above as $\mathrm{sp}(V^k_{(n+1)}) \leq H$. The clipping operation was first introduced by Hong et al. (2024). As a result, function W_t given in line 14 satisfies $W_t(s) \in [0, H]$ for any $s \in \mathcal{S}$. We choose any upper bound H on $2 \cdot \mathrm{sp}(v^*)$ where v^* is the optimal bias function from (1).

Max-Pooling The important step during the execution phase is the max-pooling step described in lines 11-12. Given state s_t in time step t, we choose an action based on the action-value function $Q_t(s_t, a)$, which is given as the maximum of $\widetilde{Q}^k_{(n)}(s_t, a)$ over $n \in [T-t+1 : N_k]$. Here, note that within the same episode, we enlarge the interval as time goes. This leads to a monotonically increasing behavior of value functions. To be precise, we argue that $V^k_{\xi_t(a_t)-1}(s_{t+1})$, which is computed at step t, is less than or equal to $Q_{t+1}(s_{t+1}, a_{t+1})$ (Section 5.3). Compared to the max-pooling step of Hong et al. (2024), we make the choice of index $\xi_t(a)$ dependent on the current state s_t. Our specific design of the max-pooling operation leads to a correct algorithm.
Deriving a monotone behavior in the sequence of value functions is indeed important to deduce a regret upper bound for RL with linear function approximation. For example, He et al. (2023) used the idea of taking the minimum of action-value functions generated up to the given time step. We may also attempt to apply the idea in our value iteration procedure, thereby replacing the max-pooling step. The resulting algorithm works for linear mixture MDPs, providing a regret upper bound of the same asymptotic scale as LSVI-DC. However, the algorithm does not work for linear MDPs, because the number of rounds for value iteration is up to T, in which case the covering number may explode. There is no such issue for linear mixture MDPs, as the linear mixture MDP case does not rely on the covering number argument. We provide and explain the algorithm with the operation of taking the minimum of value functions in Appendix F.

5 Regret Analysis of LSVI-DC

Let us state the following regret bounds of LSVI-DC for linear and linear mixture MDPs.

Theorem 1 (Linear MDP) *Set $\gamma = 1 - 1/\sqrt{T}$, $H \geq 2 \cdot \mathrm{sp}(v^*)$, $\lambda = 1$, and $\beta = 16(1+H)d\sqrt{\log(1+dT/\delta)}$. Then LSVI-DC guarantees with probability at least $1-2\delta$ that for any linear MDP with any initial state $s_1 \in \mathcal{S}$,*

$$\mathrm{Regret(T)} = \widetilde{\mathcal{O}}\left(d^{3/2}(1+H)\sqrt{T}\right)$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in dT/δ.

Theorem 2 (Linear Mixture MDP) *Set* $\gamma = 1 - 1/\sqrt{T}$, $H \geq 2 \cdot \mathrm{sp}(v^*)$, $\lambda = (B_\varphi + HB_\phi)^2$, *and* $\beta_t = H\sqrt{d\log((1+t(B_\varphi + HB_\phi)^2/\lambda)/\delta)} + B_\theta\sqrt{\lambda}$ *for* $t \geq 1$. *Then LSVI-DC guarantees with probability at least* $1 - 2\delta$ *that for any linear MDP with any initial state* $s_1 \in \mathcal{S}$,

$$\mathrm{Regret(T)} = \widetilde{\mathcal{O}}\left(d(1+H)\sqrt{T} + B_\theta(B_\varphi + HB_\phi)\sqrt{dT}\right)$$

where the $\widetilde{\mathcal{O}}(\cdot)$ *hides logarithmic factors in* dT/δ.

Theorem 1 also implies a regret upper bound of $\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)\sqrt{S^3A^3T})$ for the model-free tabular setting (Appendix G). The rest of this section outlines the proofs and explains the key techniques.

5.1 Concentration of Unknown Parameters

Let us consider the linear MDP case first. Recall that line 6 of Algorithm 1 computes the coefficient vector $w^k_{(n)}$ for a given linear MDP in each round n of value iteration in episode k. We argue that $w^k_{(n)}$ is an estimator of vector $w^{k,*}_{(n)}$ given in the following lemma.

Lemma 5.1 *For each episode* k *and round* n, *there exists* $w^{k,*}_{(n)} \in \mathbb{R}^d$ *such that for any* $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\left\langle \varphi(s,a), w^{k,*}_{(n)} \right\rangle = r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}\left[V^k_{(n)}(s')\right] - \gamma \min_{s' \in \mathcal{S}} V^k_{(n)}(s').$$

Moreover, $\|w^{k,*}_{(n)}\|_2 \leq (1+\gamma H)\sqrt{d}$ *and* $\|w^k_{(n)}\|_2 \leq (1+\gamma H)\sqrt{dt_k/\lambda}$.

We next show that $w^k_{(n)}$ is concentrated around $w^{k,*}_{(n)}$.

Lemma 5.2 *When* $\lambda = 1$, *it holds with probability at least* $1-\delta$ *that in each episode* k *and round* n,

$$\left|\left\langle \varphi(s,a), w^k_{(n)} - w^{k,*}_{(n)} \right\rangle\right| \leq \beta \|\varphi(s,a)\|_{\Sigma^{-1}_{t_k}}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ *where* $\beta = 16(1+H)d\sqrt{\log(1+dT/\delta)}$.

While Lemma 5.2 is an adaptation of (Lemma B.4, Jin et al., 2020) to the infinite-horizon setting, let us briefly explain the outline of its proof. The key step in providing the upper bound on $|\langle \varphi(s,a), w^k_{(n)} - w^{k,*}_{(n)} \rangle|$ is to consider the following quantity.

$$\left\langle \varphi(s,a), \Sigma^{-1}_{t_k} \sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau)\left(V^k_{(n)}(s_{\tau+1}) - \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_\tau,a_\tau)}\left[V^k_{(n)}(s')\right]\right) \right\rangle. \tag{3}$$

Here, the issue is that its expectation conditional on the history does not equal 0, because $V^k_{(n)}$ is not independent of (s_τ, a_τ) for $\tau \in [t_k]$. As a remedy, we take the covering number argument due to Jin et al. (2020) by considering the following class of value functions. We consider the class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the form

$$V(s) = \min\left\{\max_{a \in \mathcal{A}}\left\{\langle \varphi(s,a), w\rangle + m + \|\varphi(s,a)\|_G\right\}, M\right\}$$

where the parameters $(w, m, G, M) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^{d\times d} \times \mathbb{R}$ satisfy $\|w\|_2 \le (1+\gamma H)\sqrt{dt_k/\lambda}$, $|m| \le \gamma(1-\gamma)^{-1}$, $\|G\|_F \le \beta^2\sqrt{d}/\lambda$, and $|M| \le (1-\gamma)^{-1}$. Note that $V^k_{(n+1)}$ from line 8 of Algorithm 1 belongs to the class, because we may observe that

$$(w, m, G, M) = \left(w^k_{(n)},\ \min_{s'\in\mathcal{S}} V^k_{(n)}(s'),\ \beta^2\Sigma^{-1}_{t_k}, \min\left\{(1-\gamma)^{-1},\ \min_{s'\in\mathcal{S}} \widetilde{V}^k_{(n+1)}(s') + H\right\}\right)$$

satisfies the condition. Analyzing the covering number of the function class, we show the following lemma, based on which we provide an upper bound on term (3).

Lemma 5.3 *When $\lambda = 1$, it holds with probability at least $1-\delta$ that in each episode k and round n,*

$$\left\|\sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau)\left(V^k_{(n)}(s_{\tau+1}) - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_\tau,a_\tau)}\left[V^k_{(n)}(s')\right]\right)\right\|_{\Sigma^{-1}_{t_k}} \le 12(1+H)d\sqrt{\log\left(1+\frac{dT}{\delta}\right)}.$$

For the linear mixture MDP setting, recall that $\varphi^k_{(n)}(s,a) = \varphi(s,a) + \gamma\cdot\phi_{\bar{V}^k_{(n)}}(s,a)$ where $\bar{V}^k_{(n)} : \mathcal{S} \to \mathbb{R}$ is defined as $\bar{V}^k_{(n)}(s) = V^k_{(n)}(s) - \min_{s'\in\mathcal{S}} V^k_{(n)}(s')$ for $s \in \mathcal{S}$ (line 6 of Algorithm 1).

Lemma 5.4 *For each episode k and round n, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$\left\langle \varphi^k_{(n)}(s,a), \theta^* \right\rangle = r(s,a) + \gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[V^k_{(n)}(s')\right] - \gamma\min_{s'\in\mathcal{S}} V^k_{(n)}(s').$$

To estimate the true parameter θ^*, we use estimated parameters θ_t (line 14) and construct confidence ellipsoids containing θ^* with high probability, based on (Theorem 2, Abbasi-yadkori et al., 2011).

Lemma 5.5 *It holds with probability at least $1-\delta$ that the true parameter θ^* is contained in $\mathcal{C}_t$ for every $t \ge 1$ where*

$$\mathcal{C}_t = \{\theta \in \mathcal{B} : \|\theta - \theta_t\|_{\Sigma_t} \le \beta_t\}, \quad \beta_t = H\sqrt{d\log\left(\frac{1+t(B_\varphi + HB_\phi)^2/\lambda}{\delta}\right)} + B_\theta\sqrt{\lambda}.$$

Using Lemma 5.5, we deduce the following lemma, which is analogous to Lemma 5.2.

Lemma 5.6 *Suppose that $\theta^* \in \mathcal{C}_t$ for every $t \ge 1$. Then it holds that in every episode k and round n,*

$$\left|\left\langle \varphi^k_{(n)}(s,a), \theta_{t_k} - \theta^* \right\rangle\right| \le \beta_{t_k}\left\|\varphi^k_{(n)}(s,a)\right\|_{\Sigma^{-1}_{t_k}}$$

for any $(s,a) \in \mathcal{S}\times\mathcal{A}$.

5.2 Regret Decomposition

We consider $\text{Regret}(T) = T \cdot J^* - \sum_{t=1}^{T} r(s_t, a_t)$. In this section, provide a decomposition of the regret function based on our discussion from the previous section about parameter estimation. The main idea is to deduce an upper bound on the immediate reward $r(s_t, a_t)$ of each time step t.

For simplicity, we use notation τ_t for $t \in [t_k : t_{k+1} - 1]$ to denote $\xi_t(a_t)$ where $\xi_t(a_t)$ is given by a time index in $\text{argmax}_{n \in [T-t+1:N_k]} \widetilde{Q}^k_{(n)}(s_t, a_t)$ (line 11 of Algorithm 1). For the linear MDP case,

$$\begin{aligned} Q_t(s_t, a_t) &\leq \left\langle \varphi(s_t, a_t), w^k_{(\tau_t - 1)} \right\rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V^k_{(\tau_t - 1)}(s') + \beta \|\varphi(s_t, a_t)\|_{\Sigma^{-1}_{t_k}} \\ &\leq \left\langle \varphi(s_t, a_t), w^{k,*}_{(\tau_t - 1)} \right\rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V^k_{(\tau_t - 1)}(s') + 2\beta \|\varphi(s_t, a_t)\|_{\Sigma^{-1}_{t_k}} \end{aligned}$$

where the first inequality holds because $Q_t(s_t, a_t) = \widetilde{Q}^k_{(\tau_t)}(s_t, a_t)$ and the second inequality is from Lemma 5.2. Then Lemma 5.1 implies that

$$-r(s_t, a_t) \leq -Q_t(s_t, a_t) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[V^k_{(\tau_t - 1)}(s') \right] + 2\beta \|\varphi(s_t, a_t)\|_{\Sigma^{-1}_{t_k}}.$$

For the linear mixture MDP setting, we have

$$\begin{aligned} Q_t(s_t, a_t) &\leq \left\langle \varphi^k_{(\tau_t - 1)}(s_t, a_t), \theta_{t_k} \right\rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V^k_{(\tau_t - 1)}(s') + \beta_{t_k} \|\varphi^k_{(\tau_t - 1)}(s_t, a_t)\|_{\Sigma^{-1}_{t_k}} \\ &\leq \left\langle \varphi^k_{(\tau_t - 1)}(s_t, a_t), \theta^* \right\rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V^k_{(\tau_t - 1)}(s') + 2\beta_{t_k} \|\varphi^k_{(\tau_t - 1)}(s_t, a_t)\|_{\Sigma^{-1}_{t_k}} \end{aligned}$$

where the first inequality equality is due to $Q_t(s_t, a_t) = \widetilde{Q}^k_{(\tau_t)}(s_t, a_t)$ while the second inequality holds due to Lemma 5.6. Then it follows from Lemma 5.4 that

$$-r(s_t, a_t) \leq -Q_t(s_t, a_t) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[V^k_{(\tau_t - 1)}(s') \right] + 2\beta_{t_k} \left\| \varphi^k_{(\tau_t - 1)}(s_t, a_t) \right\|_{\Sigma^{-1}_{t_k}}.$$

Here, note that $\varphi^k_{(\tau_t - 1)}(s_t, a_t) = \varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t)$ because W_t (line 14) denotes $\bar{V}^k_{\tau_t - 1}$. Based on these upper bounds on $-r(s_t, a_t)$, we derive the following regret decomposition. Denoting by $K(T)$ the total number of distinct episodes over T steps,

$$\begin{aligned} &\text{Regret}(T) \left(= T \cdot J^* - \sum_{t=1}^{T} r(s_t, a_t) \right) \\ &\leq \underbrace{\sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \left(J^* - (1 - \gamma) V^k_{(\tau_t - 1)}(s_{t+1}) \right)}_{I_1} + \underbrace{\sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \left(V^k_{(\tau_t - 1)}(s_{t+1}) - Q_t(s_t, a_t) \right)}_{I_2} \\ &\quad + \underbrace{\gamma \sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \left(\mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t, a_t)} \left[V^k_{(\tau_t - 1)}(s') \right] - V^k_{(\tau_t - 1)}(s_{t+1}) \right)}_{I_3} + I_4 \end{aligned}$$

where

$$I_4 = \begin{cases} 2\gamma \sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \beta \|\varphi(s_t, a_t)\|_{\Sigma_{t_k}^{-1}}, & \text{(linear MDP)}, \\ 2\gamma \sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \beta_{t_k} \|\varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t)\|_{\Sigma_{t_k}^{-1}}, & \text{(linear mixture MDP)}. \end{cases}$$

5.3 Deriving Upper Bounds on the Regret Terms

In this section, we sketch how the regret terms I_1–I_4 can be bounded.

Regret Term I_1: Errors from Approximation to the discounted-reward MDP First, we show that $V_{(n)}^k$ and $\widetilde{Q}_{(n)}^k$ are optimistic estimators of V^* and Q^* for every episode k and round n.

Lemma 5.7 *Suppose that the statement of Lemma 5.2 holds for the linear MDP setting and that $\theta^* \in \mathcal{C}_t$ for every $t \geq 1$ for the linear mixture MDP setting. Then for both the linear MDP and linear mixture MDP settings, Algorithm 1 with $H \geq 2 \cdot \mathrm{sp}(v^*)$ guarantees that in every episode k and round n, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$V^*(s) \leq V_{(n)}^k(s) \leq (1-\gamma)^{-1}, \quad Q^*(s,a) \leq \widetilde{Q}_{(n)}^k(s,a) \leq (1-\gamma)^{-1}.$$

It follows from Lemma 5.7 that each term $J^* - (1-\gamma)V_{(\tau_t-1)}^k(s_{t+1})$ is at most $J^* - (1-\gamma)V^*(s_{t+1})$, which can be further bounded above based on the following lemma.

Lemma 5.8 (Lemma 2, Wei et al., 2020) *Let J^* and v^* be the optimal average reward and the optimal bias function given in* (1)*, and let V^* be the optimal discounted value function given in* (2) *with discount factor $\gamma \in [0, 1)$. Then it holds that*

$$\max_{s \in \mathcal{S}} |J^* - (1-\gamma)V^*(s)| \leq (1-\gamma)\mathrm{sp}(v^*), \quad \mathrm{sp}(V^*) \leq 2 \cdot \mathrm{sp}(v^*).$$

This lemma offers a tool to bridge an infinite-horizon average-reward MDP and a discounted-reward MDP. In particular, we deduce that $I_1 \leq (1-\gamma)\mathrm{sp}(v^*)T = \mathrm{sp}(v^*)\sqrt{T}$.

Regret Term I_2: Value Iteration and Max-Pooling Recall that we set $\tau_t = \xi_t(a_t) \in [T - t + 1 : N_k]$. Then for $t \in [t_k : t_{k+1} - 1)$, note that

$$V_{(\tau_t-1)}^k(s_{t+1}) \leq \widetilde{V}_{(\tau_t-1)}^k(s_{t+1}) = \max_{a \in \mathcal{A}} \widetilde{Q}_{(\tau_t-1)}^k(s_{t+1}, a) \leq \max_{a \in \mathcal{A}} \max_{n \in [T-t:N_k]} \widetilde{Q}_{(n)}^k(s_{t+1}, a)$$

Note that the right-most side can be rewritten as follows.

$$\max_{a \in \mathcal{A}} \max_{n \in [T-t:N_k]} \widetilde{Q}_{(n)}^k(s_{t+1}, a) = \max_{a \in \mathcal{A}} \widetilde{Q}_{(\xi_{t+1}(a))}^k(s_{t+1}, a) = \max_{a \in \mathcal{A}} Q_{t+1}(s_{t+1}, a).$$

This implies that $V_{(\tau_t-1)}^k(s_{t+1}) \leq Q_{t+1}(s_{t+1}, a_{t+1})$ for any $t \in [t_k : t_{k+1}-1)$. For $t = t_{k+1}-1$, we apply the bound $V_{\tau_t-1}^k(s_{t+1}) \leq (1-\gamma)^{-1}$ by Lemma 5.7. Then

$$I_2 \leq \sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-2} (Q_{t+1}(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)) + \sum_{k=1}^{K(T)} \left(\frac{1}{1-\gamma} - Q_{t_{k+1}-1}(s_{t_{k+1}-1}, a_{t_{k+1}-1}) \right).$$

Here, the right-hand side hand has a telescoping sum, and in fact, it equals $-\sum_{k=1}^{K(T)} Q_{t_k}(s_{t_k}, a_{t_k}) + K(T)(1-\gamma)^{-1}$. Therefore, we deduce that $I_2 \leq K(T)(1-\gamma)^{-1}$.

Regret Term I_3: martingale Difference Sequence We first observe that

$$I_3 = \gamma \sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \eta_t \quad \text{where} \quad \eta_t = \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t,a_t)} \left[W_t(s') \right] - W_t(s_{t+1})$$

because $W_t = V^k_{(\tau_t - 1)} - \min_{s' \in \mathcal{S}} V^k_{(\tau_t - 1)}(s')$. Here, we may argue that $\{\eta_t\}_{t=1}^{\infty}$ is a martingale difference sequence. Moreover, as $W_t(s) \in [0, H]$ for any $s \in \mathcal{S}$, we have $|\eta_t| \leq H$. Then, applying the Azuma-Hoeffding inequality, we deduce that $I_3 \leq H\sqrt{2T \log(1/\delta)}$ holds with probability at least $1 - \delta$. We provide a formal proof in Appendix C.2.

Regret Term I_4: Errors from Estimation of Unknown Parameters The last step is to consider the regret term I_4. Based on (Lemmas 10, 11 and 12, Abbasi-yadkori et al., 2011), we may prove the following for the linear MDP case.

Lemma 5.9 *When $\lambda = 1$, it holds that*

$$\sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \|\varphi(s_t, a_t)\|_{\Sigma_{t_k}^{-1}} \leq 2\sqrt{dT \log(1 + T/d)}.$$

For the linear mixture MDP setting, we have $\beta_{t_k} \leq \beta_T$ for any k. Then it is sufficient to prove the following result, which can be also deduced by applying the lemmas of Abbasi-yadkori et al. (2011).

Lemma 5.10 *Let $\lambda \geq (B_\varphi + HB_\phi)^2$, we have*

$$\sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \|\varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t)\|_{\Sigma_{t_k}^{-1}} \leq 2\sqrt{dT \log(1 + T/d)}.$$

Completing the Regret Bounds Combining the Upper Bounds on the Regret Terms I_1–I_4, we have

$$\text{Regret}(T) \leq (1 - \gamma)\text{sp}(v^*)T + \frac{K(T)}{1 - \gamma} + H\sqrt{2T \log(1/\delta)} + 2\widehat{\beta}\sqrt{dT \log(1 + T/d)}$$

where $\widehat{\beta} = \beta$ for the linear MDP case and $\widehat{\beta} = \beta_T$ for the linear mixture MDP case. The last ingredient is to upper bound the total number of episodes for running LSVI-DC.

Lemma 5.11 *Set $\lambda = 1$ for the linear MDP setting and $\lambda \geq (B_\varphi + HB_\phi)^2$ for the linear mixture MDP setting. Then $K(T) \leq 1 + d \log_2(1 + T/d)$.*

Plugging in the choice of parameters, we finally obtain the desired regret bounds for the linear MDP setting and the linear mixture MDP case (Appendices D and E).

6 Conclusion

This paper develops a provably efficient algorithm, LSVI-DC for learning infinite-horizon average-reward linear and linear mixture MDPs under the Bellman optimality condition. LSVI-DC is the first correct tractable algorithm that guarantees the best-known regret upper bound for the linear MDP setting, and for the linear mixture MDP setting, it provides the first regret upper bound under the Bellman optimality condition. Moreover, as a corollary, the regret upper bound for linear MDPs leads to the state-of-the-art regret bound for model-free tabular MDPs. Furthermore, we expect that some novel components of our algorithm, clipping and max-pooling, will be useful for infinite-horizon average-reward reinforcement under the Bellman optimality condition.
Although we provide tight regret bounds for both linear and linear mixture MDPs. There still exist gaps from the best-known lower bound. To close the gap, one direction is to explore variance-aware parameter estimation schemes, such as Berstein-type bounds and weighted ridge regression techniques deployed for deriving minimax optimal algorithms for the finite-horizon setting.

References

Y. Abbasi-yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Y. Abbasi-Yadkori, P. Bartlett, K. Bhatia, N. Lazic, C. Szepesvari, and G. Weisz. POLITEX: Regret bounds for policy iteration using expert prediction. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3692–3702. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lazic19a.html.

A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun. Reinforcement learning: Theory and algorithms, 2021. URL https://rltheorybook.github.io/.

A. Agarwal, Y. Jin, and T. Zhang. Voql: Towards optimal regret in model-free rl with nonlinear function approximation. In G. Neu and L. Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 987–1063. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/agarwal23a.html.

S. Agrawal and R. Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3621f1454cacf995530ea53652ddf8fb-Paper.pdf.

P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Ad-*

vances in Neural Information Processing Systems, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/e4a6222cdb5b34375400904f03d8e6a5-Paper.pdf.

A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang. Model-based reinforcement learning with value-targeted regression. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 463–474. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/ayoub20a.html.

P. L. Bartlett and A. Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 35–42, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

V. Boone and Z. Zhang. Achieving tractable minimax optimal regret in average reward mdps, 2024. URL https://arxiv.org/abs/2406.01234.

H. Bourel, O. Maillard, and M. S. Talebi. Tightening exploration in upper confidence reinforcement learning. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1056–1066. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bourel20a.html.

Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1283–1294. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/cai20d.html.

C. Dann, N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. On oracle-efficient pac rl with rich observations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5f0f5e5f33945135b874349cfbed4fb9-Paper.pdf.

S. Du, S. Kakade, J. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. Bilinear classes: A structural framework for provable generalization in rl. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2826–2836. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/du21a.html.

S. Filippi, O. Cappé, and A. Garivier. Optimism in reinforcement learning and kullback-leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122, 2010. doi: 10.1109/ALLERTON.2010.5706896.

R. Fruit, M. Pirotta, A. Lazaric, and R. Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1578–1586. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/fruit18a.html.

R. Fruit, M. Pirotta, and A. Lazaric. Improved analysis of ucrl2 with empirical bernstein inequality, 2020. URL https://arxiv.org/abs/2007.05456.

J. He, D. Zhou, and Q. Gu. Logarithmic regret for reinforcement learning with linear function approximation. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4171–4180. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/he21c.html.

J. He, H. Zhao, D. Zhou, and Q. Gu. Nearly minimax optimal reinforcement learning for linear Markov decision processes. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12790–12822. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/he23d.html.

J. He, H. Zhong, and Z. Yang. Sample-efficient learning of infinite-horizon average-reward MDPs with general function approximation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fq1wNrC2ai.

O. Hernandez-Lerma. *Adaptive Markov Control Processes*. Springer New York, NY, 2012. ISBN 0387969667.

K. Hong, Y. Zhang, and A. Tewari. Provably efficient reinforcement learning for infinite-horizon average-reward linear mdps, 2024. URL https://arxiv.org/abs/2405.15050.

P. Hu, Y. Chen, and L. Huang. Nearly minimax optimal reinforcement learning with linear function approximation. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8971–9019. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hu22a.html.

Z. Jia, L. Yang, C. Szepesvari, and M. Wang. Model-based reinforcement learning with value-targeted regression. In A. M. Bayen, A. Jadbabaie, G. Pappas, P. A. Parrilo, B. Recht, C. Tomlin, and M. Zeilinger, editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 666–686. PMLR, 10–11 Jun 2020. URL https://proceedings.mlr.press/v120/jia20a.html.

N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In D. Precup and Y. W.

Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/jiang17c.html.

C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013. doi: 10.1177/0278364913495721. URL https://doi.org/10.1177/0278364913495721.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236. URL https://doi.org/10.1038/nature14236.

A. Modi, N. Jiang, A. Tewari, and S. Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2010–2020. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/modi20a.html.

Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain. Learning unknown markov decision processes: A thompson sampling approach. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/51ef186e18dc00c2d31982567235c559-Paper.pdf.

M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017. doi: 10.1038/nature24270. URL https://doi.org/10.1038/nature24270.

W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In A. Beygelzimer and D. Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2898–2933. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/sun19a.html.

M. S. Talebi and O.-A. Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In F. Janoos, M. Mohri, and K. Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 770–805. PMLR, 07–09 Apr 2018. URL https://proceedings.mlr.press/v83/talebi18a.html.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices, 2011. URL https://arxiv.org/abs/1011.3027.

Y. Wang, R. Wang, S. S. Du, and A. Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CBmJwzneppz.

C.-Y. Wei, M. J. Jahromi, H. Luo, H. Sharma, and R. Jain. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10170–10180. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wei20c.html.

C.-Y. Wei, M. Jafarnia Jahromi, H. Luo, and R. Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3007–3015. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/wei21d.html.

G. Weisz, P. Amortila, and C. Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In V. Feldman, K. Ligett, and S. Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 1237–1264. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/weisz21a.html.

Y. Wu, D. Zhou, and Q. Gu. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3883–3913. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/wu22a.html.

L. Yang and M. Wang. Sample-optimal parametric q-learning using linearly additive features. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6995–7004. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/yang19b.html.

L. Yang and M. Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning*

Research, pages 10746–10756. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/yang20h.html.

E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020. doi: 10.1109/ACCESS.2020.2983149.

A. Zanette, D. Brandfonbrener, E. Brunskill, M. Pirotta, and A. Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1954–1964. PMLR, 26–28 Aug 2020a. URL https://proceedings.mlr.press/v108/zanette20a.html.

A. Zanette, A. Lazaric, M. Kochenderfer, and E. Brunskill. Learning near optimal policies with low inherent Bellman error. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10978–10989. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/zanette20a.html.

Z. Zhang and X. Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/9e984c108157cea74c894b5cf34efc44-Paper.pdf.

Z. Zhang and Q. Xie. Sharper model-free reinforcement learning for average-reward markov decision processes. In G. Neu and L. Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5476–5477. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/zhang23b.html.

D. Zhou and Q. Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36337–36349. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ebba182cb97864368fdb6ae00773a5e4-Paper-Conference.pdf.

D. Zhou, Q. Gu, and C. Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In M. Belkin and S. Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4532–4576. PMLR, 15–19 Aug 2021a. URL https://proceedings.mlr.press/v134/zhou21a.html.

D. Zhou, J. He, and Q. Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In M. Meila and T. Zhang, editors, *Proceedings of the*

38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 12793–12802. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/zhou21a.html.

Appendix A. Discussion of a Previous Algorithm with Clipping and Max-Pooling

In this section, we discuss the algorithm of (Hong et al., 2024, Algorithm 2) for linear MDPs. Although the original version of Algorithm 2 works under the assumption that the

Algorithm 2 γ-LSVI-UCB

1: **Input:** discount factor $\gamma \in (0,1)$, regularization parameter $\lambda > 0$, upper bound H of $2 \cdot \mathrm{sp}(v^*)$, bonus factor β
2: **Initialize:**initial state s_1, $t \leftarrow 1$, $\Sigma_1 \leftarrow \lambda I$, $V_1(\cdot) \leftarrow (1-\gamma)^{-1}$, $Q_1(\cdot,\cdot) \leftarrow (1-\gamma)^{-1}$
3: **for** time step $t = 1, 2, \cdots, T$ **do**
4: Take $a_t \in \operatorname{argmax}_{a \in \mathcal{A}} Q_t(s_t, a)$, receive $r(s_t, a_t)$, and obtain $s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$
5: Set $w_t \leftarrow \Sigma_{t_k}^{-1} \sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau)(r(s_\tau, a_\tau) + \gamma(V_t(s_{\tau+1}) - \min_{s' \in \mathcal{S}} V_t(s')))$
6: $\widetilde{Q}_{t+1}(\cdot,\cdot) \leftarrow \min\left\{ \langle \varphi(\cdot,\cdot), w_t \rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V_t(s') + \beta \|\varphi(\cdot,\cdot)\|_{\Sigma_{t_k}^{-1}}, (1-\gamma)^{-1} \right\}$
7: $\widetilde{V}_{t+1}(\cdot) \leftarrow \max_{a \in \mathcal{A}} \widetilde{Q}_{t+1}(\cdot, a)$
8: $V_{t+1}(\cdot) \leftarrow \min\left\{ \widetilde{V}_{t+1}(\cdot),\ \min_{s' \in \mathcal{S}} \widetilde{V}_{t+1}(s') + H \right\}$
9: $Q_{t+1}(\cdot,\cdot) \leftarrow \max_{\tau \in [t_k : t]} \widetilde{Q}_{\tau+1}(\cdot,\cdot)$
10: $\Sigma_{t+1} \leftarrow \Sigma_t + \varphi(s_t, a_t)\varphi(s_t, a_t)^\top$
11: **if** $\det(\Sigma_t) > 2 \det(\Sigma_{t_k})$ **then**
12: $k \leftarrow k + 1$
13: $t_k \leftarrow t + 1$
14: **end if**
15: **end for**

reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is known, it can also deal with the case where the reward function is unknown and parameterized as $r(s,a) = \langle \varphi(s,a), \theta \rangle$. The algorithm of Hong et al. (2024), γ-LSVI-UCB also runs with an approximation of the given infinite-horizon average-reward MDP by a discounted-reward MDP. Moreover, it has the clipping step (line 8) and the max-pooling step (line 9). Here, note that the max-pooling of γ-LSVI-UCB is different from our max-pooling step.

Hong et al. (2024) argued that the regret function can be decomposed as

$$\begin{aligned}
&\mathrm{Regret}(T) \left(= T \cdot J^* - \sum_{t=1}^{T} r(s_t, a_t) \right) \\
&\leq \underbrace{\sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \left(J^* - (1-\gamma) V_{\tau_t(s_t,a_t)}(s_{t+1}) \right)}_{(a)} + \underbrace{\sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \left(V_{\tau_t(s_t,a_t)}(s_{t+1}) - Q_{t+1}(s_t, a_t) \right)}_{(b)} \\
&\quad + \underbrace{\gamma \sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \left(\mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t,a_t)} \left[V_{\tau_t(s_t,a_t)}(s') \right] - V_{\tau_t(s_t,a_t)}(s_{t+1}) \right)}_{(c)} + \underbrace{\sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} 2\beta \|\varphi(s_t, a_t)\|_{\Sigma_t^{-1}}}_{(d)}
\end{aligned}$$

where $\tau_t(s_t, a_t)$ denotes the index such that $Q_{t+1}(s_t, a_t) = \widetilde{Q}_{\tau_t(s_t,a_t)+1}(s_t, a_t)$.
The issue is with the regret term (b). Hong et al. (2024) attempted to show that $V_{\tau_t(s_t,a_t)}(s_{t+1}) \leq Q_{t+1}(s_{t+1}, a_{t+1})$ by arguing that

$$V_{\tau_t(s_t,a_t)}(s_{t+1}) \leq \widetilde{V}_{\tau_t(s_t,a_t)}(s_{t+1}) = \max_{a \in \mathcal{A}} \widetilde{Q}_{\tau_t(s_t,a_t)}(s_{t+1}, a) \leq \max_{a \in \mathcal{A}} Q_{t+1}(s_{t+1}, a).$$

However, the second inequality does not necessarily hold because $\tau_t(s_t, a_t)$ can be t_k in which case $\widetilde{Q}_{\tau_t(s_t,a_t)}(s_{t+1}, a) = \widetilde{Q}_{t_k}(s_{t+1}, a)$ while $Q_{t+1}(s_{t+1}, a_{t+1}) = \max_{\tau \in [t_k+1, t+1]} \widetilde{Q}_\tau(s_{t+1}, a_{t+1})$. In such a case, $\tau_t(s_t, a_t) \notin [t_k+1, t+1]$, and thus we cannot compare $\max_{a \in \mathcal{A}} \widetilde{Q}_{\tau_t(s_t,a_t)}(s_{t+1}, a)$ and $Q_{t+1}(s_{t+1}, a_{t+1})$.
In fact, $V_{\tau_t(s_t,a_t)}(s_{t+1}) \leq Q_{t+1}(s_{t+1}, a_{t+1})$ does not necessarily hold true. Consider a situation where the value function V_{t_k} generates large values so that for some $\epsilon > 0$, we have $V_{t_k}(s) \geq \max_{a \in \mathcal{A}} \widetilde{Q}_{t+1}(s, a) + \epsilon$ for any $t \in [t_k, t_{k+1})$ and $s \in \mathcal{S}$. Hence, it is possible that $\tau_t(s_t, a_t)$ for $t \in [t_k, t_{k+1})$ is chosen to be t_k. In this case,

$$V_{\tau_t(s_t,a_t)}(s_{t+1}) = V_{t_k}(s_{t+1}) \geq \max_{a \in \mathcal{A}} \widetilde{Q}_{t+1}(s_{t+1}, a) + \epsilon \geq Q_{t+1}(s_{t+1}, a_{t+1}) + \epsilon.$$

This leads to a regret of ϵT from term (b).
Note that our algorithm has several differences compared to Algorithm 2. First, we have a separate planning phase where we run value iteration for many rounds. In contrast, Algorithm 2 runs a one-round value function update in each time time step. As a result, our max-pooling step is designed to be different from that of Algorithm 2. To remedy the issue of the analysis of Hong et al. (2024), we have a different regret decomposition. In particular, our regret term I_2 is a counterpart of term (b):

$$\sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \left(V^k_{(\tau_t - 1)}(s_{t+1}) - Q_t(s_t, a_t) \right).$$

In term I_2, we have $Q_t(s_t, a_t)$ while term (b) has $Q_{t+1}(s_t, a_t)$. With our design of the max-pooling operation, we can argue that $V^k_{(\tau_t - 1)}(s_{t+1}) \leq Q_{t+1}(s_{t+1}, a_{t+1})$, which naturally leads to a telescoping argument.

Appendix B. Proofs for the Concentration Results of Unknown Parameters

B.1 Proof of Lemma 5.1: Existence of a Coefficient Vector for an Action-Value Function

Note that

$$\begin{aligned}\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[V_{(n)}^k(s')\right]-\min_{s'\in\mathcal{S}}V_{(n)}^k(s') &= \int_{s'\in\mathcal{S}}\left(V_{(n)}^k(s')-\min_{s''\in\mathcal{S}}V_{(n)}^k(s'')\right)\mathbb{P}(ds'\mid s,a)\\ &= \int_{s'\in\mathcal{S}}\left(V_{(n)}^k(s')-\min_{s''\in\mathcal{S}}V_{(n)}^k(s'')\right)\left\langle\varphi(s,a),\mu(ds')\right\rangle\\ &= \left\langle\varphi(s,a),\int_{s'\in\mathcal{S}}\left(V_{(n)}^k(s')-\min_{s''\in\mathcal{S}}V_{(n)}^k(s'')\right)d\mu(s')\right\rangle.\end{aligned}$$

Since $r(s,a)=\langle\varphi(s,a),\theta\rangle$, taking

$$w_{(n)}^{k,*}=\theta+\gamma\int_{s'\in\mathcal{S}}\left(V_{(n)}^k(s')-\min_{s''\in\mathcal{S}}V_{(n)}^k(s'')\right)d\mu(s')$$

satisfies the desired condition. Note that

$$\left\|w_{(n)}^{k,*}\right\|_2\le\|\theta\|_2+\gamma\left\|\int_{s'\in\mathcal{S}}\left(V_{(n)}^k(s')-\min_{s''\in\mathcal{S}}V_{(n)}^k(s'')\right)d\mu(s')\right\|_2\le\sqrt{d}+\gamma H\|\mu(\mathcal{S})\|_2,$$

which implies that $\|w_{(n)}^{k,*}\|_2\le(1+\gamma H)\sqrt{d}$, as required.
Next we show that $\|w_{(n)}^k\|_2\le(1+\gamma H)\sqrt{dt_k/\delta}$. For ease of notation, set $w=w_{(n)}^k$, $u=w/\|w\|_2$, $x_\tau=\varphi(s_\tau,a_\tau)$, and $\Sigma_{t_k}=\lambda I+\sum_{\tau=1}^{t_k-1}x_\tau x_\tau^\top$, and $y_\tau=r(s_\tau,a_\tau)+\gamma(V_{(n)}^k(s_{\tau+1})-\min_{s'\in\mathcal{S}}V_{(n)}^k(s'))$. Then

$$\|w\|_2=\left|u^\top w\right|=\left|u^\top\Sigma_{t_k}^{-1}\sum_{\tau=1}^{t_k-1}x_\tau y_\tau\right|\le(1+\gamma H)\sum_{\tau=1}^{t_k-1}\left|u^\top\Sigma_{t_k}^{-1}x_\tau\right|$$

where the inequality holds because $y_\tau\le1+\gamma H$. The right-most side can be further bounded above as

$$(1+\gamma H)\sum_{\tau=1}^{t_k-1}\left|u^\top\Sigma_{t_k}^{-1}x_\tau\right|\le(1+\gamma H)\sum_{\tau=1}^{t_k-1}\sqrt{u^\top\Sigma_{t_k}^{-1}u}\sqrt{x_\tau^\top\Sigma_{t_k}^{-1}x_\tau}\le\frac{1+\gamma H}{\sqrt{\lambda}}\sum_{\tau=1}^{t_k-1}\sqrt{x_\tau^\top\Sigma_{t_k}^{-1}x_\tau}$$

where the first inequality follows from the Cauchy-Schwarz inequality and the second inequality holds because u is a unit vector and the maximum eigenvalue of $\Sigma_{t_k}^{-1}$ is at most $1/\lambda$. The last step is to bound apply the following lemma.

Lemma B.1 (Lemma D.1, Jin et al., 2020) *Let $x_1,\ldots,x_t\in\mathbb{R}^d$ and $A_t=\lambda I+\sum_{i=1}^t x_i x_i^\top$ for some $\lambda>0$. Then it holds that*

$$\sum_{i=1}^t x_i^\top A_t^{-1}x_i\le d.$$

Then we deduce that

$$\sum_{\tau=1}^{t_k-1}\sqrt{x_\tau^\top\Sigma_{t_k}^{-1}x_\tau}\le\sqrt{t_k-1}\sqrt{\sum_{\tau=1}^{t_k-1}x_\tau^\top\Sigma_{t_k}^{-1}x_\tau}\le\sqrt{t_kd}$$

where the first inequality is by the Cauchy-Schwarz inequality and the second inequality comes from Lemma B.1. Consequently, we deduce that

$$\left\|w_{(n)}^k\right\|_2=\|w\|_2\le(1+\gamma H)\sqrt{dt_k/\lambda},$$

as required.

B.2 Proof of Lemma 5.3: Self-Normalization Inequality

Recall that the ϵ-covering number $\mathcal{N}_\epsilon$ of a function class $\mathcal{V}$ is defined as the minimum size of an ϵ-cover $\widetilde{\mathcal{V}}$ of $\mathcal{V}$ that satisfies that for any $V\in\mathcal{V}$, there exists $\widetilde{V}\in\widetilde{\mathcal{V}}$ such that $\mathrm{dist}(V,\widetilde{V}):=\sup_{s\in\mathcal{S}}|V(s)-\widetilde{V}(s)|\le\epsilon$.

Lemma B.2 (Lemma D.4, Jin et al., 2020) *Let $\{x_\tau\}_{\tau=1}^\infty$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$. Let $\{\varphi_\tau\}_{\tau=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process where $\varphi_\tau\in\mathcal{F}_{\tau-1}$, and $\|\varphi_\tau\|_2\le1$. Let $\Sigma_{n+1}=\lambda I_d+\sum_{\tau=1}^n\varphi_\tau\varphi_\tau^\top$. Then for any $\delta>0$ and any given function class $\mathcal{V}$, with probability at least $1-\delta$, for all $k\in\mathbb{N}$, and any $V\in\mathcal{V}$ satisfying $sp(V)\le B$, we have*

$$\left\|\sum_{\tau=1}^n\varphi_\tau\left(V(x_\tau)-\mathbb{E}[V(x_\tau)|\mathcal{F}_{\tau-1}]\right)\right\|_{\Sigma_{n+1}^{-1}}^2\le4B^2\left(\frac{d}{2}\log\left(\frac{n+\lambda}{\lambda}\right)+\log\frac{\mathcal{N}_\epsilon}{\delta}\right)+\frac{8n^2\epsilon^2}{\lambda}$$

where $\mathcal{N}_\epsilon$ is the ϵ-covering number of $\mathcal{V}$ with respect to the distance $\mathrm{dist}(V,V')=\sup_{s\in\mathcal{S}}|V(s)-V'(s)|$ for $V,V'\in\mathcal{V}$.

In particular, we consider the class $\mathcal{V}$ of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ defined as

$$\mathcal{V}=\left\{V:\mathcal{S}\to\mathbb{R}:\begin{array}{l}V(\cdot)=\min\left\{\max\limits_{a\in\mathcal{A}}\left\{\langle\varphi(\cdot,a),w\rangle+m+\|\varphi(\cdot,a)\|_G\right\},M\right\}\\ \|w\|_2\le(1+\gamma H)\sqrt{dt_k/\lambda},\ |m|\le\gamma(1-\gamma)^{-1},\\ \|G\|_F\le\beta^2\sqrt{d}/\lambda,\ |M|\le(1-\gamma)^{-1}\end{array}\right\}\tag{4}$$

For this specific choice of $\mathcal{V}$, we can prove the following bound on the ϵ-covering number.

Lemma B.3 *The ϵ-covering number $\mathcal{N}_\epsilon$ of the class $\mathcal{V}$ defined as in Equation (4) is bounded above as*

$$\mathcal{N}_\epsilon\le\left(1+\frac{8(1+H)\sqrt{dt_k}}{\epsilon\sqrt{\lambda}}\right)^d\left(1+\frac{32\beta^2\sqrt{d}}{\lambda\epsilon^2}\right)^{d^2}\left(1+\frac{8}{\epsilon(1-\gamma)}\right)^2.$$

Proof. Let $V_1, V_2 \in \mathcal{V}$. Then there exist (w_1, m_1, G_1, M_1) and (w_2, m_2, G_2, M_2) which satisfy the condition in the description of (4) and correspond to V_1 and V_2, respectively. Let us define q_1 and q_2 as

$$\begin{aligned} q_1(s,a) &= \langle \varphi(s,a), w_1 \rangle + m_1 + \|\varphi(s,a)\|_{G_1}, \\ q_2(s,a) &= \langle \varphi(s,a), w_2 \rangle + m_2 + \|\varphi(s,a)\|_{G_2}. \end{aligned}$$

Then it follows that

$$\begin{aligned} \mathrm{dist}(V_1, V_2) &= \sup_{s\in\mathcal{S}} \left| \min\left\{ \max_{a\in\mathcal{A}} q_1(s,a), M_1 \right\} - \min\left\{ \max_{a\in\mathcal{A}} q_2(s,a), M_2 \right\} \right| \\ &\le \sup_{s\in\mathcal{S}} \left| \max_{a\in\mathcal{A}} q_1(s,a) - \max_{a\in\mathcal{A}} q_2(s,a) \right| + |M_1 - M_2| \\ &\le \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |q_1(s,a) - q_2(s,a)| + |M_1 - M_2| \\ &\le \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left\{ |\langle \varphi(s,a), w_1 - w_2 \rangle| + \left| \|\varphi(s,a)\|_{G_1} - \|\varphi(s,a)\|_{G_2} \right| \right\} \\ &\quad + |m_1 - m_2| + |M_1 - M_2| \\ &\le \sup_{\varphi:\|\varphi\|_2\le 1} |\langle \varphi, w_1 - w_2 \rangle| + \sup_{\varphi:\|\varphi\|_2\le 1} \sqrt{|\varphi^\top (G_1 - G_2)\varphi|} \\ &\quad + |m_1 - m_2| + |M_1 - M_2| \\ &= \|w_1 - w_2\|_2 + \sqrt{\|G_1 - G_2\|_2} + |m_1 - m_2| + |M_1 - M_2| \\ &\le \|w_1 - w_2\|_2 + \sqrt{\|G_1 - G_2\|_F} + |m_1 - m_2| + |M_1 - M_2| \end{aligned} \tag{5}$$

where the first inequality employs $|\min\{a, M_1\} - \min\{b, M_2\}| \le |a - b| + |M_1 - M_2|$, the second inequality is because taking the maximum is a contraction operator, the third one is by the triangle inequality, the fourth one comes from the assumption that $\|\varphi(s,a)\|_2 \le 1$ and the fact that $\left|\sqrt{x} - \sqrt{y}\right| \le \sqrt{|x-y|}$ for any $x, y \ge 0$, and the fifth inequality holds due to the definition of the ℓ_2-norm, and the last inequality holds because $\|G_1 - G_2\|_2 \le \|G_1 - G_2\|_F$. Let $\mathcal{C}_w$ be a minimum $\epsilon/4$-cover of $\{w \in \mathbb{R}^d : \|w\|_2 \le (1+\gamma H)\sqrt{dt_k/\lambda}\}$ with respect to the ℓ_2-norm, let $\mathcal{C}_m$ be a minimum $\epsilon/4$-cover of $\{C \in \mathbb{R} : |m| \le \gamma(1-\gamma)^{-1}\}$, let $\mathcal{C}_G$ be a minimum $\epsilon^2/16$-cover of $\{G \in \mathbf{R}^{d\times d} : \|G\|_F \le \beta^2\sqrt{d}/\lambda\}$ with respect to the Frobenius norm, and let $\mathcal{C}_M$ be a minimum $\epsilon/4$-cover of $\{M \in \mathbb{R} : |M| \le (1-\gamma)^{-1}\}$. It is a classical result that for any $\epsilon > 0$, the ϵ-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is bounded above by $(1 + 2R/\epsilon)^d$ (Lemma 5.2, Vershynin, 2011). Then it follows that

$$\begin{aligned} |\mathcal{C}_w| &\le \left(1 + 8(1+\gamma H)\sqrt{dt_k/\lambda}/\epsilon\right)^d, \\ |\mathcal{C}_m| &\le \left(1 + 8\gamma(1-\gamma)^{-1}/\epsilon\right), \\ |\mathcal{C}_G| &\le \left(1 + 32\beta^2\sqrt{d}/(\lambda\epsilon^2)\right)^{d^2}, \\ |\mathcal{C}_M| &\le \left(1 + 8(1-\gamma)^{-1}/\epsilon\right). \end{aligned}$$

By Equation (5), the set of functions V parameterized by $(w, m, G, M) \in \mathcal{C}_w \times \mathcal{C}_m \times \mathcal{C}_G \times \mathcal{C}_G$ is an ϵ-cover of $\mathcal{V}$. This implies that $|\mathcal{N}_\epsilon| \leq |\mathcal{C}_w| \cdot |\mathcal{C}_m| \cdot |\mathcal{C}_G| \cdot |\mathcal{C}_G|$. Therefore,

$$\mathcal{N}_\epsilon \leq \left(1 + \frac{8(1+\gamma H)\sqrt{dt_k}}{\epsilon\sqrt{\lambda}}\right)^d \left(1 + \frac{8\gamma(1-\gamma)^{-1}}{\epsilon}\right) \left(1 + \frac{32\beta^2\sqrt{d}}{\lambda\epsilon^2}\right)^{d^2} \left(1 + \frac{8(1-\gamma)^{-1}}{\epsilon}\right).$$

Since $\gamma \leq 1$, we have

$$\mathcal{N}_\epsilon \leq \left(1 + \frac{8(1+H)\sqrt{dt_k}}{\epsilon\sqrt{\lambda}}\right)^d \left(1 + \frac{32\beta^2\sqrt{d}}{\lambda\epsilon^2}\right)^{d^2} \left(1 + \frac{8}{\epsilon(1-\gamma)}\right)^2,$$

as required. ■

As $\lambda = 1$, by Lemmas B.2 and B.3 with $\epsilon = (1+H)d/t_k$, we have

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau)\left(V_{(n)}^k(s_{\tau+1}) - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_\tau,a_\tau)}\left[V_{(n)}^k(s')\right]\right)\right\|_{\Sigma_{t_k}^{-1}}^2 \\
&\leq 4H^2\left(\frac{d}{2}\log(t_k+1) + d\log\left(1 + \frac{8t_k^{3/2}}{\sqrt{d}}\right) + d^2\log\left(1 + \frac{32\beta^2 t_k^2}{(1+H)^2 d^{3/2}}\right)\right. \\
&\qquad \left. + 2\log\left(1 + \frac{8t_k(1-\gamma)^{-1}}{(1+H)d}\right) + \log\left(\frac{1}{\delta}\right)\right) + 8(1+H)^2 d^2 \qquad (6) \\
&\leq 4H^2\left(\frac{d}{2}\log(1+T) + d\log\left(1 + \frac{8T^{3/2}}{\sqrt{d}}\right) + d^2\log\left(1 + \frac{32\beta^2 T^2}{(1+H)^2 d^{3/2}}\right)\right. \\
&\qquad \left. + 2\log\left(1 + \frac{8T^{3/2}}{(1+H)d}\right) + \log\left(\frac{1}{\delta}\right)\right) + 8(1+H)^2 d^2
\end{aligned}$$

where the second inequality holds because $\gamma = 1 - 1/\sqrt{T}$ and $t_k \leq T$. Note that

$$\begin{aligned}
\log\left(1 + \frac{8T^{3/2}}{\sqrt{d}}\right) &\leq \frac{7}{2}\log(1+T) + \log(1+T) \\
&\leq \frac{9}{2}\log(1+T)
\end{aligned} \qquad (7)$$

where the first inequality holds because $8\sqrt{T} \leq (1+T)^{7/2}$ and $d \geq 1$. Next,

$$\log\left(1 + \frac{8T^{3/2}}{(1+H)d}\right) \leq \log\left(1 + \frac{8T^{3/2}}{\sqrt{d}}\right) \leq \frac{9}{2}\log(1+T) \qquad (8)$$

where the first inequality holds because $d \geq 1$ while the second inequality follows from Equation (7). Combining Equations (6) to (8), we deduce that

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau)\left(V_{(n)}^k(s_{\tau+1}) - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_\tau,a_\tau)}\left[V_{(n)}^k(s')\right]\right)\right\|_{\Sigma_{t_k}^{-1}}^2 \\
&\leq 4H^2\left(14d\log(1+T) + d^2\log\left(1 + \frac{32\beta^2 T^2}{(1+H)^2 d^{3/2}}\right) + \log\left(\frac{1}{\delta}\right)\right) + 8(1+H)^2 d^2.
\end{aligned} \qquad (9)$$

Recall that our choice of β is given by

$$\beta = 16(1+H)d\sqrt{\log\left(1+\frac{dT}{\delta}\right)}.$$

Note that

$$\frac{32\beta^2T^2}{(1+H)^2d^{3/2}} \leq 2^{13}\sqrt{d}T^2\left(\frac{dT}{\delta}\right) = \frac{2^{13}d^{3/2}T^3}{\delta}$$

where the inequality holds because $\log(1+x) \leq x$ for any $x \geq 0$. Moreover,

$$\begin{aligned}\log\left(1+\frac{2^{13}d^{3/2}T^3}{\delta}\right) &\leq 15\log\left(1+\frac{dT}{\delta}\right)+\log\left(1+\frac{dT}{\delta}\right)\\ &\leq 16\log\left(1+\frac{dT}{\delta}\right)\end{aligned} \tag{10}$$

because

$$2^{13}\sqrt{d}T^2 \leq 2^{13}d^2T^2 \leq \left(1+\frac{dT}{\delta}\right)^{15}.$$

Combining Equations (9) and (10), we obtain

$$\begin{aligned}&\left\|\sum_{\tau=1}^{t_k-1}\varphi(s_\tau,a_\tau)\left(V_{(n)}^k(s_{\tau+1})-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_\tau,a_\tau)}\left[V_{(n)}^k(s')\right]\right)\right\|^2_{\Sigma_{t_k}^{-1}}\\ &\leq 4H^2\left(14d\log(1+T)+16d^2\log\left(1+\frac{dT}{\delta}\right)+\log\left(\frac{1}{\delta}\right)\right)+8(1+H)^2d^2\\ &\leq 136(1+H)^2d^2\log\left(1+\frac{dT}{\delta}\right)\end{aligned}$$

where the second inequality holds because $\log 2 \geq 0.69$, as required.

B.3 Proof of Lemma 5.2: Concentration of Coefficient Vectors

Note that

$$\begin{aligned}&\left\langle\varphi(s,a),w_{(n)}^k\right\rangle\\ &=\underbrace{\left\langle\varphi(s,a),\Sigma_{t_k}^{-1}\sum_{\tau=1}^{t_k-1}\varphi(s_\tau,a_\tau)\left(r(s_\tau,a_\tau)+\gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_\tau,a_\tau)}\left[V_{(n)}^k(s')\right]-\gamma\min_{s'\in\mathcal{S}}V_{(n)}^k(s')\right)\right\rangle}_{(a)}\\ &\quad+\gamma\underbrace{\left\langle\varphi(s,a),\Sigma_{t_k}^{-1}\sum_{\tau=1}^{t_k-1}\varphi(s_\tau,a_\tau)\left(V_{(n)}^k(s_{\tau+1})-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_\tau,a_\tau)}\left[V_{(n)}^k(s')\right]\right)\right\rangle}_{(b)}.\end{aligned}$$

Here, the term (a) can be rewritten

$$\begin{aligned}\text{Term } (a) &= \left\langle \varphi(s,a), \Sigma_{t_k}^{-1} \sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau)\varphi(s_\tau, a_\tau)^\top w_{(n)}^{k,*} \right\rangle \\ &= \left\langle \varphi(s,a), w_{(n)}^{k,*} \right\rangle - \lambda \left\langle \varphi(s,a), \Sigma_{t_k}^{-1} w_{(n)}^{k,*} \right\rangle\end{aligned}$$

where the first equality is by Lemma 5.1 while the second equality is by the definition of Σ_{t_k}. This implies that

$$\begin{aligned}\left|\left\langle \varphi(s,a), w_{(n)}^{k} - w_{(n)}^{k,*} \right\rangle\right| &\leq \lambda \left\langle \varphi(s,a), \Sigma_{t_k}^{-1} w_{(n)}^{k,*} \right\rangle + \gamma \left|\text{term } (b)\right| \\ &\leq \lambda \left\|\varphi(s,a)\right\|_{\Sigma_{t_k}^{-1}} \left\|w_{(n)}^{k,*}\right\|_{\Sigma_{t_k}^{-1}} + \gamma \left|\text{term } (b)\right| \\ &\leq \sqrt{\lambda} \left\|\varphi(s,a)\right\|_{\Sigma_{t_k}^{-1}} \left\|w_{(n)}^{k,*}\right\|_{2} + \gamma \left|\text{term } (b)\right| \\ &\leq H\sqrt{d\lambda} \left\|\varphi(s,a)\right\|_{\Sigma_{t_k}^{-1}} + \gamma \left|\text{term } (b)\right|\end{aligned}$$

where the second inequality is due to the Cauchy-Schwarz inequality, the third one holds because $\Sigma_{t_k}^{-1} \preceq (1/\lambda) I_d$ where I_d is the $d \times d$ identity matrix, and the last inequality is due to Lemma 5.1.

For the term (b), we apply Lemma 5.3 as follows.

$$\begin{aligned}\left|\text{Term (b)}\right| &\leq \left\|\varphi(s,a)\right\|_{\Sigma_{t_k}^{-1}} \left\| \sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau) \left(V_{(n)}^{k}(s_{\tau+1}) - \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_\tau, a_\tau)} \left[V_{(n)}^{k}(s') \right] \right) \right\|_{\Sigma_{t_k}^{-1}} \\ &\leq 12(1+H)d \sqrt{\log\left(1 + \frac{dT}{\delta}\right)} \left\|\varphi(s,a)\right\|_{\Sigma_{t_k}^{-1}}\end{aligned}$$

where the first inequality is due to the Cauchy-Schwarz inequality while the second one follows from Lemma 5.3. Then we deduce that

$$\left|\left\langle \varphi(s,a), w_{(n)}^{k} - w_{(n)}^{k,*} \right\rangle\right| \leq 14(1+H)d \sqrt{\log\left(1 + \frac{dT}{\delta}\right)} \left\|\varphi(s,a)\right\|_{\Sigma_{t_k}^{-1}}$$

because $\lambda = 1$ and $\log 2 \geq 0.69$.

B.4 Proof of Lemma 5.4: Property of the True Coefficient Vector

Recall that the true parameter θ^* induces that

$$\langle \phi(s,a,s'), \theta^* \rangle = \mathbb{P}(s' \mid s,a).$$

Note that for any function $V : \mathcal{S} \to \mathbb{R}$, we have

$$\begin{aligned}\langle \phi_V(s,a), \theta^* \rangle &= \left\langle \int_{s'} \phi(s,a,s') V(s') ds', \theta^* \right\rangle \\ &= \int_{s'} \left\langle \phi(s,a,s'), \theta^* \right\rangle V(s') ds' \\ &= \int_{s'} \mathbb{P}(s' \mid s,a) V(s') ds' \\ &= \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[V(s') \right].\end{aligned}$$

This implies that

$$\begin{aligned}\left\langle \varphi^k_{(n)}(s,a), \theta^* \right\rangle &= \langle \varphi(s,a), \theta^* \rangle + \gamma \left\langle \phi_{\bar{V}^k_{(n)}}(s,a), \theta^* \right\rangle \\ &= r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[V^k_{(n)}(s') - \min_{s'' \in \mathcal{S}} V^k_{(n)}(s'') \right] \\ &= r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[V^k_{(n)}(s') \right] - \gamma \min_{s'' \in \mathcal{S}} V^k_{(n)}(s''),\end{aligned}$$

as required.

B.5 Proof of Lemma 5.5: Confidence Ellipsoids for the Coefficient Vector

Lemma 5.5 is an immediate consequence of the following result.

Lemma B.4 (Theorem 2, Abbasi-yadkori et al., 2011) *Let $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$ be a filtration, and let $\{\eta_\tau\}_{\tau=1}^\infty$ be a real-valued stochastic process such that η_τ is $\mathcal{F}_\tau$-measurable and η_τ is conditionally R-sub-Gaussian for some $R > 0$. Moreover, let $\{\varphi_\tau\}_{\tau=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that φ_τ is $\mathcal{F}_{\tau-1}$-measurable and $\|\varphi_\tau\| \leq B_\varphi$. Let $\Sigma_{\tau+1} = \lambda I_d + \sum_{n=1}^\tau \varphi_n \varphi_n^\top$. Define $y_\tau = \langle \varphi_\tau, \theta^* \rangle + \eta_\tau$, and assume that $\|\theta^*\|_2 \leq B_\theta$. Then it holds with probability at least $1-\delta$ that for all $\tau \geq 1$,*

$$\theta^* \in \left\{ \theta \in \mathbb{R}^d : \ \|\theta - \theta_\tau\|_{\Sigma_\tau} \leq R\sqrt{d \log\left(\frac{1 + \tau B_\varphi^2/\lambda}{\delta} \right)} + B_\theta \sqrt{\lambda} \right\}$$

where

$$\theta_\tau = \Sigma_\tau^{-1} \sum_{n=1}^{\tau-1} \varphi_n y_n.$$

To apply Lemma B.4, we take

$$\begin{aligned}y_\tau &= r(s_\tau, a_\tau) + \gamma W_\tau(s_{\tau+1}), \\ \varphi_\tau &= \varphi(s_\tau, a_\tau) + \gamma \cdot \phi_{W_\tau}(s_\tau, a_\tau), \\ \eta_\tau &= r(s_\tau, a_\tau) + \gamma W_\tau(s_{\tau+1}) - \langle \varphi(s_\tau, a_\tau) + \gamma \cdot \phi_{W_\tau}(s_\tau, a_\tau), \theta^* \rangle \\ &= \gamma \left(W_\tau(s_{\tau+1}) - \langle \phi_{W_\tau}(s_\tau, a_\tau), \theta^* \rangle \right).\end{aligned}$$

Here, note that

$$\langle \phi_{W_\tau}(s_\tau, a_\tau), \theta^* \rangle = \sum_{s' \in \mathcal{S}} {\theta^*}^\top \phi(s_\tau, a_\tau, s') W_\tau(s') = \mathbb{E}_{s' \sim \mathbb{P}(\cdot | s_\tau, a_\tau)} \left[W_\tau(s') \right]$$

where the second equality holds because ${\theta^*}^\top \phi(s_\tau, a_\tau, s') = \mathbb{P}(s' \mid s_\tau, a_\tau)$. Then it follows that

$$\eta_\tau = \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot | s_\tau, a_\tau)} \left[W_\tau(s_{\tau+1}) - W_\tau(s') \right],$$

and therefore, η_τ is H-sub-Gaussian. Moreover, we take

$$\begin{aligned}
\Sigma_{\tau+1} &= \lambda I_d + \sum_{n=1}^{\tau} \left(\varphi(s_n, a_n) + \gamma \cdot \phi_{W_n}(s_n, a_n) \right) \left(\varphi(s_n, a_n) + \gamma \cdot \phi_{W_n}(s_n, a_n) \right)^\top \\
&= \lambda I_d + \sum_{n=1}^{\tau} \varphi_n \varphi_n^\top, \\
\theta_{\tau+1} &= \Sigma_{\tau+1}^{-1} \sum_{n=1}^{\tau} \left(\varphi(s_n, a_n) + \gamma \cdot \phi_{W_n}(s_n, a_n) \right) \left(r(s_n, a_n) + \gamma W_n(s_{n+1}) \right) \\
&= \Sigma_{\tau+1}^{-1} \sum_{n=1}^{\tau} \varphi_n y_n.
\end{aligned}$$

Then it follows from Lemma B.4 that for all $t \geq 1$, the true parameter θ^* belongs to

$$\mathcal{C}_t := \left\{ \theta \in \mathcal{B} : \ \| \theta - \theta_t \|_{\Sigma_t} \leq \beta_t \right\}$$

where

$$\beta_t = H \sqrt{d \log \left(\frac{1 + t(B_\varphi + H B_\phi)^2 / \lambda}{\delta} \right)} + B_\theta \sqrt{\lambda},$$

as required.

B.6 Proof of Lemma 5.6: Bounding the Errors from Extended Value Iteration

Note that

$$\begin{aligned}
\left| \left\langle \varphi_{(n)}^k(s, a), \theta_{t_k} - \theta^* \right\rangle \right| &= \left| \left\langle \varphi(s, a) + \gamma \cdot \phi_{\bar{V}_{(n)}^k}(s, a), \theta_{t_k} - \theta^* \right\rangle \right| \\
&\leq \left\| \varphi(s, a) + \gamma \cdot \phi_{\bar{V}_{(n)}^k}(s, a) \right\|_{\Sigma_{t_k}^{-1}} \| \theta_{t_k} - \theta^* \|_{\Sigma_{t_k}} \\
&\leq \beta_{t_k} \left\| \varphi(s, a) + \gamma \cdot \phi_{\bar{V}_{(n)}^k}(s, a) \right\|_{\Sigma_{t_k}^{-1}}
\end{aligned}$$

where the first inequality is by applying the Cauchy-Schwarz inequality and the second one is due to Lemma 5.5, as required.

Appendix C. Common Lemmas

In this section, we provide and prove some results that are necessary to prove Theorems 1 and 2.

C.1 Proof of Lemma 5.7: Optimistic Estimators for Value Functions

For a fixed episode k, we prove the statement by induction on n. For $n = 1$, it is trivial that

$$V_{(1)}^k = \frac{1}{1-\gamma} \geq V^*(s), \quad \widetilde{Q}_{(1)}^k(s,a) = \frac{1}{1-\gamma} \geq Q^*(s,a)$$

for every $(s,a) \in \mathcal{S} \times \mathcal{A}$. Next, we assume that for $\tau = n$, the inequalities

$$\frac{1}{1-\gamma} \geq V_{(\tau)}^k(s) \geq V^*(s) \quad \text{and} \quad \frac{1}{1-\gamma} \geq \widetilde{Q}_{(\tau)}^k(s,a) \geq Q^*(s,a)$$

hold for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. First of all, by the clipping operation by $(1-\gamma)^{-1}$, we automatically have $\widetilde{Q}_{(n+1)}^k(s,a) \leq (1-\gamma)^{-1}$. Then

$$V_{(n+1)}^k(s) \leq \widetilde{V}_{(n+1)}^k(s) = \max_{a \in \mathcal{A}} \widetilde{Q}_{(n+1)}^k(s,a) \leq (1-\gamma)^{-1}.$$

Next, we show that $\widetilde{Q}_{(n+1)}^k(s,a) \geq Q*(s,a)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. For the linear MDP case,

$$\begin{aligned}
\widetilde{Q}_{(n+1)}^k(s,a) &= \min\left\{\left\langle \varphi(s,a), w_{(n)}^k \right\rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V_{(n)}^k(s') + \beta \|\varphi(s,a)\|_{\Sigma_{t_k}^{-1}}, (1-\gamma)^{-1}\right\} \\
&\geq \min\left\{\left\langle \varphi(s,a), w_{(n)}^{k,*} \right\rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V_{(n)}^k(s'), (1-\gamma)^{-1}\right\} \\
&= \min\left\{r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}\left[V_{(n)}^k(s')\right], (1-\gamma)^{-1}\right\} \\
&\geq \min\left\{r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}\left[V^*(s')\right], (1-\gamma)^{-1}\right\} \\
&= Q^*(s,a)
\end{aligned}$$

where the first inequality comes from Lemma 5.2, the second equality is by Lemma 5.1, the second inequality is due to the induction hypothesis that $V_{(n)}^k(s') \geq V^*(s')$ for any $s' \in \mathcal{S}$, and the last equality follows from the fact that $Q^*(s,a) \leq (1-\gamma)^{-1}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and the Bellman optimality equation (2). For the linear mixture MDP case,

$$\begin{aligned}
\widetilde{Q}_{(n+1)}^k(s,a) &= \min\left\{\left\langle \varphi_{(n)}^k(\cdot,\cdot), \theta_{t_k} \right\rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V_{(n)}^k(s') + \beta_{t_k} \|\varphi_{(n)}^k(\cdot,\cdot)\|_{\Sigma_{t_k}^{-1}}, (1-\gamma)^{-1}\right\} \\
&\geq \min\left\{\left\langle \varphi_{(n)}^k(\cdot,\cdot), \theta^* \right\rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V_{(n)}^k(s'), (1-\gamma)^{-1}\right\} \\
&= \min\left\{r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}\left[V_{(n)}^k(s')\right], (1-\gamma)^{-1}\right\} \\
&\geq \min\left\{r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}\left[V^*(s')\right], (1-\gamma)^{-1}\right\} \\
&= Q^*(s,a)
\end{aligned}$$

where the first inequality comes from Lemma 5.6, the second equality is by Lemma 5.4, the second inequality is due to the induction hypothesis that $V_{(n)}^k(s') \geq V^*(s')$ for any $s' \in \mathcal{S}$, and the last equality follows from the fact that $Q^*(s,a) \leq (1-\gamma)^{-1}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and the Bellman optimality equation (2).

Let us also consider $\widetilde{V}^k_{(n+1)}$. Let $a^*_s = \text{argmax}_{a\in\mathcal{A}}\, Q^*(s,a)$. Then

$$\begin{aligned}\widetilde{V}^k_{(n+1)}(s) - V^*(s) &= \max_{a\in\mathcal{A}} \widetilde{Q}^k_{(n+1)}(s,a) - Q^*(s,a^*_s)\\ &\geq \widetilde{Q}^k_{(n+1)}(s,a^*_s) - Q^*(s,a^*_s)\\ &\geq 0,\end{aligned}$$

implying in turn that $\widetilde{V}^k_{(n+1)}(s) \geq V^*(s)$ for any $s\in\mathcal{S}$. This further implies that

$$\begin{aligned}V^k_{(n+1)}(s) &= \min\left\{\widetilde{V}^k_{(n+1)}(s), \min_{s'\in\mathcal{S}} \widetilde{V}^k_{(n+1)}(s') + H\right\}\\ &\geq \min\left\{V^*(s), \min_{s'\in\mathcal{S}} V^*(s') + H\right\}\\ &= V^*(s),\end{aligned}$$

where the first inequality comes from our observation that $\widetilde{V}^k_{(n+1)}(s) \geq V^*(s)$ for any $s\in\mathcal{S}$ while the second equality holds because $\text{sp}(V^*) \leq 2{\cdot}\text{sp}(v^*) \leq H$, as supported by Lemma 5.8. Since k was chosen arbitrarily, we conclude that in every episode, for all $n\in[N]$ and for all $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$V^k_{(n)}(s) \geq V^*(s), \quad \widetilde{Q}^k_{(n)}(s,a) \geq Q^*(s,a),$$

as required.

C.2 Regret Term I_3

The term I_3 is a sum of martingale difference sequence $\{\eta_t\}_{t=1}^{\infty}$ with regard to a filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$, where

$$\eta_t = \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}\left[V^k_{(\tau_t-1)}(s_t,a_t)\right] - V^k_{(\tau_t-1)}(s_{t+1})$$

and $\mathcal{F}_t = \sigma(s_1,a_1,\cdots,s_t,a_t,s_{t+1})$ for $t\in[t_k : t_{k+1}-1]$. This is because η_t is $\mathcal{F}_t$-measurable, $\mathbb{E}[|\eta_t|]<\infty$, and $\mathbb{E}[\eta_t|\mathcal{F}_{t-1}]=0$, which we will show in the following paragraphs.
The first condition holds, because all the randomnesses in η_t, which are $s_t, a_t, \tau_t, s_{t+1}$, are captured given the filtration $\mathcal{F}_t$.
For the second condition, as $W_t(\cdot) = V^k_{(\tau_t-1)}(\cdot) - \min_{s'\in\mathcal{S}} V^k_{(\tau_t-1)}(s')$ for $t\in[t_k : t_{k+1}-1]$, we can rewrite η_t as

$$\eta_t = \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}\left[W_t(s_t,a_t)\right] - W_t(s_{t+1}).$$

As we constrained the span of $V^k_{(n)}$ in Algorithm 1 to be within H, $W_t(s)$ lies in $[0,H]$ for all $s\in\mathcal{S}$. Therefore, $|\eta_t|\leq H<\infty$, which proves the second condition.
Finally, the last condition can be checked as follows. Given $\mathcal{F}_{t-1}$, s_t is fixed, which naturally determines the value of $\xi_t(a)$ for all $a\in\mathcal{A}$. Then it follows that $Q_t(s_t,a)$ is defined for all $a\in\mathcal{A}$. As a result, we can determine a_t only by the knowledge of s_t. This means that $\mathcal{F}_{t-1}$ encodes the information of a_t,τ_t as well as s_t in it. Therefore, once conditioned on $\mathcal{F}_{t-1}$,

$$\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}\left[V^k_{(\tau_t-1)}(s_t,a_t)\right]$$

is fixed. Based on this, we may evaluate the conditional expectation of η_t given $\mathcal{F}_{t-1}$ as follows.

$$\begin{aligned}\mathbb{E}[\eta_t|\mathcal{F}_{t-1}] &= \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}\left[V^k_{(\tau_t-1)}(s_t,a_t)\right] - \mathbb{E}\left[V^k_{(\tau_t-1)}(s_{t+1})|\mathcal{F}_{t-1}\right] \\ &= \int_{s'\in\mathcal{S}} \mathbb{P}(s' \mid s_t,a_t)V^k_{(\tau_t-1)}(s')ds' - \int_{s'\in\mathcal{S}} \mathbb{P}(s' \mid s_t,a_t)V^k_{(\tau_t-1)}(s')ds' \\ &= 0.\end{aligned}$$

So far, we proved that the term I_3 is the sum of martingale difference sequence. Then we can bound the term I_3 based on the Azuma-Hoeffding inequality.

Lemma C.1 (Azuma-Hoeffding inequality) *Let $\{X_k\}_{k=0}^{\infty}$ be a discrete-parameter real-valued martingale sequence such that for every $k \in \mathbb{N}$, the condition $|X_k - X_{k-1}| \leq \mu$ holds for some non-negative constant μ. Then with probability at least $1-\delta$, we have*

$$X_n - X_0 \leq \mu\sqrt{2n\log(1/\delta)}.$$

Since $X_t = \sum_{n=1}^{t}\eta_t$ for $t \geq 1$ and X_0 give rise to a martingale sequence with $|\eta_t| \leq H$, it follows from Lemma C.1 that

$$I_3 \leq H\sqrt{2T\log(1/\delta)}$$

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

C.3 Proof of Lemma 5.11: Upper Bound on the Number of Episodes

Note that $\det(\Sigma_1) = \lambda^d$ because $\Sigma_1 = \lambda I_d$. To upper bound $\det(\Sigma_{T+1})$, we apply the following lemma.

Lemma C.2 (Lemma 10, Abbasi-yadkori et al., 2011) *For any $x_1, \ldots, x_T \in \mathbb{R}^d$ such that $\|\mathbf{x}_t\|_2 \leq L$, let $\mathbf{A}_1 = \lambda I_d$ and $\mathbf{A}_{t+1} = \lambda I_d + \sum_{i=1}^{t}\mathbf{x}_i\mathbf{x}_i^\top$ for $t \geq 1$. Then*

$$\det(\Sigma_{T+1}) \leq \left(\lambda + \frac{TL^2}{d}\right)^d.$$

Recall that for the linear MDP case,

$$\Sigma_{T+1} = \lambda I_d + \sum_{k=1}^{K(T)}\sum_{t=t_k}^{t_{k+1}-1}\varphi(s_t,a_t)\varphi(s_t,a_t)^\top$$

and for the linear mixture MDP case,

$$\Sigma_{T+1} = \lambda I_d + \sum_{k=1}^{K(T)}\sum_{t=t_k}^{t_{k+1}-1}(\varphi(s_t,a_t) + \gamma\cdot\phi_{W_t}(s_t,a_t))(\varphi(s_t,a_t) + \gamma\cdot\phi_{W_t}(s_t,a_t))^\top.$$

Here, we have $\|\varphi(s_t,a_t)\|_2 \leq 1$ for the linear MDP setting while $\|\varphi(s_t,a_t)+\gamma\cdot\phi_{W_t}(s_t,a_t)\|_2 \leq B_\varphi + HB_\phi$ as $W_t(s) \in [0,H]$ for any $s \in \mathcal{S}$. By applying Lemma C.2, we get

$$\det(\Sigma_{T+1}) \leq \left(1 + \frac{T}{d}\right)^d$$

for the linear MDP case with $\lambda = 1$, in which case

$$\frac{\det(\Sigma_{T+1})}{\det(\Sigma_1)} \leq \left(1 + \frac{T}{d}\right)^d$$

For the linear mixture MDP setting, we get

$$\det(\Sigma_{T+1}) \leq \left(\lambda + \frac{T(B_\varphi + HB_\phi)^2}{d}\right)^d.$$

When $\lambda \geq (B_\varphi + HB_\phi)^2$, it follows that

$$\frac{\det(\Sigma_{T+1})}{\det(\Sigma_1)} \leq \left(1 + \frac{T(B_\varphi + HB_\phi)^2}{d\lambda}\right)^d \leq \left(1 + \frac{T}{d}\right)^d.$$

Moreover, note that

$$\det(\Sigma_{T+1}) \geq \det(\Sigma_T) \geq \det(\Sigma_{t_{K(T)}}) \geq \cdots \geq 2^{K(T)-1} \det(\Sigma_{t_1}),$$

implying in turn that

$$K(T) \leq 1 + \log_2(\det(\Sigma_{T+1})/\det(\Sigma_{t_1})) \leq 1 + d\log(1 + T/d),$$

as required.

Appendix D. Proof of Theorem 1: Regret Bound for Linear MDPs

D.1 Proof of Lemma 5.9: Regret Term I_4 for the Linear MDP Setting

Let us prove the desired upper bound on the term

$$\sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \|\varphi(s_t, a_t)\|_{\Sigma_{t_k}^{-1}}.$$

First, we argue that $\|\varphi(s_t, a_t)\|_{\Sigma_{t_k}^{-1}} \leq \sqrt{2}\|\varphi(s_t, a_t)\|_{\Sigma_t^{-1}}$. Observe that for $t \in [t_k : t_{k+1} - 1]$, we have $\Sigma_t \succeq \Sigma_{t_k}$. This implies that $\Sigma_{t_k}^{-1} \succeq \Sigma_t^{-1}$.

Lemma D.1 (Lemma 12, Abbasi-yadkori et al., 2011) *Let $A, B \in \mathbb{R}^{d\times d}$ be positive semidefinite matrices such that $A \succeq B$. Then for any $x \in \mathbb{R}^d$, we have $\|x\|_A \leq \|x\|_B \sqrt{\det(A)/\det(B)}$.*

By Lemma D.1, we have that

$$\|\varphi(s_t, a_t)\|_{\Sigma_{t_k}^{-1}} \leq \|\varphi(s_t, a_t)\|_{\Sigma_t^{-1}} \sqrt{\frac{\det(\Sigma_{t_k}^{-1})}{\det(\Sigma_t^{-1})}} = \|\varphi(s_t, a_t)\|_{\Sigma_t^{-1}} \sqrt{\frac{\det(\Sigma_t)}{\det(\Sigma_{t_k})}} \leq \sqrt{2}\|\varphi(s_t, a_t)\|_{\Sigma_t^{-1}}$$

where the second inequality holds because $\det(\Sigma_t) \leq 2\det(\Sigma_{t_k})$ for any $t < t_{k+1}$. Therefore, it follows that

$$\sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \|\varphi(s_t, a_t)\|_{\Sigma_{t_k}^{-1}} \leq \sqrt{2} \sum_{t=1}^{T} \|\varphi(s_t, a_t)\|_{\Sigma_t^{-1}}.$$

For the right-hand side, we apply the following lemma.

Lemma D.2 (Lemma 11, Abbasi-yadkori et al., 2011) *For any $x_1, \ldots, x_T \in \mathbb{R}^d$ such that $\|\mathbf{x}_t\|_2 \leq L$, let $\mathbf{A}_1 = \lambda I_d$ and $\mathbf{A}_{t+1} = \lambda I_d + \sum_{i=1}^{t} \mathbf{x}_i \mathbf{x}_i^\top$ for $t \geq 1$. If $\lambda \geq \max\{1, L^2\}$, then we have*

$$\sum_{t=1}^{T} \|\mathbf{x}_t\|^2_{\mathbf{A}_t^{-1}} \leq 2 \log \frac{\det(A_{T+1})}{\lambda^d}.$$

Note that $\Sigma_t \succeq \Sigma_1 = \lambda I_d$, we have $(1/\lambda) I_d = \Sigma_1^{-1} \succeq \Sigma_t^{-1}$. Then

$$\|\varphi(s_t, a_t)\|^2_{\Sigma_t^{-1}} \leq \|\varphi(s_t, a_t)\|^2_{\Sigma_1^{-1}} = \frac{1}{\lambda} \|\varphi(s_t, a_t)\|_2^2 \leq 1$$

where the first inequality holds as $\Sigma_1^{-1} \succeq \Sigma_t^{-1}$ while the second inequality holds because $\lambda = 1$ and $\|\varphi(s_t, a_t)\|_2 \leq 1$. Note that

$$\sum_{t=1}^{T} \|\varphi(s_t, a_t)\|_{\Sigma_t^{-1}} \leq \sqrt{T \sum_{t=1}^{T} \|\varphi(s_t, a_t)\|^2_{\Sigma_t^{-1}}} \leq \sqrt{2T \log(\det(\Sigma_{T+1}/\lambda^d))}.$$

where the first inequality is by the Cauchy-Schwarz inequality and the second inequality is due to Lemma D.2. By Lemma C.2, it follows that

$$\frac{\det(\Sigma_{t+1})}{\lambda^d} \leq \left(1 + \frac{T}{d}\right)^d$$

as $\lambda = 1$. Finally, we deduce that

$$\sum_{t=1}^{T} \|\varphi(s_t, a_t)\|_{\Sigma_{t_k}^{-1}} \leq \sqrt{2} \sum_{t=1}^{T} \|\varphi(s_t, a_t)\|_{\Sigma_t^{-1}} \leq 2\sqrt{dT \log(1 + T/d)},$$

as required.

D.2 Completing the Regret Bound for the Linear MDP Case

As explained in Section 5, we have the following regret upper bound for the linear MDP setting.

$$\text{Regret}(T) \leq (1 - \gamma)\text{sp}(v^*)T + \frac{K(T)}{1 - \gamma} + H\sqrt{2T \log(1/\delta)} + 2\beta\sqrt{dT \log\left(1 + \frac{T}{d}\right)}.$$

Since $1 - \gamma = 1/\sqrt{T}$, $K(T) \leq 2d\sqrt{\log_2(1 + T/d)}$ by Lemma 5.11, and $\beta = 16(1 + H)d\sqrt{\log(1 + dT/\delta)}$, we obtain

$$\begin{aligned}
\text{Regret}(T) &\leq \text{sp}(v^*)\sqrt{T} + 2d\sqrt{T \log_2\left(1 + \frac{T}{d}\right)} + H\sqrt{2T \log(1/\delta)} \\
&\quad + 32(1 + H)d\sqrt{dT \log\left(1 + \frac{T}{d}\right) \log(1 + dT/\delta)} \\
&\leq 37(1 + H)d^{3/2}\sqrt{T} \log\left(1 + \frac{dT}{\delta}\right).
\end{aligned}$$

Appendix E. Proof of Theorem 2: Regret Bound for Linear Mixture MDPs

E.1 Proof of Lemma 5.10: Regret Term I_4 for the Linear Mixture MDP Setting

In this section, we prove the required upper bound on the term

$$\sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \|\varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t)\|_{\Sigma_{t_k}^{-1}}.$$

For simplicity, we use notation φ_t to denote

$$\varphi_t = \varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t).$$

As in Appendix D.1 for the linear MDP setting, we may argue by Lemma D.1 that

$$\sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \|\varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t)\|_{\Sigma_{t_k}^{-1}} = \sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \|\varphi_t\|_{\Sigma_{t_k}^{-1}} \leq \sqrt{2} \sum_{t=1}^{T} \|\varphi_t\|_{\Sigma_t^{-1}}.$$

Moreover, we have

$$\|\varphi_t\|_{\Sigma_t^{-1}}^2 \leq \|\varphi_t\|_{\Sigma_1^{-1}}^2 = \frac{1}{\lambda} \|\varphi_t\|_2^2 \leq \frac{1}{\lambda}(B_\varphi + HB_\phi)^2.$$

Since $\lambda \geq (B_\varphi + HB_\phi)^2$, we have that $\|\varphi_t\|_{\Sigma_t^{-1}} \leq 1$. Then applying Lemma D.2, we deduce that

$$\sum_{t=1}^{T} \|\varphi_t\|_{\Sigma_t^{-1}} \leq \sqrt{T \sum_{t=1}^{T} \|\varphi_t\|_{\Sigma_t^{-1}}^2} \leq \sqrt{2T \log(\det(\Sigma_{T+1}/\lambda^d))}.$$

By Lemma C.2, we have

$$\frac{\det(\Sigma_{t+1})}{\lambda^d} \leq \left(1 + \frac{T(B_\varphi + HB_\phi)^2}{d\lambda}\right)^d \leq \left(1 + \frac{T}{d}\right)^d$$

as $\lambda \geq (B_\varphi + HB_\phi)^2$. Finally, we deduce that

$$\sum_{t=1}^{T} \|\varphi_t\|_{\Sigma_{t_k}^{-1}} \leq \sqrt{2} \sum_{t=1}^{T} \|\varphi_t\|_{\Sigma_t^{-1}} \leq 2\sqrt{dT \log(1 + T/d)},$$

E.2 Completing the Regret Bound for the Linear Mixture MDP Case

As explained in Section 5, we have the following regret upper bound for the linear mixture MDP setting.

$$\text{Regret}(T) \leq (1-\gamma)\text{sp}(v^*)T + \frac{K(T)}{1-\gamma} + H\sqrt{2T \log(1/\delta)} + 2\beta_T \sqrt{dT \log\left(1 + \frac{T}{d}\right)}.$$

Since $1-\gamma = 1/\sqrt{T}$, $K(T) \leq 2d\sqrt{\log_2(1+T/d)}$ by Lemma 5.11, and $\beta_T = H\sqrt{d\log((1+T(B_\varphi + HB_\phi)^2/\lambda)/\delta)}$ $B_\theta\sqrt{\lambda}$, we obtain

$$\begin{aligned}
\mathrm{Regret}(T) &\leq \mathrm{sp}(v^*)\sqrt{T} + 2d\sqrt{T\log_2\left(1+\frac{T}{d}\right)} + H\sqrt{2T\log(1/\delta)} \\
&\quad + 2Hd\sqrt{T\log\left(1+\frac{T}{d}\right)\log\left(\frac{1+T(B_\varphi+HB_\phi)^2/\lambda}{\delta}\right)} \\
&\quad + 2B_\theta\sqrt{\lambda dT\log\left(1+\frac{T}{d}\right)} \\
&\leq 7(1+H)d\sqrt{T}\log\left(1+\frac{dT}{\delta}\right) + 2B_\theta(B_\varphi+HB_\phi)\sqrt{dT\log\left(1+\frac{T}{d}\right)}
\end{aligned}$$

where the second inequality holds because $\lambda = (B_\varphi + HB_\phi)^2$.

Appendix F. An Algorithm with No Max-Pooling Step

In this section, we present an algorithm that does not apply max-pooling. The basic outline of Algorithm 3 is similar to that of LSVI-DC, but it has some distinct components.
In line 8, the clipping operation considers three terms, and the new term is $V^k_{(n)}$, which is the value function from the previous round. Due to this design of the clipping step, we deduce a monotone behavior such that

$$V^k_{(n+1)}(s) \leq V^k_{(n)}(s) \leq \cdots \leq V^k_{(1)}(s)$$

for any $s \in \mathcal{S}$. Line 11 replaces the max-pooling step of Algorithm 1. Note that the index τ_t is not meant for computing the maximum of some action-value functions.
The regret function under Algorithm 3 is given as follows.

$$\begin{aligned}
&\mathrm{Regret}(T)\left(= T\cdot J^* - \sum_{t=1}^{T} r(s_t,a_t)\right) \\
&\leq \underbrace{\sum_{k=1}^{K(T)}\sum_{t=t_k}^{t_{k+1}-1}\left(J^* - (1-\gamma)V^k_{(\tau_t-1)}(s_t)\right)}_{I_1} + \underbrace{\sum_{k=1}^{K(T)}\sum_{t=t_k}^{t_{k+1}-1}\left(V^k_{(\tau_t-1)}(s_t) - Q_t(s_t,a_t)\right)}_{I_2} \\
&\quad + \underbrace{\gamma\sum_{k=1}^{K(T)}\sum_{t=t_k}^{t_{k+1}-1}\left(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}\left[V^k_{(\tau_t-1)}(s')\right] - V^k_{(\tau_t-1)}(s_{t+1})\right)}_{I_3} + I_4 \\
&\quad + \underbrace{\gamma\sum_{k=1}^{K(T)}\sum_{t=t_k}^{t_{k+1}-1}\left(V^k_{(\tau_t-1)}(s_{t+1}) - V^k_{(\tau_t-1)}(s_t)\right)}_{I_5}
\end{aligned}$$

Algorithm 3 LSVI-DC with No Max-Pooling

Input: discount factor $\gamma \in (0,1)$, regularization parameter $\lambda > 0$, upper bound H of $2 \cdot \mathrm{sp}(v^*)$, bonus factor β, radius values $\beta_1, \ldots, \beta_T$

Initialize:initial state s_1, $t \leftarrow 1$, $\Sigma_1 \leftarrow \lambda I$, $b_1 = 0$, $\theta_1 = 0$

for episode $k = 1, 2, \cdots$ **do**

  $t_k \leftarrow t$, $V^k_{(1)} \leftarrow (1-\gamma)^{-1}$, $\widetilde{Q}^k_{(1)} \leftarrow (1-\gamma)^{-1}$, $N \leftarrow T^{3/2}(1-\gamma)^{-1}$, $\tau_{t_k - 1} \leftarrow N$

  for round $n = 1, \cdots, N$ **do**

    (Linear MDP):

    $w^k_{(n)} \leftarrow \Sigma^{-1}_{t_k} \sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau)(r(s_\tau, a_\tau) + \gamma(V^k_{(n)}(s_{\tau+1}) - \min_{s' \in \mathcal{S}} V^k_{(n)}(s')))$

    $\widetilde{Q}^k_{(n+1)}(\cdot,\cdot) \leftarrow \min\left\{\left\langle \varphi(\cdot,\cdot), w^k_{(n)} \right\rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V^k_{(n)}(s') + \beta \|\varphi(\cdot,\cdot)\|_{\Sigma^{-1}_{t_k}}, \frac{1}{1-\gamma}\right\}$

    (Linear mixture MDP):

    $\varphi^k_{(n)}(\cdot,\cdot) \leftarrow \varphi(\cdot,\cdot) + \gamma \cdot \phi_{\bar{V}^k_{(n)}}(\cdot,\cdot)$ where $\bar{V}^k_{(n)}(\cdot) = V^k_{(n)}(\cdot) - \min_{s' \in \mathcal{S}} V^k_{(n)}(s')$

    $\widetilde{Q}^k_{(n+1)}(\cdot,\cdot) \leftarrow \min\left\{\left\langle \varphi^k_{(n)}(\cdot,\cdot), \theta_{t_k} \right\rangle + \gamma \cdot \min_{s' \in \mathcal{S}} V^k_{(n)}(s') + \beta_{t_k} \|\varphi^k_{(n)}(\cdot,\cdot)\|_{\Sigma^{-1}_{t_k}}, \frac{1}{1-\gamma}\right\}$

    $\widetilde{V}^k_{(n+1)}(\cdot) \leftarrow \max_{a \in \mathcal{A}} \widetilde{Q}^k_{(n+1)}(\cdot, a)$

    $V^k_{(n+1)}(\cdot) \leftarrow \min\left\{\widetilde{V}^k_{(n+1)}(\cdot),\ \min_{s' \in \mathcal{S}} \widetilde{V}^k_{(n+1)}(s') + H, V^k_{(n)}\right\}$

  end for

  while $\det(\Sigma_t) \le 2\det(\Sigma_{t_k})$ **do**

    Take $\tau_t \leftarrow \max\{\tau \in [2, \tau_{t-1}) : \max_{a \in \mathcal{A}} \widetilde{Q}^k_{(\tau-1)}(s_t, a) \le \max_{a \in \mathcal{A}} \widetilde{Q}^k_{(\tau)}(s_t, a) + 1/\sqrt{T}\}$

    Set $Q_t(s,a) \leftarrow \widetilde{Q}^k_{(\tau_t)}(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

    Take $a_t \in \operatorname{argmax}_{a \in \mathcal{A}} Q_t(s_t, a)$, receive $r(s_t, a_t)$, and obtain $s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$

    (Linear MDP):

    $\Sigma_{t+1} \leftarrow \Sigma_t + \varphi(s_t, a_t)\varphi(s_t, a_t)^\top$

    (Linear mixture MDP) :

    Take $W_t(\cdot) \leftarrow V^k_{(\tau_t - 1)}(s) - \min_{s' \in \mathcal{S}} V^k_{(\tau_t - 1)}(s')$

    Update $\Sigma_{t+1} \leftarrow \Sigma_t + (\varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t))(\varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t))^\top$

    Update $b_{t+1} \leftarrow b_t + (\varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t))(r(s_t, a_t) + \gamma \cdot W_t(s_{t+1}))$

    Set $\theta_{t+1} \leftarrow \Sigma^{-1}_{t+1} b_{t+1}$

    $t \leftarrow t + 1$

  end while

end for

where

$$I_4 = \begin{cases} 2\gamma \sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \beta \|\varphi(s_t, a_t)\|_{\Sigma^{-1}_{t_k}}, & \text{(linear MDP)}, \\ 2\gamma \sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \beta_{t_k} \|\varphi(s_t, a_t) + \gamma \cdot \phi_{W_t}(s_t, a_t)\|_{\Sigma^{-1}_{t_k}}, & \text{(linear mixture MDP)}. \end{cases}$$

Note that regret term I_1 has $V^k_{(\tau_t - 1)}(s_t)$, not $V^k_{(\tau_t - 1)}(s_{t+1})$. As a result, regret term I_2 also has $V^k_{(\tau_t - 1)}(s_t)$, and we get an additional regret term I_5. Nevertheless, the regret term I_5

can be upper bounded, because $\tau_{t+1} - 1 \leq \tau_t - 1$ and thus $V^k_{(\tau_{t+1}-1)}(s_{t+1}) \geq V^k_{(\tau_t-1)}(s_{t+1})$. This leads to

$$I_5 \leq \gamma \sum_{k=1}^{K(T)} \sum_{t=t_k}^{t_{k+1}-1} \left(V^k_{(\tau_{t+1}-1)}(s_{t+1}) - V^k_{(\tau_t-1)}(s_t) \right)$$

where the right-hand side has a telescoping structure. For term I_2, note that

$$V^k_{(\tau_t-1)}(s_t) \leq \widetilde{V}^k_{(\tau_t-1)}(s_t) = \max_{a \in \mathcal{A}} \widetilde{Q}^k_{(\tau_t-1)}(s_t, a) \leq \max_{a \in \mathcal{A}} \widetilde{Q}^k_{(\tau_t)}(s_t, a) + \frac{1}{\sqrt{T}}$$

where the second inequality is due to the choice of τ_t. Then it follows that

$$V^k_{(\tau_t-1)}(s_t) \leq \max_{a \in \mathcal{A}} \widetilde{Q}^k_{(\tau_t)}(s_t, a) + \frac{1}{\sqrt{T}} = \max_{a \in \mathcal{A}} Q_t(s_t, a) + \frac{1}{\sqrt{T}} \leq Q_t(s_t, a_t) + \frac{1}{\sqrt{T}}.$$

This implies that $I_2 \leq \sqrt{T}$. The terms I_1 and I_3 can be dealt with similarly as done for Algorithm 1.

For linear mixture MDPs, term I_4 can be bounded above using β that has the same asymptotic magnitude as for the case of Algorithm 1. However, for linear MDPs, due to our clipping operation in line 8, the covering number for Algorithm 3 would be exponential in t. That means that we would have to choose β_t of magnitude $\widetilde{\mathcal{O}}(Hd\sqrt{t})$, which prevents us from deducing a sublinear regret.

Appendix G. Regret Bound for the Model-Free Tabular Setting

The tabular setting is indeed a special case of the linear MDP framework, as explained by Jin et al. (2020). To make our paper self-contained, let us explain the connection here. Consider a tabular MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$ with $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$. Then we take SA-dimensional feature map $\varphi(s, a)$ defined as the unit vector whose component corresponding to $(s, a) \in \mathcal{S} \times \mathcal{A}$ is 1. Next, the parameter $\theta \in \mathbb{R}^{SA}$ is given by

$$\theta_{(s,a)} = r(s, a),$$

in which case we have

$$\theta^\top \varphi(s, a) = \theta_{(s,a)} = r(s, a).$$

Moreover, we take $\mu(s') \in \mathbb{R}^d$ as

$$\mu(s')_{(s,a)} = \mathbb{P}(s' \mid s, a).$$

Then we get

$$\mu(s')^\top \varphi(s, a) = \mu(s')_{(s,a)} = \mathbb{P}(s' \mid s, a).$$

Note that we have $\|\theta\|_2 \leq \sqrt{d}$ and $\|\mu(\mathcal{S})\|_2 \leq \sqrt{d}$ as $\|\mu(s')\|_2 \leq \sqrt{d}$ for any $s' \in \mathcal{S}$. Therefore, LSVI-DC guarantees that the regret is $\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)(SA)^{3/2}\sqrt{T})$ as the dimension equals SA.