Reinforcement Learning for Infinite-Horizon Average-Reward Linear MDPs via Approximation by Discounted-Reward MDPs

Kihyuk Hong
University of Michigan
kihyukh@umich.edu

Woojin Chae
KAIST
woojeeny02@kaist.ac.kr

Yufan Zhang
University of Michigan
yufanzh@umich.edu

Dabeen Lee
KAIST
dabeenl@kaist.ac.kr

Ambuj Tewari
University of Michigan
tewaria@umich.edu

Abstract

We study the infinite-horizon average-reward reinforcement learning with linear MDPs. Previous approaches either suffer from computational inefficiency or require strong assumptions on dynamics, such as ergodicity, for achieving a regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$. In this paper, we propose an algorithm that achieves the regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$ and is computationally efficient in the sense that the time complexity is polynomial in problem parameters. Our algorithm runs an optimistic value iteration on a discounted-reward MDP that approximates the average-reward setting. With an appropriately tuned discounting factor γ, the algorithm attains the desired $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. The challenge in our approximation approach is to get a regret bound with a sharp dependency on the effective horizon $1/(1-\gamma)$. We address this challenge by clipping the value function obtained at each value iteration step to limit the span of the value function.

1 Introduction

Reinforcement learning (RL) in the infinite-horizon average-reward setting aims to learn a policy that maximizes the average reward in the long run. This setting is relevant for applications where the interaction between the agent and environment lacks a natural termination, such as in network routing [21] or inventory management [12]. However, designing and analyzing algorithms in this setting is challenging because the associated Bellman operator is not a contraction. This prevents the use of optimistic value iteration-based algorithms, which add bonus terms to the value function, a widely used approach in the finite-horizon episodic and infinite-horizon discounted settings.

Faced with this challenge, the seminal work of [16] adopts a model-based approach in the tabular setting. Their algorithm constructs a confidence set on the transition model and runs value iterations by choosing an optimistic model from the confidence set at each iteration, without adding bonus terms. Most works in the tabular setting follow this principle of constructing confidence sets for the model [7, 10]. Other works either have suboptimal regret bound [29] or assume a strong ergodic assumption [29].

The model-based approach used in the tabular setting does not generalize well to function approximation settings, where the state space can be arbitrarily large, making sample-efficient model estimation challenging. As a result, works on infinite-horizon average-reward RL with function approximation are sparse. Even in the linear MDP setting, existing works either impose strong ergodic assumptions [28] or use computationally inefficient algorithms [15, 28] to achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret.

A recent line of work employs the technique of approximating infinite-horizon average-reward RL in tabular MDPs using the discounted setting. The key insight is that the discounted setting can closely approximates the average-reward setting when the discounting factor is close to 1. The approximation is beneficial because the discounted setting allows for greater flexibility in algorithm choice, thanks to the contraction property of the corresponding Bellman operator. However, there is a trade-off in choosing γ: as γ approaches 1, the approximation error decreases, but the regret in the discounted setting increases due to greater variability in the value function, which scales polynomially with $1/(1-\gamma)$.

Wei et al. [29] apply this approximation technique to design a model-free, Q-learning based algorithm, though they obtain a suboptimal regret bound of $\widetilde{\mathcal{O}}(T^{2/3})$ due to suboptimal order of $1/(1-\gamma)$. More recently, Zhang et al.

Table 1: Comparison of algorithms for infinite-horizon average-reward linear MDP

Algorithm	Regret $\widetilde{\mathcal{O}}(\cdot)$	Assumption	Computation poly$(\cdot)$
FOPO [28]	$\text{sp}(v^*)\sqrt{d^3T}$	Bellman optimality equation	T^d, A, d
OLSVI.FH [28]	$\sqrt{\text{sp}(v^*)}(dT)^{\frac{3}{4}}$	Bellman optimality equation	T, A, d
LOOP [15]	$\sqrt{\text{sp}(v^*)^3 d^3 T}$	Bellman optimality equation	T^d, A, d
MDP-EXP2 [28]	$d\sqrt{t_{\text{mix}}^3 T}$	Uniform Mixing	T, A, d
γ-LSCVI-UCB (Ours)	$\text{sp}(v^*)\sqrt{d^3T}$	Bellman optimality equation	T, S, A, d
Lower Bound [30]	$\Omega(d\sqrt{\text{sp}(v^*)T})$		

[32] achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret by improving the dependence on $1/(1-\gamma)$, demonstrating that the discounted setting approximation idea can give $\widetilde{\mathcal{O}}(\sqrt{T})$ regret for infinite-horizon average-reward RL.

In this paper, we apply the discounted setting approximation idea to the linear MDP setting and make the following contributions.

- **Theoretical result.** We design the first computationally efficient algorithm for infinite-horizon average-reward setting with linear MDPs that achieves $\widetilde{\mathcal{O}}(\text{sp}(v^*)\sqrt{d^3T})$ regret.
- **Algorithm design for tabular MDPs.** We demonstrate in the tabular setting that approximating the average-reward setting using the discounted setting, and running a simple value iteration-based algorithm with span-constrained value function estimates can achieve a $\widetilde{\mathcal{O}}(\sqrt{T})$ regret for the average-reward setting when the discount factor is appropriately tuned. The simplicity of the algorithm allows extending to the linear MDP setting.
- **Algorithm design for linear MDPs.** We adapt the algorithm design for tabular MDPs to work with linear MDPs. We show that a naive extension gives a vacuous regret bound. We propose a novel algorithm structure that decouples value iteration steps and decision making steps. Specifically, we generate a sequence of action value functions by running value iterations for T iterations in advance, then at decision making time step, choose action value function in reverse order.

1.1 Related Work

A comparison of our work to previous works on infinite-horizon average-reward linear MDPs is shown in Table 1. We say an algorithm is computationally efficient if it can be run with computational complexity poly(d, S, A, T) where d is the dimension of the feature vector, S is the size of the state space, A is the size of action space and T is the number of time steps. Our work is the first computationally efficient algorithm with computation complexity polynomial in the parameters d, S, A, T to achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret without making a strong ergodicity assumption. Our regret matches that of FOPO [28], an algorithm that requires solving a computationally intractable optimization problem for finding optimistic value function. A brute-force approach for solving their optimization problem requires computations polynomial in T^d, which is exponential in d. We defer the comparison for the tabular setting and discussion on other related work to Appendix E.

Reduction of Average-Reward to Discounted Setting The technique of reducing the average-reward setting to the discounted setting is used by Jin et al. [19], Wang et al. [25], Wang et al. [26], and Zurek et al. [33] to solve the sample complexity problem of producing a nearly optimal policy given access to a simulator in the tabular setting. Wei et al. [29] and Zhang et al. [32]

Infinite-Horizon Average-Reward Setting with Linear Mixture MDPs The linear mixture MDP setting is closely related to the linear MDP setting in that the Bellman operator admits a compact representation. However, linear mixture MDP parameterizes the probability transition model such that the transition probability is linear in the low-dimensional feature representation of state-action-state triplets. Such a structure allows a sample efficient estimation of the model, enabling the design of an optimism-based algorithm using a confidence set on the model, much like the model-based approach for the tabular setting. Wu et al. [30] and Ayoub et al. [5] design algorithms for this setting using a confidence set on the model.

2 Preliminaries

Notations Let $\|\boldsymbol{x}\|_A = \sqrt{x^T A x}$ for $\boldsymbol{x} \in \mathbb{R}^d$ and a psd matrix $A \in \mathbb{R}^{d\times d}$. Let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Let $\Delta(\mathcal{X})$ be the set of probability measures on $\mathcal{X}$. Let $[n] = \{1, \ldots, n\}$ and $[m : n] = \{m, m+1, \ldots, n\}$. Let $\mathrm{sp}(v) = \max_{s,s'} |v(s) - v(s')|$.

2.1 Infinite-Horizon Average-Reward MDPs

We consider a Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the probability transition kernel, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function. We assume $\mathcal{S}$ is a measurable space with possibly infinite number of elements and $\mathcal{A}$ is a finite set. We assume the reward is deterministic and the reward function r is known to the learner. The probability transition kernel P is unknown to the learner.

The interaction protocol between the learner and the MDP is as follows. The learner interacts with the MDP for T steps, starting from an arbitrary state $s_1 \in \mathcal{S}$ chosen by the environment. At each step $t = 1, \ldots, T$, the learner chooses an action $a_t \in \mathcal{A}$ and observes the reward $r(s_t, a_t)$ and the next state s_{t+1}. The next state s_{t+1} is drawn by the environment from $P(\cdot|s_t, a_t)$.

Consider a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ where $\pi(a|s)$ specifies the probability of choosing action a at state s. The performance measure of our interest for the policy π is the long-term average reward starting from an initial state s defined as

$$J^\pi(s) \coloneqq \liminf_{T\to\infty} \tfrac{1}{T}\mathbb{E}^\pi \left[\textstyle\sum_{t=1}^T r(s_t, a_t)|s_1 = s\right]$$

where $\mathbb{E}^\pi[\cdot]$ is the expectation with respect to the probability distribution on the trajectory $(s_1, a_1, s_2, a_2, \ldots)$ induced by the interaction between P and π. The performance of the learner is measured by the regret against the best stationary policy π^* that maximizes $J^\pi(s_1)$. Writing $J^*(s_1) \coloneqq J^{\pi^*}(s_1)$, the regret is

$$R_T \coloneqq \textstyle\sum_{t=1}^T (J^*(s_1) - r(s_t, a_t)).$$

As discussed by Bartlett et al. [7], without an additional assumption on the structure of the MDP, if the agent enters a bad state from which reaching the optimally rewarding states is impossible, the agent may suffer a linear regret. To avoid this pathological case, we follow Wei et al. [28] and make the following structural assumption on the MDP.

Assumption A (Bellman optimality equation)**.** *There exist $J^* \in \mathbb{R}$ and functions $v^* : \mathcal{S} \to \mathbb{R}$ and $q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have*

$$\begin{gathered} J^* + q^*(s, a) = r(s, a) + Pv^*(s, a) \\ v^*(s) = \max_{a\in\mathcal{A}} q^*(s, a). \end{gathered}$$

As shown by Wei et al. [28], under the assumption above, the policy π^* that deterministically selects an action from $\mathrm{argmax}_a\, q^*(s, a)$ at each state $s \in \mathcal{S}$ is an optimal policy. Moreover, π^* always gives an optimal average reward $J^{\pi^*}(s_1) = J^*$ for all initial states $s_1 \in \mathcal{S}$. Since the optimal average reward is independent of the initial state, we can simply write the regret as $R_T = \sum_{t=1}^T (J^* - r(s_t, a_t))$. Functions $v^*(s)$ and $q^*(s, a)$ are the relative advantage of starting with s and (s, a) respectively. We can expect a problem with large $\mathrm{sp}(v^*)$ to be more difficult since starting with a bad state can be more disadvantageous. As is common in the literature [7, 29], we assume $\mathrm{sp}(v^*)$ is known to the learner.

2.2 Approximation by Infinite-Horizon Discounted Setting

The key idea of this paper, inspired by Zhang et al. [32], is to approximate the infinite-horizon average-reward setting by the infinite-horizon discounted setting with a discount factor $\gamma \in [0, 1)$ tuned carefully. Introducing the discounting factor allows for a computationally efficient algorithm design that exploits the contraction property of the Bellman operator for the infinite-horizon discounted setting. When γ is close to 1, we expect the optimal policy for the discounted setting to be nearly optimal for the average-reward setting, given the classical result [24] that says the average reward of a stationary policy is equal to the limit of the discounted cumulative reward as γ goes to 1. Before stating a lemma that relates the infinite-horizon average-reward setting and the infinite-horizon discounted setting, we define the value function under the discounted setting. For a policy π, define

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r(s_t,a_t)|s_1=s\right]$$
$$Q^{\pi}(s,a) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r(s_t,a_t)|s_1=s,a_1=a\right].$$

We suppress the dependency of the value functions on the discounting factor γ for notational convenience. We write the optimal value functions under the discounted setting as

$$V^*(s) = \max_{\pi} V^{\pi}(s), \quad Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a).$$

The following lemma relates the infinite-horizon average-reward setting and the discounted setting.

Lemma 1 (Lemma 2 in Wei et al. [29]). *For any $\gamma \in [0,1)$, the optimal value function V^* for the infinite-horizon discounted setting with discounting factor γ satisfies*

(i) $sp(V^*) \leq 2sp(v^*)$ and

(ii) $|(1-\gamma)V^*(s) - J^*| \leq (1-\gamma)sp(v^*)$ for all $s \in \mathcal{S}$.

The lemma above suggests that the difference between the optimal average reward J^* and the optimal discounted cumulative reward normalized by the factor $(1-\gamma)$ is small as long as γ is close to 1. Hence, we can expect the policy optimal under the discounted setting will be nearly optimal for the average-reward setting, provided γ is sufficiently close to 1.

3 Warmup: Tabular Setting

In this section, we introduce an algorithm designed for the tabular setting, where the state space $\mathcal{S}$ and action space $\mathcal{A}$ are both finite, and no specific structure is assumed for the reward function or the transition probabilities. The structure of the algorithm, along with the accompanying analysis, will lay the groundwork for extending these results to the linear MDP setting.

3.1 Algorithm

Our algorithm, called *discounted upper confidence bound clipped value iteration* (γ-UCB-CVI), adapts UCBVI [6], which was originally designed for the finite-horizon episodic setting, to the infinite-horizon discounted setting. At each time step, the algorithm performs an approximate Bellman backup with an added bonus term $\beta\sqrt{1/N_t(s,a)}$ (Line 8) where $N_t(s,a)$ is the number of times the state-action pair (s,a) is visited. The bonus term is designed to guarantee optimism, ensuring that $Q_t \geq Q^*$ for all $t = 1,\ldots,T$. A key modification from UCBVI is the clipping step (Line 10), which bounds span of the value function estimate V_t by H. Without clipping, V_t can range from 0 to $\frac{1}{1-\gamma}$, while with clipping, the range is restricted to $[0,H]$. As we will see in the analysis, this clipping step is crucial to achieving a sharp dependence on $\frac{1}{1-\gamma}$ in the regret bound, which enables the $\widetilde{O}(\sqrt{T})$ regret through tuning γ. With the discounting factor set to $\gamma = 1 - 1/\sqrt{T}$ and the span to $H = 2 \cdot \mathrm{sp}(v^*)$, we get the following regret bound.

Theorem 2. *Under Assumption A, there exists a constant $c > 0$ such that, for any fixed $\delta \in (0,1)$, if Algorithm 1 is run with $\gamma = 1 - \sqrt{1/T}$, $H = 2 \cdot sp(v^*)$, and $\beta = cH\sqrt{S\log(SAT/\delta)}$, then with probability at least $1-\delta$, the total regret is bounded by*

$$R_T \leq \mathcal{O}\left(sp(v^*)\sqrt{S^2AT\log(SAT/\delta)}\right).$$

In the theorem above, the constant c used for defining β can be chosen to be the one determined in Lemma 3. Our regret bound matches the best previously known regret bound for computationally efficient algorithm in this setting. See Appendix E for a full comparison with previous works on infinite-horizon average-reward tabular MDPs. We believe we can improve our bound by a factor of $\sqrt{S}$ with a refined analysis using Bernstein inequality, following the idea of the refined analysis for UCBVI provided by Azar et al. [6]. We leave the improvement to future work.

Algorithm 1: γ-UCB-CVI for Tabular Setting

Input: Discounting factor $\gamma \in [0, 1)$, span H, bonus factor β.
Initialize: $Q_1(s,a), V_1(s) \leftarrow \frac{1}{1-\gamma}$; $N_0(s,a,s') \leftarrow 0, N_0(s,a) \leftarrow 1$, for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.

1 Receive initial state s_1.
2 **for** *time step* $t = 1, \dots, T$ **do**
3   Take action $a_t = \operatorname{argmax}_a Q_t(s_t, a)$. Receive reward $r(s_t, a_t)$. Receive next state s_{t+1}.
4     $N_t(s_t, a_t, s_{t+1}) \leftarrow N_{t-1}(s_t, a_t, s_{t+1}) + 1$
5     $N_t(s_t, a_t) \leftarrow N_{t-1}(s_t, a_t) + 1$.
6   (Other entries of N_t remain the same as N_{t-1}.)
7     $\widehat{P}_t(s'|s,a) \leftarrow N_t(s,a,s')/N_t(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
8     $Q_{t+1}(s,a) \leftarrow (r(s,a) + \gamma(\widehat{P}_t V_t)(s,a) + \beta\sqrt{1/N_t(s,a)}) \wedge Q_t(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$
9     $\widetilde{V}_{t+1}(s) \leftarrow (\max_a Q_{t+1}(s,a)) \wedge V_t(s), \quad \forall s \in \mathcal{S}$.
10     $V_{t+1}(s) \leftarrow \widetilde{V}_{t+1}(s) \wedge (\min_{s'} \widetilde{V}_{t+1}(s') + H), \quad \forall s \in \mathcal{S}$.

3.2 Analysis

In this section, we give a sketch of the proof of Theorem 2. The key to the proof is the following concentration inequality for the model estimate.

Lemma 3. *Under the setting of Theorem 2, there exists a constant c such that for any fixed $\delta \in (0,1)$, we have with probability at least $1-\delta$ that*

$$|(\widehat{P}_t - P)V_t(s,a)| \le c \cdot sp(v^*)\sqrt{S\log(SAT/\delta)/N_t(s,a)}$$

for all $(s,a,t) \in \mathcal{S} \times \mathcal{A} \times [T]$.

See Appendix A.1 for a proof. With the concentration inequality, we can show the following optimism result. See Appendix A.2 for a proof.

Lemma 4 (Optimism)**.** *Under the setting of Theorem 2, we have with probability at least $1-\delta$ that*

$$V_t(s) \ge V^*(s), \quad Q_t(s,a) \ge Q^*(s,a)$$

for all $(s,a,t) \in \mathcal{S} \times \mathcal{A} \times [T]$.

Now, we show the regret bound under the event that the high probability events in the previous two lemmas (Lemma 3, Lemma 4) hold. By the value iteration step (Line 8) of Algorithm 1 and the concentration inequality in Lemma 3, we have for all $t = 2, \dots, T$ that

$$\begin{aligned} r(s_t, a_t) &\ge Q_t(s_t,a_t) - \gamma(\widehat{P}_{t-1}V_{t-1})(s_t,a_t) - \beta\sqrt{1/N_{t-1}(s_t,a_t)} \\ &\ge V_t(s_t) - \gamma(PV_{t-1})(s_t,a_t) - 2\beta\sqrt{1/N_{t-1}(s_t,a_t)} \end{aligned}$$

where the second inequality follows by $V_t(s_t) \le \widetilde{V}_t(s_t) \le \max_a Q_t(s_t,a) = Q_t(s_t,a_t)$. Hence, the regret can be bounded by the following decomposition.

$$\begin{aligned} R_T &= TJ^* - \sum_{t=1}^T r(s_t,a_t) \\ &\le J^* - \sum_{t=2}^T (J^* - V_t(s_t) + \gamma(PV_{t-1})(s_t,a_t) + 2\beta\sqrt{1/N_{t-1}(s_t,a_t)}) \\ &= J^* + \underbrace{\sum_{t=2}^T (J^* - (1-\gamma)V_t(s_t))}_{(a)} + \underbrace{\gamma\sum_{t=2}^T (V_{t-1}(s_{t+1}) - V_t(s_t))}_{(b)} \\ &\quad + \underbrace{\gamma\sum_{t=2}^T (PV_{t-1}(s_t,a_t) - V_{t-1}(s_{t+1}))}_{(c)} + \underbrace{2\beta\sum_{t=2}^T 1/\sqrt{1/N_{t-1}(s_t,a_t)}}_{(d)}. \end{aligned}$$

We bound the terms $(a), (b), (c), (d)$ separately as follows.

Bounding Term (a) We can bound

$$\sum_{t=2}^{T}(J^* - (1-\gamma)V_t(s_t)) \le \sum_{t=2}^{T}(J^* - (1-\gamma)V^*(s_t))$$
$$\le T(1-\gamma)\mathrm{sp}(v^*),$$

where the first inequality follows from optimism (Lemma 4), and the second inequality uses Lemma 1 that relates J^* to $(1-\gamma)V^*(s)$.

Bounding Term (b) Note that $V_t(s)$ is decreasing in t for any fixed $s \in \mathcal{S}$ by Line 9-10 in Algorithm 1, and the most it can decrease as t increases from 1 to T is $\frac{1}{1-\gamma}$. Hence,

$$\sum_{t=2}^{T}(V_{t-1}(s_{t+1}) - V_t(s_t)) \le \frac{1}{1-\gamma} + \sum_{t=2}^{T}(V_{t-1}(s_{t+1}) - V_{t+1}(s_{t+1}))$$
$$\le \frac{1}{1-\gamma} + \sum_{s\in\mathcal{S}}\sum_{t=2}^{T}(V_{t-1}(s) - V_{t+1}(s)) \le \frac{2S}{1-\gamma}.$$

The bound above is polynomial in S, the size of the state space, which is undesirable in the linear MDP setting where S can be arbitrarily large. The main challenge of this paper, as we will see in the later section, is sidestepping this issue for linear MDPs.

Bounding Term (c) Term (c) is the sum of a martingale difference sequence where each term is bounded by $\mathrm{sp}(v^*)$. Hence, by the Azuma-Hoeffding inequality, with probability at least $1-\delta$, term (c) is bounded by $\mathrm{sp}(v^*)\sqrt{2T\log\frac{1}{\delta}}$. Note that without clipping, each term of the martingale difference sequence can only be bounded by $\frac{1}{1-\gamma}$ leading to a bound of $\widetilde{\mathcal{O}}(\frac{1}{1-\gamma}\sqrt{T})$, which is too loose for achieving a regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$.

Bounding Term (d) We can bound the sum of the bonus terms (d) by

$$\beta\sum_{t=2}^{T}\sqrt{1/N_{t-1}(s_t,a_t)} \le \beta\sum_{s\in\mathcal{S},a\in\mathcal{A}}\sum_{n=1}^{N_T(s,a)}\sqrt{\frac{1}{n}}$$
$$\le 2c\cdot\mathrm{sp}(v^*)\sqrt{S^2AT\log(SAT/\delta)},$$

where the last inequality holds because $\sum_{s,a}N_T(s,a) = T$ and the left-hand side of the last inequality is maximized when $N_T(s,a) = T/SA$ for all s, a.

Combining the above, and rescaling δ, it follows that with probability at least $1-\delta$, we have

$$R_T \le \mathcal{O}\Big(T(1-\gamma)\mathrm{sp}(v^*) + \frac{S}{1-\gamma} + \mathrm{sp}(v^*)\sqrt{T\log(1/\delta)} + \mathrm{sp}(v^*)\sqrt{S^2AT\log(SAT/\delta)}\Big).$$

Choosing γ such that $\frac{1}{1-\gamma} = \sqrt{T}$, we get

$$R_T \le \mathcal{O}\left(\mathrm{sp}(v^*)\sqrt{S^2AT\log(SAT/\delta)}\right),$$

which completes the proof of Theorem 2.

4 Linear MDP Setting

In this section, we apply the key ideas developed from the previous section to the linear MDP setting. The linear MDP setting is defined as follows.

Assumption B (Linear MDP [18])**.** *We assume that the transition and the reward functions can be expressed as a linear function of a known d-dimensional feature map $\boldsymbol{\varphi} : \mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$ such that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have*

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

where $\boldsymbol{\mu}(s') = (\mu_1(s'),\ldots,\mu_d(s'))$ for $s'\in\mathcal{S}$ is a vector of d unknown (signed) measures on $\mathcal{S}$ and $\boldsymbol{\theta}\in\mathbb{R}^d$ is a known parameter for the reward function.

Algorithm 2: γ-LSCVI-UCB for linear MDP setting with minimum oracle

Input: Discounting factor $\gamma \in (0,1)$, regularization $\lambda > 0$, span H, bonus factor β.
Initialize: $t \leftarrow 1$, $k \leftarrow 1$, $t_k \leftarrow 1$, $\Lambda_1 \leftarrow \lambda I$, $\bar{\Lambda}_0 \leftarrow \lambda I$, $Q_t^1(\cdot,\cdot) \leftarrow \frac{1}{1-\gamma}$ for $t \in [T]$.

1 Receive state s_1.
2 **for** *time step* $t = 1, \dots, T$ **do**
3   Take action $a_t = \operatorname{argmax}_a Q_t^k(s_t, a)$. Receive reward $r(s_t, a_t)$. Receive next state s_{t+1}.
4     $\bar{\Lambda}_t \leftarrow \bar{\Lambda}_{t-1} + \boldsymbol{\varphi}(s_t, a_t)\boldsymbol{\varphi}(s_t, a_t)^T$.
5   **if** $2\det(\Lambda_k) < \det(\bar{\Lambda}_t)$ **then**
6         $k \leftarrow k+1$, $t_k \leftarrow t+1$, $\Lambda_k \leftarrow \bar{\Lambda}_t$.
        `// Run value iteration to plan for remaining` $T - t_k + 1$ `time steps in the new episode.`
7         $\widetilde{V}_{T+1}^k(\cdot) \leftarrow \frac{1}{1-\gamma}$, $V_{T+1}^k(\cdot) \leftarrow \frac{1}{1-\gamma}$.
8   **for** $u = T, T-1, \dots, t_k$ **do**
9             $\boldsymbol{w}_{u+1}^k \leftarrow \Lambda_k^{-1} \sum_{\tau=1}^{t_k - 1} \boldsymbol{\varphi}(s_\tau, a_\tau)(V_{u+1}^k(s_{\tau+1}) - \min_{s'} \widetilde{V}_{u+1}^k(s'))$.
10             $Q_u^k(\cdot,\cdot) \leftarrow \Big(r(\cdot,\cdot) + \gamma(\langle \boldsymbol{\varphi}(\cdot,\cdot), \boldsymbol{w}_{u+1}^k \rangle + \min_{s'} \widetilde{V}_{u+1}^k(s') + \beta \|\boldsymbol{\varphi}(\cdot,\cdot)\|_{\Lambda_k^{-1}})\Big) \wedge \frac{1}{1-\gamma}$.
11             $\widetilde{V}_u^k(\cdot) \leftarrow \max_a Q_u^k(\cdot, a)$.
12             $V_u^k(\cdot) \leftarrow \widetilde{V}_u^k(\cdot) \wedge (\min_{s'} \widetilde{V}_u^k(s') + H)$.

As is commonly done in works on linear MDPs[18], we further assume, without loss of generality (see [28] for justification), the following boundedness conditions:

$$\begin{aligned} \|\boldsymbol{\varphi}(s,a)\|_2 \le 1 \text{ for all } (s,a) \in \mathcal{S} \times \mathcal{A}, \\ \|\boldsymbol{\theta}\|_2 \le \sqrt{d}, \quad \|\boldsymbol{\mu}(\mathcal{S})\|_2 \le \sqrt{d}. \end{aligned} \tag{1}$$

As discussed by Jin et al. [18], although the transition model P is linear in the d-dimensional feature mapping $\boldsymbol{\varphi}$, P still has $|\mathcal{S}|$ degrees of freedom as the measure $\boldsymbol{\mu}$ is unknown, making the estimation of the model P difficult. For sample efficient learning, we rely on the fact that $Pv(s,a)$ is linear in $\boldsymbol{\varphi}(s,a)$ for any function $v : \mathcal{S} \to \mathbb{R}$ so that $Pv(s,a) = \langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_v^* \rangle$ for some $\boldsymbol{w}_v^*$, since

$$\begin{aligned} Pv(s,a) &\coloneqq \textstyle\int_{s' \in \mathcal{S}} v(s') P(ds'|s,a) \\ &= \int_{s' \in \mathcal{S}} v(s') \langle \boldsymbol{\varphi}(s,a), \boldsymbol{\mu}(ds') \rangle \\ &= \langle \boldsymbol{\varphi}(s,a), \int_{s' \in \mathcal{S}} v(s') \boldsymbol{\mu}(ds') \rangle. \end{aligned}$$

Challenges Naively adapting the algorithm design and analysis for the tabular setting to the linear MDP setting will give a regret bound polynomial in S, the size of the state space, when bounding $\sum_{t=1}^T (V_{t-1}(s_{t+1}) - V_t(s_t))$. Also, algorithmically making the state value function monotonically decrease in t by taking minimum with the previous estimate every iteration, as is done in the tabular setting for the telescoping sum argument, will make the covering number of the function class for the value function exponential in either T or S [13]. A major challenge in algorithm design and analysis is to sidestep these issues.

Now, we present our algorithm for the linear MDP setting.

4.1 Algorithm

Our algorithm, called *discounted least-squares clipped value iteration with upper confidence bound* (γ-LSCVI-UCB), adapts LSVI-UCB [18] developed for the episodic setting to the discounted setting. We highlight key differences from LSVI-UCB below.

Clipping the Value Function We clip the state value function estimates to restrict the span (Line 12), which is crucial for statistical efficiency that saves a factor of $1/(1-\gamma)$ in the regret bound.

Restricting the Range of Value Target When regressing $V_u^k(s')$ on $\boldsymbol{\varphi}(s,a)$ we subtract the value target by $\min_{s'} \widetilde{V}_u^k(s')$ and use $V_u^k(\cdot) - \min_{s'} \widetilde{V}_u^k(s')$ as the value target instead of $V_u^k(\cdot)$ (Line 9). This adjustment of the value target guarantees a bound on $\|\boldsymbol{w}_u^k\|_2$ that scales with $\mathrm{sp}(v^*)$ instead of $1/(1-\gamma)$, which is necessary for achieving $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. To compensate for the adjustment, we add back $\min_{s'} \widetilde{V}_u^k(s')$ when estimating the value target using the regression coefficient $\boldsymbol{w}_u^k$: $\langle \boldsymbol{\varphi}(\cdot,\cdot), \boldsymbol{w}_u^k \rangle + \min_{s'} \widetilde{V}_u^k(s')$ (Line 10).

Decoupling Value Iteration Step and Decision Making Step In our previous algorithm γ-UCB-CVI, designed for the tabular setting, the value iteration step alternates with the decision making step. At each time step t, a greedy action is selected based on the most recently constructed action value function Q_t. This structure is common in value iteration based algorithms and Q-learning algorithms for both infinite-horizon average-reward tabular MDPs [32] and infinite-horizon discounted tabular MDPs [14, 20]. With the coupling of value iteration and decision making steps, bounding the term $\sum_t V_{t-1}(s_{t+1}) - V_t(s_t)$ required enforcing V_t to be decreasing in t algorithmically since V_t is one Bellman operation ahead of V_{t-1}. However, as discussed previously, in the linear MDP setting, making V_t monotonically decreasing by taking the minimum with previous value functions would cause the log covering number of the function class for the value function to scale with T, making regret bound vacuous. To sidestep this issue, we use a novel algorithm structure that decouples the value iteration step and the decision making step. Specifically, before taking any action at time t, we generate a sequence of action value functions $Q_T, Q_{T-1}, \ldots, Q_t$ by running $T-t$ value iterations (Line 7-12). Then, at each decision time step t, take a greedy action with respect to Q_t. With the new algorithm structure, Q_{t-1} is now one Bellman operation ahead of Q_t, and the term of interest becomes $\sum_{t=1}^T V_{t+1}(s_{t+1}) - V_t(s_t)$, which can be bounded by telescoping sum.

Planning Periodically If we generate all T action value functions to be used at time step 1 and use them for decision-making for T steps, we cannot make use of the trajectory data collected. To address this, and still use the scheme of pregenerating action value functions, we periodically restart the process of running value iterations for T steps. This allows us to incorporate the newly collected trajectory data into subsequent decision-making. We adopt the rarely-switching covariance matrix trick [27], which triggers a restart when the determinant of the empirical covariance matrix doubles (Line 5).

The algorithm has the following guarantee.

Theorem 5. *Under Assumptions A and B, running Algorithm 2 with inputs $\gamma = 1 - \sqrt{\log(T)/T}$, $\lambda = 1$, $H = 2 \cdot sp(v^*)$ and $\beta = 2c_\beta \cdot sp(v^*) d\sqrt{\log(dT/\delta)}$ guarantees with probability at least $1-\delta$,*

$$R_T \leq \mathcal{O}(sp(v^*)\sqrt{d^3 T \log(dT/\delta)}).$$

The constant c_β is an absolute constant defined in Lemma 6. We expect careful analysis of the variance of the value estimate [13] may improve our regret by a factor of $\sqrt{d}$. We leave the improvement to future work.

4.2 Computational Complexity

Our algorithm runs in episodes and since the episode doubles when the covariance matrix $\bar{\Lambda}_t$ doubles, there can be at most $\mathcal{O}(d \log_2 T)$ episodes (see Lemma 20). In each episode, we run at most T value iterations. In each iteration step u, the algorithm computes $\min_{s'} \widetilde{V}_u^k(s')$ which requires evaluating $\widetilde{V}_u^k(s')$ at all $s' \in \mathcal{S}$, which requires $\mathcal{O}(d^2 SA)$ computations. All other operations runs in $\mathcal{O}(d^2 + A)$ per value iteration. In total, the algorithm runs in $\mathcal{O}((\log_2 T) d^2 SAT)$. See Appendix C for detailed analysis.

The FOPO algorithm by the most relevant work Wei et al. [28] has time complexity of $\mathcal{O}((\log_2 T) T^d)$. Although our time complexity is an improvement over previous work in the sense that the time complexity is polynomial in problem parameters, we introduce linear dependency on S. The dependency on S is from taking the minimum of value functions for clipping. We believe that we can get rid of the dependency on S by guessing the minimum instead of taking global minimum. For example, it can be shown that using $\min_{s'} V^*(s')$ instead of $\min_{s'} V_u^k(s')$ for clipping allows the same regret bound (see Appendix B.3). A promising approach is to use $\min_{s' \in \{s_1, \ldots, s_t\}} V_u^k(s')$, minimum over states visited so far, instead of the global minimum. We leave getting rid of S dependency on time complexity to future work.

4.3 Analysis

In this section, we provide a sketch of the proof of the regret bound given in Theorem 5. We first show that the value iteration step in Line 10 with the bonus term $\beta\|\boldsymbol{\phi}(\cdot,\cdot)\|_{\Lambda_k^{-1}}$ with appropriately chosen β ensures the value function estimates V_t and Q_t to be optimistic estimates of V^* and Q^* respectively. The argument is based on the following concentration inequality for the regression coefficients. See Appendix B.1 for a proof.

Lemma 6 (Concentration of regression coefficients)**.** *With probability at least $1-\delta$, there exists an absolute constant c_β such that for $\beta = c_\beta \cdot Hd\sqrt{\log(dT/\delta)}$, we have*

$$|\langle \boldsymbol{\phi}, \boldsymbol{w}_u^k - \boldsymbol{w}_u^{k*}\rangle| \leq \beta \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}}$$

for all episode index k and for all vector $\boldsymbol{\phi} \in \mathbb{R}^d$ where $\boldsymbol{w}_u^{k*} \coloneqq \int (V_u^k(s) - \min_{s'} V_u^k(s'))d\boldsymbol{\mu}(s)$ is a parameter that satisfies $\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_u^{k*}\rangle = PV_u^k(s,a) - \min_{s'} V_u^k(s')$.

With the concentration inequality, we can show the following optimism result. See Appendix B.2 for an induction-based proof.

Lemma 7 (Optimism)**.** *Under the linear MDP setting, running Algorithm 2 with input $H = 2\cdot sp(v^*)$ guarantees with probability at least $1-\delta$ that for all episodes $k = 1, 2, \ldots, u = t_k, \ldots, T+1$ and for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have*

$$V_u^k(s) \geq V^*(s), \quad Q_u^k(s,a) \geq Q^*(s,a).$$

Now, we show the regret bound under the event that the high probability events in the previous two lemmas (Lemma 6, Lemma 7) hold. Let t be a time step in episode k such that both t and $t+1$ are in episode k. By the definition of $Q_u^k(\cdot,\cdot)$ (Line 10), we have for all $t = t_k, \ldots, T+1$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$ that

$$\begin{aligned} r(s,a) &\geq Q_t^k(s,a) - \gamma(\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_{t+1}^k\rangle + \min_{s'} V_{t+1}^k(s') \\ &\qquad - \beta\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_k^{-1}}) \\ &\geq Q_t^k(s,a) - \gamma P V_{t+1}^k(s,a) - 4\beta\|\boldsymbol{\varphi}(s,a)\|_{\bar{\Lambda}_t^{-1}} \end{aligned}$$

where the second inequality uses the concentration bound for the regression coefficients in Lemma 6. It also uses $\|\boldsymbol{x}\|_{\Lambda_k^{-1}} \leq 2\|\boldsymbol{x}\|_{\Lambda_t^{-1}}$ (Lemma 19). Hence, we can bound the regret in episode k by

$$\begin{aligned} R^k &= \sum_{t=t_k}^{t_{k+1}-1} (J^* - r(s_t,a_t)) \\ &\leq \sum_{t=t_k}^{t_{k+1}-1} (J^* - Q_t^k(s_t,a_t) + \gamma P V_{t+1}^k(s_t,a_t) + 4\beta\|\boldsymbol{\varphi}(s_t,a_t)\|_{\Lambda_t^{-1}}) \\ &= \underbrace{\sum_{t=t_k}^{t_{k+1}-1} (J^* - (1-\gamma)V_{t+1}^k(s_{t+1}))}_{(a)} + \gamma \underbrace{\sum_{t=t_k}^{t_{k+1}-1} (V_{t+1}^k(s_{t+1}) - Q_t^k(s_t,a_t))}_{(b)} \\ &\qquad + \gamma \underbrace{\sum_{t=t_k}^{t_{k+1}-1} (PV_{t+1}^k(s_t,a_t) - V_{t+1}^k(s_{t+1}))}_{(c)} + 4\beta \underbrace{\sum_{t=t_k}^{t_{k+1}-1} \|\boldsymbol{\varphi}(s_t,a_t)\|_{\Lambda_t^{-1}}}_{(d)} \end{aligned}$$

where the first inequality uses the bound for $r(s_t,a_t)$. With the same argument as in the tabular case, the term (a) summed over all episodes can be bounded by $T(1-\gamma)\mathrm{sp}(v^*)$ using the optimism $V_u^k(s_{t+1}) \geq V^*(s_{t+1})$. Term (d) summed over all episodes can be bounded by $\mathcal{O}(\beta\sqrt{dT\log T})$ using Cauchy-Schwartz and Lemma 18. The term (c) summed over all episodes is a sum of a martingale difference sequence and can be bounded by $\mathcal{O}(\mathrm{sp}(v^*)\sqrt{T\log(1/\delta)})$ using $\mathrm{sp}(V_u^k) \leq 2\cdot \mathrm{sp}(v^*)$ by the clipping step in Line 12. To bound term (b) note that

$$\begin{aligned} V_{t+1}^k(s_{t+1}) &\leq \widetilde{V}_{t+1}^k(s_{t+1}) \\ &= \max_a Q_{t+1}^k(s_{t+1},a) \\ &= Q_{t+1}^k(s_{t+1},a_{t+1}) \end{aligned}$$

as long as both t and $t+1$ are in episode k since the algorithm chooses a_{t+1} that maximizes $Q^k_{t+1}(s_{t+1},\cdot)$. Hence,

$$\sum_{t=t_k}^{t_{k+1}-1}(V^k_{t+1}(s_{t+1})-Q^k_t(s_t,a_t))\le\frac{1}{1-\gamma}+\sum_{t=t_k}^{t_{k+1}-2}(Q^k_{t+1}(s_{t+1},a_{t+1})-Q^k_t(s_t,a_t))$$
$$\le\mathcal{O}\left(\frac{1}{1-\gamma}\right)$$

where the second inequality uses telescoping sum and the fact that $Q^k_t\le\frac{1}{1-\gamma}$. Since the episode is advanced when the determinant of the covariance matrix doubles, it can be shown that the number of episodes is bounded by $\mathcal{O}(d\log(T))$ (Lemma 20). Combining all the bounds, and using $\beta=\mathcal{O}(\mathrm{sp}(v^*)d\sqrt{\log(dT/\delta)}$, we get

$$R_T\le\mathcal{O}\Big(T(1-\gamma)\mathrm{sp}(v^*)+\frac{d}{1-\gamma}\log(T)+\mathrm{sp}(v^*)\sqrt{T\log(1/\delta)}+\mathrm{sp}(v^*)\sqrt{d^3T\log(dT/\delta)}\Big).$$

Setting $\gamma=1-\frac{\sqrt{\log T}}{\sqrt{T}}$, we get

$$R_T\le\mathcal{O}\left(\mathrm{sp}(v^*)\sqrt{d^3T\log(dT/\delta)}\right).$$

5 Conclusion

In this paper, we propose an algorithm with time complexity polynomial in the problem parameters that achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret for infinite-horizon average-reward linear MDPs without making a strong ergodicity assumption on the dynamics. Our algorithm approximates the average-reward setting by the discounted setting with a carefully tuned discounting factor. A key technique that allows for order optimal regret bound is bounding the span of the value function in each value iteration step via clipping. Additionally, we precompute a sequence of action value functions by running value iterations, then use them in reverse order for taking actions. A promising direction for future work would be to extend these methods to the general function approximation setting.

6 Acknowledgements

KH and AT acknowledge the support of the National Science Foundation (NSF) under grant IIS-2007055. DL acknowledges the support of the National Research Foundation of Korea (NRF) under grant RS-2024-00350703.

References

[1] Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. "Online-to-confidence-set conversions and application to sparse stochastic bandits". In: *Artificial Intelligence and Statistics*. PMLR. 2012, pp. 1–9.

[2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. "Improved algorithms for linear stochastic bandits". In: *Advances in neural information processing systems* 24 (2011).

[3] Priyank Agrawal and Shipra Agrawal. "Optimistic Q-learning for average reward and episodic reinforcement learning". In: *arXiv preprint arXiv:2407.13743* (2024).

[4] Peter Auer, Thomas Jaksch, and Ronald Ortner. "Near-optimal regret bounds for reinforcement learning". In: *Advances in neural information processing systems* 21 (2008).

[5] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. "Model-based reinforcement learning with value-targeted regression". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 463–474.

[6] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. "Minimax regret bounds for reinforcement learning". In: *International conference on machine learning*. PMLR. 2017, pp. 263–272.

[7] Peter Bartlett and Ambuj Tewari. "REGAL: a regularization based algorithm for reinforcement learning in weakly communicating MDPs". In: *Uncertainty in Artificial Intelligence: Proceedings of the 25th Conference*. AUAI Press. 2009, pp. 35–42.

[8] Liyu Chen, Rahul Jain, and Haipeng Luo. "Learning infinite-horizon average-reward Markov decision process with constraints". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 3246–3270.

[9] Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. "Improved analysis of ucrl2 with empirical bernstein inequality". In: *arXiv preprint arXiv:2007.05456* (2020).

[10] Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. "Efficient bias-span-constrained exploration-exploitation in reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1578–1586.

[11] Arnob Ghosh, Xingyu Zhou, and Ness Shroff. "Achieving sub-linear regret in infinite horizon average reward constrained mdp with linear function approximation". In: *The Eleventh International Conference on Learning Representations*. 2023.

[12] Ilaria Giannoccaro and Pierpaolo Pontrandolfo. "Inventory management in supply chains: a reinforcement learning approach". In: *International Journal of Production Economics* 78.2 (2002), pp. 153–161.

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

[14] Jiafan He, Dongruo Zhou, and Quanquan Gu. "Nearly minimax optimal reinforcement learning for discounted MDPs". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22288–22300.

[15] Jianliang He, Han Zhong, and Zhuoran Yang. "Sample-efficient Learning of Infinite-horizon Average-reward MDPs with General Function Approximation". In: *The Twelfth International Conference on Learning Representations*. 2024.

[16] Thomas Jaksch, Ronald Ortner, and Peter Auer. "Near-optimal Regret Bounds for Reinforcement Learning". In: *Journal of Machine Learning Research* 11 (2010), pp. 1563–1600.

[17] Xiang Ji and Gen Li. "Regret-optimal model-free reinforcement learning for discounted mdps with short burn-in time". In: *Advances in Neural Information Processing Systems* 36 (2024).

[18] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. "Provably efficient reinforcement learning with linear function approximation". In: *Conference on learning theory*. PMLR. 2020, pp. 2137–2143.

[19] Yujia Jin and Aaron Sidford. "Towards tight bounds on the sample complexity of average-reward MDPs". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5055–5064.

[20] Shuang Liu and Hao Su. "Regret bounds for discounted mdps". In: *arXiv preprint arXiv:2002.05138* (2020).

[21] Zoubir Mammeri. "Reinforcement learning based routing in networks: Review and classification of approaches". In: *Ieee Access* 7 (2019), pp. 55916–55950.

[22] Ronald Ortner. "Regret bounds for reinforcement learning via markov chain concentration". In: *Journal of Artificial Intelligence Research* 67 (2020), pp. 115–128.

[23] Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. "Learning unknown markov decision processes: A thompson sampling approach". In: *Advances in neural information processing systems* 30 (2017).

[24] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[25] Jinghan Wang, Mengdi Wang, and Lin F Yang. "Near sample-optimal reduction-based policy learning for average reward mdp". In: *arXiv preprint arXiv:2212.00603* (2022).

[26] Shengbo Wang, Jose Blanchet, and Peter Glynn. "Optimal Sample Complexity for Average Reward Markov Decision Processes". In: *arXiv preprint arXiv:2310.08833* (2023).

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

[28] Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. "Learning infinite-horizon average-reward mdps with linear function approximation". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 3007–3015.

[29] Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. "Model-free reinforcement learning in infinite-horizon average-reward markov decision processes". In: *International conference on machine learning*. PMLR. 2020, pp. 10170–10180.

[30] Yue Wu, Dongruo Zhou, and Quanquan Gu. "Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 3883–3913.

[31] Zihan Zhang and Xiangyang Ji. "Regret minimization for reinforcement learning by evaluating the optimal bias function". In: *Advances in Neural Information Processing Systems* 32 (2019).

[32] Zihan Zhang and Qiaomin Xie. "Sharper Model-free Reinforcement Learning for Average-reward Markov Decision Processes". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 2023, pp. 5476–5477.

[33] Matthew Zurek and Yudong Chen. "Span-Based Optimal Sample Complexity for Average Reward MDPs". In: *arXiv preprint arXiv:2311.13469* (2023).

A TABULAR SETTING

Central to the analysis of the concentration bound for the approximate Bellman backup is the following concentration bound for scalar-valued self-normalized processes.

Lemma 8 (Concentration of Scalar-Valued Self-Normalized Processes [1])**.** *Let $\{\varepsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Let $\varepsilon_t|\mathcal{F}_{t-1}$ be zero-mean and σ-subgaussian. Let $\{Z_t\}_{t=0}^{\infty}$ be an $\mathbb{R}$-valued stochastic process where $Z_t \in \mathcal{F}_{t-1}$. Assume $W > 0$ is deterministic. Then for any $\delta > 0$, with probability at least $1-\delta$, we have for all $t \geq 0$ that*

$$\frac{(\sum_{s=1}^{t} Z_s \varepsilon_s)^2}{W + \sum_{s=1}^{t} Z_s^2} \leq 2\sigma^2 \log\left(\frac{\sqrt{W + \sum_{s=1}^{t} Z_t^2}}{\delta\sqrt{W}}\right).$$

A.1 Proof of Lemma 3

To show a bound for $|(\widehat{P}_t - P)V_t(s,a)|$ uniformly on $t \in [T]$, we use a covering argument on the function class that captures V_t. Note that the value functions V_t defined in the algorithm always lie in the following function class.

$$\mathcal{V}_{\text{tabular}} = \{v \in \mathbb{R}^{\mathcal{S}} : \text{sp}(v) \leq 2\text{sp}(v^*),\ v(s) \in [0, \tfrac{1}{1-\gamma}] \text{ for all } s \in \mathcal{S}\}.$$

We first bound the error for a fixed value function in $\mathcal{V}_{\text{tabular}}$. Afterward, we will use a covering argument to get a uniform bound over $\mathcal{V}_{\text{tabular}}$.

Lemma 9. *Fix any $V \in \mathcal{V}_{tabular}$. There exists some constant C such that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have:*

$$|(\widehat{P}_t - P)V(s,a)| \leq C sp(v^*)\sqrt{\frac{\log(SAT/\delta)}{N_t(s,a)}}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $t = 1, \ldots, T$.

Proof. Fix any $(s,a) \in \mathcal{S} \times \mathcal{A}$. By definition, we have:

$$(\widehat{P}_t - P)V(s,a) = \frac{1}{N_t(s,a)} \sum_{\tau=1}^{t} \mathbb{I}\{s_\tau = s, a_\tau = a\}[V(s_{\tau+1}) - PV(s,a)].$$

Let $\varepsilon_t = \mathbb{I}\{s_t = s, a_t = a\}[V(s_{t+1}) - PV(s,a)]$, $Z_t = \mathbb{I}\{s_t = s, a_t = a\}$, and $W = 1$. Since the range of ε_t is bounded by $2\text{sp}(v^*)$, it is $\text{sp}(v^*)$-subgaussian. By Lemma 8, we know for some constant C, with probability at least $1 - \frac{\delta}{SA}$, for all $t = 1, \ldots, T$, we have

$$\begin{aligned}
|(\widehat{P}_t - P)V(s,a)| &\leq \frac{2\sum_{s=1}^{t} Z_s \varepsilon_s}{1 + \sum_{s=1}^{t} Z_t^2} \\
&\leq C\text{sp}(v^*)\sqrt{\frac{\log(SA\sqrt{N_t(s,a)}/\delta)}{N_t(s,a)}} \\
&\leq C\text{sp}(v^*)\sqrt{\frac{\log(SAT/\delta)}{N_t(s,a)}}.
\end{aligned}$$

Applying a union bound for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ gives us the desired inequality.

□

We use $\mathcal{N}_\epsilon$ to denote the ϵ-covering number of $\mathcal{V}_{\text{tabular}}$ with respect to the distance $\text{dist}(V, V') = \|V - V'\|_\infty$. Since $\mathcal{V}_{\text{tabular}}$ is contained in $\{v \in \mathbb{R}^{\mathcal{S}} : v(s) \in [0, \frac{1}{1-\gamma}] \text{ for all } s \in \mathcal{S}\}$, of which ϵ-covering number is bounded by $(\frac{1}{\epsilon(1-\gamma)})^{|\mathcal{S}|}$, we have

$$\log \mathcal{N}_\epsilon \leq S \log \frac{1}{\epsilon(1-\gamma)}.$$

Finally, we prove the uniform concentration desired.

Proof of Lemma 3. For any $V \in \mathcal{V}$, we know there exists a $\widetilde{V}$ in the ϵ-covering such that $V = \widetilde{V} + \Delta_V$ where $\sup_s |\Delta_V(s)| \leq \epsilon$. Thus we have

$$|(\widehat{P}_t - P)V_t(s,a)| \leq |(\widehat{P}_t - P)\widetilde{V}(s,a)| + |(\widehat{P}_t - P)\Delta_V(s,a)| \leq |(\widehat{P}_t - P)\widetilde{V}(s,a)| + 2\epsilon.$$

We then apply Lemma 9 and a union bound to obtain:

$$|(\widehat{P}_t - P)V_t(s,a)| \leq C\mathrm{sp}(v^*)\sqrt{\frac{\log(SAT\mathcal{N}_\epsilon/\delta)}{N_t(s,a)}} + 2\epsilon \leq C\mathrm{sp}(v^*)\sqrt{\frac{S\log\frac{SAT}{\delta\epsilon(1-\gamma)}}{N_t(s,a)}} + 2\epsilon.$$

Picking $\epsilon = \frac{1}{\sqrt{T}}$ concludes the proof. $\square$

A.2 Proof of Lemma 4

Proof of Lemma 4. We prove by induction on $t \geq 1$. The base case $t = 1$ is trivial since Algorithm 1 initializes $V_1(\cdot) = \frac{1}{1-\gamma}$, $Q_1(\cdot,\cdot) = \frac{1}{1-\gamma}$. Now, suppose $V_1, \ldots, V_t \geq V^*$ and $Q_1, \ldots, Q_t \geq Q^*$.

We first show that $Q_{t+1}(s,a) \geq Q^*(s,a)$. By the Bellman optimality equation for the discounted setting, we have for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$Q^*(s,a) = r(s,a) + \gamma(PV^*)(s,a).$$

Fix any pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. By the definition of Q_{t+1} in Line 8 of Algorithm 1, we have either $Q_{t+1}(s,a) = Q_t(s,a)$ or $Q_{t+1}(s,a) = r(s,a) + \gamma(\widehat{P}_t V_t)(s,a) + \beta\sqrt{1/N_t(s,a)}$. The former case is done by the induction hypothesis. Suppose $Q_{t+1}(s,a) = r(s,a) + \gamma(\widehat{P}_t V_t - PV^*)(s,a) + \beta\sqrt{1/N_t(s,a)}$. Then,

$$Q_{t+1}(s,a) - Q^*(s,a) = \gamma(\widehat{P}_t V_t - PV^*)(s,a) + \beta\sqrt{1/N_t(s,a)}.$$

By Lemma 3,

$$\begin{aligned}(\widehat{P}_t V_t - PV^*)(s,a) &= (\widehat{P}_t - P)V_t(s,a) + P(V_t - V^*)(s,a) \\ &\geq P(V_t - V^*)(s,a) - C \cdot \mathrm{sp}(v^*)\sqrt{\frac{S\log(SAT/\delta)}{N_t(s,a)}}.\end{aligned}$$

By our choice of the bonus factor β and the induction hypothesis, we get

$$Q_{t+1}(s,a) - Q^*(s,a) \geq P(V_t - V^*)(s,a) \geq 0.$$

Then we show $V_{t+1}(s) \geq V^*(s)$. By the definition of $\widetilde{V}_{t+1}$ in Line 9 of Algorithm 1, we either have $\widetilde{V}_{t+1}(s) = V_t(s)$ or $\widetilde{V}_{t+1}(s) = \max_a Q_{t+1}(s,a)$. The former case is done by the induction hypothesis. Suppose $\widetilde{V}_{t+1}(s) = \max_a Q_{t+1}(s,a)$, then by the optimism of the Q function we just proved, we have

$$\widetilde{V}_{t+1}(s) - V^*(s) = \max_a Q_{t+1}(s,a) - Q^*(s,a_s^*) \geq Q_{t+1}(s,a_s^*) - Q^*(s,a_s^*) \geq 0,$$

where $a_s^* = \mathrm{argmax}_a Q^*(s,a)$. Finally, by the definition of V_{t+1} in Line 10 of Algorithm 1, we have $V_{t+1}(s) = \widetilde{V}_{t+1}(s) \wedge (\min_{s'} \widetilde{V}_{t+1}(s') + \mathrm{sp}(v^*)) \geq \widetilde{V}_{t+1}(s) \geq V^*(s)$. $\square$

B LINEAR MDP SETTING

Central to the analysis of the concentration bound for the approximate Bellman backup is the following concentration bound for scalar-valued self-normalized processes.

Lemma 10 (Concentration of vector-valued self-normalized processes [2])**.** *Let $\{\varepsilon_t\}_{t=1}^\infty$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^\infty$. Let $\varepsilon_t|\mathcal{F}_{t-1}$ be zero-mean and σ-subgaussian. Let $\{\phi_t\}_{t=0}^\infty$ be an $\mathbb{R}^d$-valued stochastic process where $\phi_t \in \mathcal{F}_{t-1}$. Assume Λ_0 is a $d\times d$ positive definite matrix, and let $\Lambda_t = \Lambda_0 + \sum_{s=1}^t \phi_s\phi_s^T$. Then for any $\delta > 0$, with probability at least $1-\delta$, we have for all $t \geq 0$ that*

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

B.1 Concentration Bound for Regression Coefficients

Lemma 11. *Let $V : \mathcal{S} \to [0, B]$ be a bounded function. Then, there exists a parameter $\boldsymbol{w}_V^* \in \mathbb{R}^d$ such that $PV(s,a) = \langle \varphi(s,a), \boldsymbol{w}_V \rangle$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and*

$$\|\boldsymbol{w}_V^*\|_2 \leq B\sqrt{d}.$$

Proof. By Assumption B, we have

$$PV(s,a) = \int_{\mathcal{S}} V(s')P(ds'|s,a) = \int_{\mathcal{S}} V(s')\langle \boldsymbol{\varphi}(s,a), \boldsymbol{\mu}(ds')\rangle = \langle \boldsymbol{\varphi}(s,a), \int_{\mathcal{S}} V(s')d\boldsymbol{\mu}(s')\rangle.$$

Hence, $\boldsymbol{w}_V^* = \int_{\mathcal{S}} V(s')d\boldsymbol{\mu}(s')$ satisfies $PV(s,a) = \langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_V\rangle$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Also, such $\boldsymbol{w}_V$ satisfies

$$\|\boldsymbol{w}_V\|_2 = \left\| \int_{\mathcal{S}} V(s')d\boldsymbol{\mu}(s') \right\|_2 \leq B \left\| \int_{\mathcal{S}} d\boldsymbol{\mu}(s') \right\|_2 \leq B\sqrt{d}$$

where the first inequality holds since $\boldsymbol{\mu}$ is a vector of positive measures and $V(s') \geq 0$. The last inequality is by the boundedness assumption (1) on $\boldsymbol{\mu}(\mathcal{S})$. □

Lemma 12. *Let $\boldsymbol{w}$ be a ridge regression coefficient obtained by regressing $y \in [0, B]$ on $\boldsymbol{x} \in \mathbb{R}^d$ using the dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ so that $\boldsymbol{w} = \Lambda^{-1}\sum_{i=1}^n \boldsymbol{x}_i y_i$ where $\Lambda = \sum_{i=1}^n \boldsymbol{x}\boldsymbol{x}^T + \lambda I$. Then,*

$$\|\boldsymbol{w}\|_2 \leq B\sqrt{dn/\lambda}.$$

Proof. For any unit vector $\boldsymbol{u} \in \mathbb{R}^d$ with $\|\boldsymbol{u}\|_2 = 1$, we have

$$\begin{aligned}
|\boldsymbol{u}^T\boldsymbol{w}| &= \left| \boldsymbol{u}^T\Lambda^{-1}\sum_{i=1}^n \boldsymbol{x}_i y_i \right| \\
&\leq B\sum_{i=1}^n |\boldsymbol{u}^T\Lambda^{-1}\boldsymbol{x}_i| \\
&\leq B\sum_{i=1}^n \sqrt{\boldsymbol{u}^T\Lambda^{-1}\boldsymbol{u}}\sqrt{\boldsymbol{x}_i^T\Lambda^{-1}\boldsymbol{x}_i} \\
&\leq \frac{B}{\sqrt{\lambda}}\sum_{i=1}^n \sqrt{\boldsymbol{x}_i^T\Lambda^{-1}\boldsymbol{x}_i} \\
&\leq \frac{B}{\sqrt{\lambda}}\sqrt{n}\sqrt{\sum_{i=1}^n \boldsymbol{x}_i^T\Lambda^{-1}\boldsymbol{x}_i} \\
&\leq B\sqrt{dn/\lambda}
\end{aligned}$$

where the second inequality and the fourth inequality are by Cauchy-Schwartz, the third inequality is by $\Lambda \succeq \lambda I$, and the last inequality is by Lemma 17.

The desired result follows from the fact that $\|\boldsymbol{w}\|_2 = \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1} |\boldsymbol{u}^T\boldsymbol{w}|$. □

The following self-normalized process bound is an adaptation of Lemma D.4 in Jin et al. [18]. Their proof defines the bound B to be a value that satisfies $\|V\|_\infty \leq B$. Upon observing their proof, it is easy to see that we can strengthen their result to require only $\mathrm{sp}(V) \leq B$. The following lemma is the strengthened version.

Lemma 13 (Adaptation of Lemma D.4 in Jin et al. [18])**.** *Let $\{x_t\}_{t=1}^\infty$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^\infty$. Let $\{\phi_t\}_{t=0}^\infty$ be a $\mathbb{R}^d$-valued stochastic process where $\phi_t \in \mathcal{F}_{t-1}$, and $\|\phi_t\|_2 \leq 1$. Let $\Lambda_n = \lambda I + \sum_{t=1}^n \phi_t\phi_t^T$. Then for any $\delta > 0$ and any given function class $\mathcal{V}$, with probability at least $1-\delta$, for all $n \geq 0$, and any $V \in \mathcal{V}$ satisfying $sp(V) \leq H$, we have*

$$\left\| \sum_{t=1}^n \phi_t(V(x_t) - \mathbb{E}[V(x_t)|\mathcal{F}_{t-1}]) \right\|_{\Lambda_n^{-1}}^2 \leq 4H^2\left[\frac{d}{2}\log\left(\frac{n+\lambda}{\lambda}\right) + \log\frac{\mathcal{N}_\varepsilon}{\delta}\right] + \frac{8n^2\varepsilon^2}{\lambda}$$

where $\mathcal{N}_\varepsilon$ is the ε-covering number of $\mathcal{V}$ with respect to the distance $dist(V, V') = \sup_x |V(x) - V'(x)|$.

Lemma 14 (Adaptation of Lemma D.6 in Jin et al. [18]). *Let $\mathcal{V}_{linear}$ be a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following parametric form*

$$V(\cdot) = \max_a \boldsymbol{w}^T \boldsymbol{\varphi}(\cdot, a) + v + \beta\sqrt{\boldsymbol{\varphi}(\cdot, a)^T \Lambda^{-1} \boldsymbol{\varphi}(\cdot, a)} \wedge M \tag{2}$$

where the parameters $(\boldsymbol{w}, \beta, v, \Lambda, M)$ satisfy $\|\boldsymbol{w}\| \le L$, $\beta \in [0, B]$, $v \in [0, D]$, $M \ge 0$ and the minimum eigenvalue satisfies $\lambda_{min}(\Lambda) \ge \lambda$. Assume $\|\boldsymbol{\varphi}(s, a)\| \le 1$ for all (s, a) pairs, and let $\mathcal{N}_\varepsilon$ be the ε-covering number of $\mathcal{V}$ with respect to the distance $dist(V, V') = \sup_x |V(x) - V'(x)|$. Then

$$\log \mathcal{N}_\varepsilon \le d \log(1 + 8L/\varepsilon) + \log(1 + 4D/\varepsilon) + d^2 \log[1 + 8d^{1/2} B^2/(\lambda\varepsilon^2)].$$

For the next lemma, we define value functions $V_u^{(t)}$ to be the functions obtained by the following value iteration (analogous to Line 7-12 in Algorithm 2):

$V_{T+1}^{(t)}(\cdot) \leftarrow \frac{1}{1-\gamma}$, $V_{T+1}^{(t)}(\cdot) \leftarrow \frac{1}{1-\gamma}$.
for $u = T, T-1, \ldots, 1$ **do**
  $\boldsymbol{w}_{u+1}^{(t)} \leftarrow \bar{\Lambda}_t^{-1} \sum_{\tau=1}^t \boldsymbol{\varphi}(s_\tau, a_\tau)(V_{u+1}^{(t)}(s_{\tau+1}) - \min_{s'} \widetilde{V}_{u+1}^{(t)}(s'))$.
  $\widetilde{Q}_u^{(t)}(\cdot,\cdot) \leftarrow \left(r(\cdot,\cdot) + \gamma(\langle \boldsymbol{\varphi}(\cdot,\cdot), \boldsymbol{w}_{u+1}^{(t)}\rangle + \min_{s'} \widetilde{V}_{u+1}^{(t)}(s') + \beta\|\boldsymbol{\varphi}(\cdot,\cdot)\|_{\bar{\Lambda}_t^{-1}})\right) \wedge \frac{1}{1-\gamma}$.
  $\widetilde{V}_u^{(t)}(\cdot) \leftarrow \max_a \widetilde{Q}_u^{(t)}(\cdot, a)$.
  $V_u^{(t)}(\cdot) \leftarrow \widetilde{V}_u^{(t)}(\cdot) \wedge (\min_{s'} \widetilde{V}_u^{(t)}(s') + H)$.

With this definition, we show a high-probability bound on $\| \sum_{\tau=1}^t \boldsymbol{\varphi}(s_\tau, a_\tau)[V_u^{(t)}(s_{\tau+1}) - PV_u^{(t)}(s_\tau, a_\tau)]\|_{\bar{\Lambda}_t^{-1}}$ uniformly on $u \in [T]$ and $t \in [T]$. Since the tuple $(t_k - 1, V_u^k, \Lambda_k)$ encountered in Algorithm 2 is the same as the pair $(t, V_u^{(t)}, \bar{\Lambda}_t)$ for some $t \in [T]$, the uniform bound implies bound on $\| \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau)[V_u^k(s_{\tau+1}) - PV_u^k(s_\tau, a_\tau)]\|_{\Lambda_k^{-1}}$ for all episode k.

Lemma 15 (Adaptation of Lemma B.3 in Jin et al. [18]). *Under the linear MDP setting in Theorem 5 for the γ-LSCVI-UCB algorithm with clipping oracle (Algorithm 2), let c_β be the constant in the definition of $\beta = c_\beta H d\sqrt{\log(dT/\delta)}$. There exists an absolute constant C that is independent of c_β such that for any fixed $\delta \in (0, 1)$, the event $\mathcal{E}$ defined by*

$$\begin{aligned}&\forall u \in [T],\ t \in [T]: \\ &\quad \left\| \sum_{\tau=1}^t \boldsymbol{\varphi}(s_\tau, a_\tau)[V_u^{(t)}(s_{\tau+1}) - PV_u^{(t)}(s_\tau, a_\tau)] \right\|_{\bar{\Lambda}_t^{-1}} \le C \cdot Hd\sqrt{\log((c_\beta + 1)dT/\delta)}\end{aligned}$$

satisfies $P(\mathcal{E}) \ge 1 - \delta$.

Proof. For all $t = 1, \ldots, T$, by Lemma 12, we have $\|\boldsymbol{w}_t\|_2 \le H\sqrt{dt/\lambda}$. Hence, by combining Lemma 14 and Lemma 13, for any $\varepsilon > 0$ and any fixed pair $(u, t) \in [T] \times [T]$, we have with probability at least $1 - \delta/T^2$ that

$$\begin{aligned}&\left\| \sum_{\tau=1}^t \boldsymbol{\varphi}(s_\tau, a_\tau)[V_u(s_{\tau+1}) - PV_u(s_\tau, a_\tau)] \right\|_{\bar{\Lambda}_t^{-1}}^2 \\ &\qquad \le 4H^2 \left[\frac{2}{d}\log\left(\frac{t+\lambda}{\lambda}\right) + d\log\left(1 + \frac{4H\sqrt{dt}}{\varepsilon\sqrt{\lambda}}\right) + d^2\log\left(1 + \frac{8d^{1/2}\beta^2}{\varepsilon^2\lambda}\right) + \log\left(\frac{T^2}{\delta}\right)\right] + \frac{8t^2\varepsilon^2}{\lambda}\end{aligned}$$

where we use the fact that $\tau_t \le t$. Using a union bound over $(u, t) \in [T] \times [T]$ and choosing $\varepsilon = Hd/t$ and $\lambda = 1$, there exists an absolute constant $C > 0$ independent of c_β such that, with probability at least $1 - \delta$,

$$\left\| \sum_{\tau=1}^t \varphi(s_\tau, a_\tau)[V_u(s_{\tau+1}) - PV_u(s_\tau, a_\tau)] \right\|_{\bar{\Lambda}_t^{-1}}^2 \le C^2 \cdot d^2 H^2 \log((c_\beta + 1)dT/\delta),$$

which concludes the proof. □

Proof of Lemma 6. We prove under the event $\mathcal{E}$ defined in Lemma 15. For convenience, we introduce the notation $\bar{V}_u^k(s) = V_u^k(s) - \min_{s'} V_u^k(s')$. With this notation, we can write

$$\boldsymbol{w}_u^k = \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau) \bar{V}_u^k(s_{\tau+1}).$$

We can decompose $\langle \boldsymbol{\phi}, \boldsymbol{w}_u^k \rangle$ as

$$\langle \boldsymbol{\phi}, \boldsymbol{w}_u^k \rangle = \underbrace{\langle \boldsymbol{\phi}, \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau) P\bar{V}_u^k(s_\tau, a_\tau) \rangle}_{(a)} + \underbrace{\langle \boldsymbol{\phi}, \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau) (\bar{V}_u^k(s_{\tau+1}) - P\bar{V}_u^k(s_\tau, a_\tau)) \rangle}_{(b)}.$$

Since $\boldsymbol{w}_u^{k*} = \int \bar{V}_u^k(s) d\boldsymbol{\mu}(s)$ and $\bar{V}_u^k(s) \in [0, H]$ for all $s \in \mathcal{S}$, it follows by Lemma 11 that $\|\boldsymbol{w}_u^{k*}\|_2 \leq H\sqrt{d}$. Hence, the first term (a) in the display above can be bounded as

$$\begin{aligned}
\langle \boldsymbol{\phi}, \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau) P\bar{V}_u^k(s_\tau, a_\tau) \rangle &= \langle \boldsymbol{\phi}, \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau) \boldsymbol{\varphi}(s_\tau, a_\tau)^T \boldsymbol{w}_u^{k*} \rangle \\
&= \langle \boldsymbol{\phi}, \boldsymbol{w}_u^{k*} \rangle - \lambda \langle \boldsymbol{\phi}, \Lambda_k^{-1} \boldsymbol{w}_u^{k*} \rangle \\
&\leq \langle \boldsymbol{\phi}, \boldsymbol{w}_u^{k*} \rangle + \lambda \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}} \|\boldsymbol{w}_u^{k*}\|_{\Lambda_k^{-1}} \\
&\leq \langle \boldsymbol{\phi}, \boldsymbol{w}_u^{k*} \rangle + H\sqrt{\lambda d} \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}}
\end{aligned}$$

where the first inequality is by Cauchy-Schwartz and the second inequality is by Lemma 11. Under the event $\mathcal{E}$ defined in Lemma 15, the second term (b) can be bounded by

$$\begin{aligned}
&\langle \boldsymbol{\phi}, \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau) (\bar{V}_u^k(s_{\tau+1}) - P\bar{V}_u^k(s_\tau, a_\tau)) \\
&\qquad \leq \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}} \left\| \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau) (V_u^k(s_{\tau+1}) - PV_u^k(s_\tau, a_\tau)) \right\|_{\Lambda_k^{-1}} \\
&\qquad \leq C \cdot Hd \sqrt{\log((c_\beta + 1) dT/\delta)} \cdot \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}}.
\end{aligned}$$

Combining the two bounds and rearranging, we get

$$\langle \boldsymbol{\phi}, \boldsymbol{w}_u^k - \boldsymbol{w}_u^{k*} \rangle \leq C \cdot Hd \sqrt{(\log(c_\beta + 1) dT/\delta)} \cdot \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}}$$

for some absolute constant C independent of c_β. Lower bound of $\langle \boldsymbol{\phi}, \boldsymbol{w}_u^k - \boldsymbol{w}_u^{k*} \rangle$ can be shown similarly, establishing

$$|\langle \boldsymbol{\phi}, \boldsymbol{w}_u^k - \boldsymbol{w}_u^{k*} \rangle| \leq C \cdot Hd \sqrt{\log((c_\beta + 1) dT/\delta)} \cdot \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}}.$$

It remains to show that there exists a choice of absolute constant c_β such that

$$C \sqrt{\log(c_\beta + 1) + \log(dT/\delta)} \leq c_\beta \sqrt{\log(dT/\delta)}.$$

Noting that $\log(dT/\delta) \geq \log 2$, this can be done by choosing an absolute constant c_β that satisfies $C\sqrt{\log 2 + \log(c_\beta + 1)} \leq c_\beta \sqrt{\log 2}$. □

B.2 Optimism

Proof of Lemma 7. We prove under the event $\mathcal{E}$ defined in Lemma 15. Fix any episode index $k > 1$. We prove by induction on $u = T+1, T, \ldots, 1$. The base case $u = T+1$ is trivial since $V_{T+1}^k(s) = \frac{1}{1-\gamma} \geq V^*(s)$ for all $s \in \mathcal{S}$ and $Q_{T+1}^k(s, a) = \frac{1}{1-\gamma} \geq Q^*(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Now, suppose the optimism results $V_{u+1}^k(s) \geq V^*(s)$ and $Q_{u+1}^k(s, a) \geq Q^*(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ hold for some $u \in [T]$. For convenience, we use the notation $\bar{V}_u^k(s) = V_u^k(s) - \min_{s'} V_u^k(s')$. Using the concentration bounds of regression coefficients $\boldsymbol{w}_u^k$ provided in Lemma 6, which holds under the event $\mathcal{E}$, we can lower bound $Q_u^k(s,a)$ as follows.

$$\begin{aligned}
Q_u^k(s,a) &= \Big(r(s,a) + \gamma(\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_{u+1}^k\rangle + \min_{s'} V_{u+1}^k(s') + \beta\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_k^{-1}}\Big) \wedge \frac{1}{1-\gamma} \\
&\geq \Big(r(s,a) + \gamma(\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_{u+1}^k{}^*\rangle + \min_{s'} V_{u+1}^k(s'))\Big) \wedge \frac{1}{1-\gamma} \\
&= (r(s,a) + \gamma P V_{u+1}^k(s,a)) \wedge \frac{1}{1-\gamma} \\
&\geq (r(s,a) + \gamma P V^*(s,a)) \wedge \frac{1}{1-\gamma} \\
&= Q^*(s,a)
\end{aligned}$$

where $\boldsymbol{w}_{u+1}^k{}^*$ is a parameter that satisfies $\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_{u+1}^k{}^*\rangle = P\bar{V}_{u+1}^k(s,a)$. The second inequality is by the induction hypothesis $V_{u+1}^k \geq V^*$ and the last equality is by the Bellman optimality equation for the discounted setting and the fact that $Q^* \leq \frac{1}{1-\gamma}$.

We established $Q_u^k(s,a) \geq Q^*(s,a)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. It remains to show that $V_u^k(s) \geq V^*(s)$ for all $s \in \mathcal{S}$. Recall that the algorithm defines $\widetilde{V}_u^k(\cdot) = \max_a Q_u^k(\cdot, a)$. Hence, for all $s \in \mathcal{S}$, we have

$$\begin{aligned}
\widetilde{V}_u^k(s) - V^*(s) &= \max_a Q_u^k(s,a) - V^*(s) \\
&\geq Q_u^k(s,a_s^*) - Q^*(s,a_s^*) \\
&\geq 0
\end{aligned}$$

where we use the notation $a_s^* = \operatorname{argmax}_a Q^*(s,a)$ so that $V^*(s) = Q^*(s,a_s^*)$, establishing $\widetilde{V}_u^k(s) \geq V^*(s)$ for all $s \in \mathcal{S}$. Hence, for all $s \in \mathcal{S}$, we have

$$\begin{aligned}
V_u^k(s) &= \widetilde{V}_u^k(s) \wedge (\min_{s'} \widetilde{V}_u^k(s') + 2\cdot \mathrm{sp}(v^*)) \\
&\geq V^*(s) \wedge (\min_{s'} V^*(s') + 2\cdot \mathrm{sp}(v^*)) \\
&= V^*(s)
\end{aligned}$$

where the last equality is due to $\mathrm{sp}(V^*) \leq 2\cdot\mathrm{sp}(v^*)$ by Lemma 1. By induction, the proof for the optimism results $V_u^k(s) \geq V^*(s)$ and $Q_u^k(s,a) \geq Q^*(s,a)$ for $u = T+1, T, \ldots, 1$ is complete. ☐

B.3 Access to $\min_{s'} V^*(s')$

In this section, we demonstrate that using $\min_{s'} V^*(s')$ for clipping instead of $\min_{s'} V_u^k(s')$ at the clipping step in Algorithm 2 achieves the same regret bound.

Since the only part of the proof affected by the modified clipping is the optimism proof, we provide the proof for the optimism lemma only.

Lemma 16 (Optimism)**.** *Under the linear MDP setting, consider running a modified version of Algorithm 2 that uses $\min_{s'} V^*(s')$ for clipping instead of $\min_{s'} V_u^k(s')$ at the uth value iteration step in episode k. Using the same input $H = 2\cdot sp(v^*)$ for the modified algorithm guarantees with probability at least $1-\delta$ that for all $u = 1,\ldots,T$ and for all $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$V_u^k(s) \geq V^*(s), \quad Q_u^k(s,a) \geq Q^*(s,a).$$

Proof. We prove under the event $\mathcal{E}$ defined in Lemma 15. Fix any episode index $k > 1$. We prove by induction on $u = T+1, T, \ldots, 1$. The base case $u = T+1$ is trivial since $V_{T+1}^k(s) = \frac{1}{1-\gamma} \geq V^*(s)$ for all $s \in \mathcal{S}$ and $Q_{T+1}^k(s,a) = \frac{1}{1-\gamma} \geq Q^*(s,a)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Now, suppose the optimism results $V_{u+1}^k(s) \geq V^*(s)$ and $Q_{u+1}^k(s,a) \geq Q^*(s,a)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ hold for some $u \in [T]$. For convenience, we use the notation $\bar{V}_u^k(s) = V_u^k(s) - \min_{s'} V^*(s')$. Using the concentration bounds of regression coefficients $\boldsymbol{w}_{u+1}^k$ provided in Lemma 6, which holds under the event $\mathcal{E}$, we can lower bound $Q_u^k(s,a)$ as follows.

$$
\begin{aligned}
Q_u^k(s,a) &= \Big(r(s,a) + \gamma(\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_{u+1}^k\rangle + \min_{s'} V^*(s') + \beta\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_k^{-1}}\Big) \wedge \frac{1}{1-\gamma} \\
&\geq \Big(r(s,a) + \gamma(\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_{u+1}^k{}^*\rangle + \min_{s'} V^*(s'))\Big) \wedge \frac{1}{1-\gamma} \\
&= (r(s,a) + \gamma P V_{u+1}^k(s,a)) \wedge \frac{1}{1-\gamma} \\
&\geq (r(s,a) + \gamma P V^*(s,a)) \wedge \frac{1}{1-\gamma} \\
&= Q^*(s,a)
\end{aligned}
$$

where $\boldsymbol{w}_{u+1}^k{}^*$ is a parameter that satisfies $\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_{u+1}^k{}^*\rangle = P\bar{V}_{u+1}^k(s,a)$. The second inequality is by the induction hypothesis $V_{u+1}^k \geq V^*$ and the last equality is by the Bellman optimality equation for the discounted setting and the fact that $Q^* \leq \frac{1}{1-\gamma}$.

We established $Q_u^k(s,a) \geq Q^*(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. It remains to show that $V_u^k(s) \geq V^*(s)$ for all $s \in \mathcal{S}$. Recall that the algorithm defines $\widetilde{V}_u^k(\cdot) = \max_a \widetilde{Q}_u^k(\cdot, a)$. Hence, for all $s \in \mathcal{S}$, we have

$$
\begin{aligned}
\widetilde{V}_u^k(s) - V^*(s) &= \max_a Q_u^k(s,a) - V^*(s) \\
&\geq Q_u^k(s, a_s^*) - Q^*(s, a_s^*) \\
&\geq 0
\end{aligned}
$$

where we use the notation $a_s^* = \operatorname{argmax}_a Q^*(s,a)$ so that $V^*(s) = Q^*(s, a_s^*)$, establishing $\widetilde{V}_u^k(s) \geq V^*(s)$ for all $s \in \mathcal{S}$. Hence, for all $s \in \mathcal{S}$, we have

$$
\begin{aligned}
V_u^k(s) &= \widetilde{V}_u^k(s) \wedge (\min_{s'} V^*(s') + 2 \cdot \mathrm{sp}(v^*)) \\
&\geq V^*(s) \wedge (\min_{s'} V^*(s') + 2 \cdot \mathrm{sp}(v^*)) \\
&= V^*(s)
\end{aligned}
$$

where the last equality is due to $\mathrm{sp}(V^*) \leq 2 \cdot \mathrm{sp}(v^*)$ by Lemma 1. By induction, the proof for the optimism results $V_u^k(s) \geq V^*(s)$ and $Q_u^k(s,a) \geq Q^*(s,a)$ for $u = T+1, T, \dots, 1$ is complete. ☐

C COMPUTATIONAL COMPLEXITY

C.1 γ-LSCVI-UCB (Algorithm 2)

Our algorithm γ-LSCVI-UCB runs in episodes and the number of episodes is bounded by $\mathcal{O}(d \log T)$. In each episode, value iteration is run for at most T iterations. In each iteration u in episode k, one evaluation of $\min_{s'} \widetilde{V}_u^k(s')$, t_k evaluations of $V_u^k(\cdot)$ of $V_u^k(\cdot)$ and a multiplication of $d \times d$ matrix (Λ_k^{-1}) and a d-dimensional vector is required.

One evaluation of $\min_{s'} \widetilde{V}_u^k(s')$ involves S evaluations of $\widetilde{V}_u^k(\cdot)$. One evaluation of $\widetilde{V}_u^k(\cdot)$ involves A evaluations of $Q_u^k(\cdot,\cdot)$. One evaluation of $Q_u^k(\cdot,\cdot)$ requires $\mathcal{O}(d^2)$ operations. In total, one evaluation of $\min_{s'} \widetilde{V}_u^k(s')$ requires $\mathcal{O}(d^2 SA)$ operations.

Now, computing $\boldsymbol{w}_u^k$ requires evaluating $V_{u+1}^k(\cdot)$ for at most T states, which requires $\mathcal{O}(d^2 AT)$ operations; adding at most T d-dimensional vectors, which requires Td operations; and multiplying by $d \times d$ matrix, which requires d^2 operations.

In total, computing $\boldsymbol{w}_u^k$ requires $\mathcal{O}(d^2 A(S+T))$ operations. Hence, running at most T value iterations in each episode requires $\mathcal{O}(d^2 A(S+T)T)$ operations, and since there are at most $\mathcal{O}(d \log T)$ episodes, total operations for the algorithm is $\widetilde{\mathcal{O}}(d^3 A(S+T)T)$, which is polynomial in d, S, A, T.

C.2 FOPO [28]

In this section, we provide time complexity analysis of the FOPO algorithm [28]. The algorithm is shown in Algorithm 3.

Algorithm 3: FOPO

Input: $\delta \in (0,1)$, $\lambda = 1$, $\beta = 20(2 + \mathrm{sp}(v^*))d\sqrt{\log(T/\delta)}$

Initialize: $\Lambda_1 = \lambda I$

Receive initial state s_1.

for *time step* $t = 1, \ldots, T$ **do**

  Solve the following optimization problem to get w_t:

$$
\begin{aligned}
\max_{w_t, b_t \in \mathbb{R}^d, J_t \in \mathbb{R}} \quad & J_t \\
\text{subject to} \quad & w_t = \Lambda_t^{-1} \sum_{\tau=1}^{t-1} (\boldsymbol{\varphi}(s_\tau, a_\tau)(r(s_\tau, a_\tau) - J_t + \max_a \langle \boldsymbol{\varphi}(s_{\tau+1}, a), w_t \rangle) + b_t) \\
& \|b_t\|_{\Lambda_t} \leq \beta \\
& \|w_t\| \leq (2 + \mathrm{sp}(v^*))\sqrt{d}
\end{aligned}
$$

The bottleneck of the algorithm is solving the optimization problem. The algorithm needs to solve the optimization problem $\mathcal{O}(d \log T)$ times since the number of episodes is $\mathcal{O}(d \log T)$. Since there is no efficient way of solving the fixed point optimization problem to the best of our knowledge, we provide an analysis of the time complexity of a brute force approach for approximately solving the problem. The brute force approach does a grid search on the optimization variables w_t, b_t, and J_t.

Consider the following grids:

$$
\begin{aligned}
G_w(\Delta) &= \{\Delta \cdot (k_1, \ldots, k_d) : \pm k_1, \ldots, \pm k_d \in [\lfloor (2 + \mathrm{sp}(v^*))/\Delta \rfloor]\} \\
G_b(\Delta) &= \{\Delta \cdot (k_1, \ldots, k_d) : \pm k_1, \ldots, \pm k_d \in [\lfloor T\beta/(\sqrt{d}\Delta) \rfloor]\} \\
G_J(\Delta) &= \{\Delta k : k \in [\lfloor 1/\Delta \rfloor]\}.
\end{aligned}
$$

The grids are designed such that the constraints $\|w\|_2 \leq (2 + \mathrm{sp}(v^*))\sqrt{d}$, $\|b\|_{\Lambda_t} \leq \beta$ are satisfied for all $w \in G_w(\Delta)$ and $b \in G_b(\Delta)$ and $J \in [0,1]$ for all $J \in G_J(\Delta)$. Also, $G_w(\Delta)$, $G_b(\Delta)$ and $G_J(\Delta)$ are Δ-covering with respect to $\|\cdot\|_\infty$ of $\mathcal{B}_d((2 + \mathrm{sp}(v^*))\sqrt{d})$, $\mathcal{B}_d(T\beta)$ and $[0,1]$ respectively, where $\mathcal{B}_d(r)$ is a d-dimensional ball of radius r.

Denote by $\Delta_t(w, b, J)$ the difference between the left hand side and the right hand side of the fixed point equation at time step t, i.e.,

$$
\Delta_t(w, b, J) = w - \Lambda_t^{-1} \sum_{\tau=1}^{t-1} (\boldsymbol{\varphi}(s_\tau, a_\tau)(r(s_\tau, a_\tau) - J + \max_a \langle \boldsymbol{\varphi}(s_{\tau+1}, a), w \rangle) + b).
$$

Let w_t^*, b_t^*, J_t^* be the solution to the fixed point problem, which may not lie in the grids. Then, $\Delta_t(w_t^*, b_t^*, J_t^*) = 0$. Let $\widetilde{w}_t^* \in G_w(\Delta)$, $\widetilde{b}_t^* \in G_b(\Delta)$ and $\widetilde{J}^* \in G_J(\Delta)$ be the grid points closest to w_t^*, b_t^*, J_t^*, respectively. Then,

$$
\begin{aligned}
&\|\Delta_t(\widetilde{w}_t^*, \widetilde{b}_t^*, \widetilde{J}^*)\|_2 \\
&= \|\Delta_t(\widetilde{w}_t^*, \widetilde{b}_t^*, \widetilde{J}^*) - \Delta_t(w_t^*, b_t^*, J_t^*)\|_2 \\
&= \|\widetilde{w}_t^* - w_t^* - \Lambda_t^{-1} \sum_{\tau=1}^{t-1} (\boldsymbol{\varphi}(s_\tau, a_\tau)(J_t^* - \widetilde{J}_t^* + \max_a \langle \boldsymbol{\varphi}(s_{\tau+1}, a), \widetilde{w}_t^* \rangle - \max_a \langle \boldsymbol{\varphi}(s_{\tau+1}, a), w_t^* \rangle) + \widetilde{b}_t^* - b_t^*)\|_2 \\
&\leq \|\widetilde{w}_t^* - w_t^*\|_2 + \sum_{\tau=1}^{t-1} |J_t^* - \widetilde{J}_t^*| \|\Lambda_t^{-1}\|_2 \|\boldsymbol{\varphi}(s_\tau, a_\tau)\|_2 + \sum_{\tau=1}^{t-1} \max_a |\langle \boldsymbol{\varphi}(s_{\tau+1}, a), \widetilde{w}_t^* - w_t^* \rangle| \|\Lambda_t^{-1}\|_2 \|\boldsymbol{\varphi}(s_\tau, a_\tau)\|_2 \\
&\qquad + \sum_{\tau=1}^{t-1} \|\widetilde{b}_t^* - b_t^*\|_2 \|\Lambda_t^{-1}\|_2 \\
&\leq \Delta\sqrt{d} + T\Delta + T\Delta\sqrt{d} + T\Delta\sqrt{d} \\
&\leq \mathcal{O}(T\sqrt{d}\Delta).
\end{aligned}
$$

Hence, the solution $\widetilde{w}_t, \widetilde{b}_t, \widetilde{J}_t$ obtained by the grid search satisfies

$$\|\Delta_t(\widetilde{w}_t, \widetilde{b}_t, \widetilde{J}_t)\|_2 \le \|\Delta_t(\widetilde{w}_t^*, \widetilde{b}_t^*, \widetilde{J}_t^*)\|_2 \le \mathcal{O}(T\sqrt{d}\Delta).$$

Inspecting the proof in Appendix C.1 in Wei et al. [28], it can be seen that the additional regret incurred by approximating the solution of the fixed point problem is

$$\sum_{t=1}^{T} \langle \boldsymbol{\varphi}(s_t, a_t), \Delta_t(\widetilde{w}_t, \widetilde{b}_t, \widetilde{J}_t)\rangle \le T^2\sqrt{d}\Delta.$$

Choosing the grid size Δ to be $\mathcal{O}(1/\sqrt{T^3})$ guarantees the additional regret does not affect the order of the total regret. Since the grid search requires $\mathcal{O}(((1+\mathrm{sp}(v^*))/\Delta)^d \times (T\beta/(\sqrt{d}\Delta))^d \times (1/\Delta)) = \widetilde{\mathcal{O}}((T(1+\mathrm{sp}(v^*))^2\sqrt{d})^d (1/\Delta)^{2d+1})$ evaluations of the fixed point equation, the grid search method requires $\widetilde{\mathcal{O}}(T^{4d+3/2}(1+\mathrm{sp}(v^*))^{2d} d^{d/2})$ evaluations, and each evaluation requires $\mathcal{O}(d^2+T)$ operations. In total, FOPO can be run using the brute force grid search method with time complexity $\widetilde{\mathcal{O}}(T^{4d+5/2}(1+\mathrm{sp}(v^*))^{2d} d^{d/2+3})$, which is exponential in d.

C.3 LOOP [15]

The LOOP algorithm [15] solves the following optimization problem every episode:

$$\begin{aligned}
&\max_{w_t \in \mathbb{B}_d(R), J_t \in [0,1]} \quad J_t \\
&\text{subject to} \quad \sum_{\tau=1}^{t-1} (\langle \boldsymbol{\varphi}(s_\tau, a_\tau), \boldsymbol{w}_t\rangle - r(s_\tau, a_\tau) - \max_a \langle \boldsymbol{\varphi}(s_{\tau+1}, a), \boldsymbol{w}_t\rangle + J_t)^2 \\
&\qquad\qquad \min_{w' \in \mathbb{B}_d(R), J' \in [0,1]} \sum_{\tau=1}^{t-1} (\langle \boldsymbol{\varphi}(s_\tau, a_\tau), \boldsymbol{w}_t\rangle - r(s_\tau, a_\tau) - \max_a \langle \boldsymbol{\varphi}(s_{\tau+1}, a), \boldsymbol{w}'\rangle + J')^2 \le \beta
\end{aligned}$$

where $R = \frac{1}{2}\mathrm{sp}(v^*)\sqrt{d}$ and $\beta = \widetilde{\mathcal{O}}(\mathrm{sp}(v^*)d)$. To the best of our knowledge, there is no computationally efficient way of solving this problem. Solving the problem by grid search involves looping over $\mathrm{poly}(T^d)$ grid points for $\boldsymbol{w}_t$ and J_t. Also, checking the constraint for each grid point requires $\mathrm{poly}(T, d, A)$ operations. Hence, the total time complexity of the algorithm is $\mathrm{poly}(T^d, d, A)$.

D OTHER TECHNICAL LEMMAS

Lemma 17 (Lemma D.1 in Jin et al. [18]). *Let $\Lambda_t = \sum_{i=1}^t \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T + \lambda I$ where $\boldsymbol{\phi}_i \in \mathbb{R}^d$ and $\lambda > 0$. Then,*

$$\sum_{i=1}^{t} \boldsymbol{\phi}_i^T \Lambda_t^{-1} \boldsymbol{\phi}_i \le d.$$

Lemma 18 (Lemma 11 in Abbasi-Yadkori et al. [2]). *Let $\{\boldsymbol{\phi}_t\}_{t\ge 1}$ be a bounded sequence in $\mathbb{R}^d$ with $\|\boldsymbol{\phi}_t\|_2 \le 1$ for all $t \ge 1$. Let $\Lambda_0 = I$ and $\Lambda_t = \sum_{i=1}^t \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T + I$ for $t \ge 1$. Then,*

$$\sum_{i=1}^{t} \boldsymbol{\phi}_i^T \Lambda_{i-1}^{-1} \boldsymbol{\phi}_i \le 2\log\det(\Lambda_t) \le 2d\log(1+t).$$

Lemma 19 (Lemma 12 in Abbasi-Yadkori et al. [2]). *Suppose $A, B \in \mathbb{R}^{d\times d}$ are two positive definite matrices satisfying $A \succeq B$. Then, for any $\boldsymbol{x} \in \mathbb{R}^d$, we have*

$$\|\boldsymbol{x}\|_A \le \|\boldsymbol{x}\|_B \sqrt{\frac{\det(A)}{\det(B)}}.$$

Lemma 20 (Bound on number of episodes). *The number of episodes K in Algorithm 2 is bounded by*

$$K \le d\log_2\left(1 + \frac{T}{\lambda d}\right).$$

Table 2: Comparison of algorithms for infinite-horizon average-reward RL in tabular setting

Algorithm	Regret $\widetilde{\mathcal{O}}(\cdot)$	Assumption	Computation
UCRL2 [4]	$DS\sqrt{AT}$	Bounded diameter	Efficient
REGAL [7]	$\text{sp}(v^*)\sqrt{SAT}$	Weakly communicating	Inefficient
PSRL [23]	$\text{sp}(v^*)S\sqrt{AT}$	Weakly communicating	Efficient
OSP [22]	$\sqrt{t_{\text{mix}}SAT}$	Ergodic	Inefficient
SCAL [10]	$\text{sp}(v^*)S\sqrt{AT}$	Weakly communicating	Efficient
UCRL2B [9]	$S\sqrt{DAT}$	Bounded diameter	Efficient
EBF [31]	$\sqrt{\text{sp}(v^*)SAT}$	Weakly communicating	Inefficient
γ-UCB-CVI (Ours)	$\text{sp}(v^*)S\sqrt{AT}$	Bellman optimality equation	Efficient
Optimistic Q-learning [29]	$\text{sp}(v^*)(SA)^{\frac{1}{3}}T^{\frac{2}{3}}$	Weakly communicating	Efficient
MDP-OOMD [29]	$\sqrt{t_{\text{mix}}^3\eta AT}$	Ergodic	Efficient
UCB-AVG [32]	$\text{sp}(v^*)S^5A^2\sqrt{T}$	Weakly communicating	Efficient
Lower bound [4]	$\Omega(\sqrt{DSAT})$		

Proof. Let $\{\Lambda_k\}_{k=1}^K$ and $\{\bar{\Lambda}_t\}_{t=0}^T$ be as defined in Algorithm 2. Note that

$$\text{tr}(\bar{\Lambda}_T) = \text{tr}(\lambda I_d) + \sum_{t=1}^T \text{tr}(\varphi(s_t,a_t)\varphi(s_t,a_t)^T) = \lambda d + \sum_{t=1}^T \|\varphi(s_t,a_t)\|_2^2 \leq \lambda d + T.$$

By the AM–GM inequality, we have

$$\det(\bar{\Lambda}_T) \leq \left(\frac{\text{tr}(\bar{\Lambda}_T)}{d}\right)^d \leq \left(\frac{\lambda d + T}{d}\right)^d.$$

Since we update Λ_k only when $\det(\bar{\Lambda}_t)$ doubles, $\det(\bar{\Lambda}_T) \geq \det(\Lambda_K) \geq \det(\Lambda_1) \cdot 2^K = \lambda^d \cdot 2^K$. Thus, we obtain

$$K \leq d \log_2\left(1 + \frac{T}{\lambda d}\right)$$

as desired.

□

E ADDITIONAL RELATED WORK

Infinite-Horizon Average-Reward Setting with Tabular MDP We focus on works on infinite-horizon average-reward setting with tabular MDP that assume either the MDP is weakly communicating or the MDP has a bounded diameter. For other works and comparisons, see Table 2. Seminal work by Auer et al. [4] on infinite-horizon average-reward setting in tabular MDPs laid the foundation for the problem. Their model-based algorithm called UCRL2 constructs a confidence set on the transition model and run an extended value iteration that involves choosing the optimistic model in the confidence set each iteration. They achieve a regret bound of $\mathcal{O}(DS\sqrt{AT})$ where D is the diameter of the true MDP. Bartlett et al. [7] improve the regret bound of UCRL2 by restricting the confidence set of the model to only include models such that the span of the induced optimal value function is bounded. Their algorithm, called REGAL, achieves a regret bound that scales with the span of the optimal value function $\text{sp}(v^*)$ instead of the diameter of the MDP. However, REGAL is computationally inefficient. Fruit et al. [10] propose a model-based algorithm called SCAL, which is a computationally efficient version of REGAL. Zhang et al. [31] propose a model-based algorithm called EBF that achieves the minimax optimal regret of $\mathcal{O}(\sqrt{\text{sp}(v^*)SAT})$ by maintaining a tighter model confidence set by making use of the estimate for the optimal bias function. However, their algorithm is computationally inefficient. There is another line of work on model-free algorithms for this setting. Wei et al. [29] introduce a model-free Q-learning-based algorithm called Optimistic Q-learning. Their algorithm is a reduction to the discounted setting. Although model-free, their algorithm has a suboptimal regret of $\mathcal{O}(T^{2/3})$. Recently, Zhang et al. [32] introduce a Q-learning-based algorithm called UCB-AVG that achieves regret bound of $\mathcal{O}(\sqrt{T})$. Their algorithm, which is also a reduction to the discounted setting, is the first model-free to achieve the order optimal regret bound. Their main idea is to use the optimal bias function estimate to increase statistical efficiency. Agrawal et al. [3] introduces a model-free Q-learning-based algorithm and provides a unified view of episodic setting and infinite-horizon average-reward setting. However, their algorithm requires additional assumption of the existence of a state with bounded hitting time.

Infinite-Horizon Average-Reward Setting with General Function Approximation He et al. [15] study infinite-horizon average reward with general function approximation. They propose an algorithm called LOOP which is a modified version of the fitted Q-iteration with optimistic planning and lazy policy updates. Although their algorithm when adapted to the linear MDP set up achieves $\mathcal{O}(\sqrt{\mathrm{sp}(v^*)^3 d^3 T})$, which is comparable to our work, their algorithm is computationally inefficient.

Infinite-Horizon Average-Reward Setting with Linear MDPs There is another work by Ghosh et al. [11] on the infinite-horizon average-reward setting with linear MDPs. They study a more general constrained MDP setting where the goal is to maximize average reward while minimizing the average cost. They achieve $\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)\sqrt{d^3 T})$ regret, same as our work, but they make an additional assumption that the optimal policy is in a smooth softmax policy class. Also, their algorithm requires solving an intractable optimization problem.

Reduction of Average-Reward to Finite-Horizon Episodic Setting There are works that reduce the average-reward setting to the finite-horizon episodic setting. However, in general, this reduction can only give regret bound of $\mathcal{O}(T^{2/3})$. Chen et al. [8] study the constrained tabular MDP setting and propose an algorithm that uses the finite-horizon reduction. Their algorithm gives regret bound of $\mathcal{O}(T^{2/3})$. Wei et al. [28] study the linear MDP setting and propose a finite-horizon reduction that uses the LSVI-UCB [18]. Their reduction gives regret bound of $\mathcal{O}(T^{2/3})$.

Online RL in Infinite-Horizon Discounted Setting The literature on online RL in the infinite-horizon discounted setting is sparse because there is no natural notion of regret in this setting without additional assumption on the structure of the MDP. The seminal paper by Liu et al. [20] introduce a notion of regret in the discounted setting and propose a Q-learning-based algorithm for the tabular setting and provides a regret bound. He et al. [14] propose a model-based algorithm that adapts UCBVI [6] to the discounted setting and achieve a nearly minimax optimal regret bound. Ji et al. [17] propose a model-free algorithm with nearly minimax optimal regret bound.