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

Kihyuk Hong
University of Michigan

Woojin Chae
KAIST

Yufan Zhang
University of Michigan

Dabeen Lee
KAIST

Ambuj Tewari
University of Michigan

Abstract

We study the problem of infinite-horizon average-reward reinforcement learning with linear Markov decision processes (MDPs). The associated Bellman operator of the problem not being a contraction makes the algorithm design challenging. 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 the first algorithm that achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret with computational complexity polynomial in the problem parameters, without making strong assumptions on dynamics. Our approach approximates the average-reward setting by a discounted MDP with a carefully chosen discounting factor, and then applies an optimistic value iteration. We propose an algorithmic structure that plans for a nonstationary policy through optimistic value iteration and follows that policy until a specified information metric in the collected data doubles. Additionally, we introduce a value function clipping procedure for limiting the span of the value function for sample efficiency.

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 does not terminate and continues indefinitely, such as in network routing (Mammeri 2019) or inventory management (Giannoccaro et al. 2002). Designing algorithms in this setting is challenging because the associated Bellman operator is not a contraction. This complicates the use of optimistic value iteration-based algorithms, which add bonus terms to the value function every iteration, a widely used approach in the finite-horizon episodic (Jin et al. 2020) and infinite-horizon discounted settings (He et al. 2021).

The seminal work of (Jaksch et al. 2010) adopts a model-based approach in the tabular setting that constructs a confidence set on the transition model. Their algorithm runs optimistic value iterations by choosing an optimistic model from the confidence set at each iteration. Most work in the tabular setting follows this approach of constructing confidence sets for the model (Bartlett et al. 2009; Fruit et al. 2018). Other work (Wei et al. 2020) that does not use this approach either has suboptimal regret bound or assumes the ergodicity assumption.

Adapting the approach of constructing confidence set on the transition model to function approximation settings, such as linear MDP setting, with arbitrarily large state space is challenging because sample-efficient model estimation is elusive in general unless additional assumptions on the transition model are made. Previous work on infinite-horizon average-reward RL with function approximation either imposes ergodicity assumptions (Wei et al. 2021) or uses computationally inefficient algorithms (Wei et al. 2021; He et al. 2024) with time complexity exponential in problem parameters to achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. The following question remains open:

> *Does there exist a polynomial-time algorithm for infinite-horizon average-reward lin-*

ear MDPs that achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret without requiring the ergodicity assumption?

In this paper, we answer the question in the affirmative. We draw insights from a recent line of work (Wei et al. 2020; Zhang et al. 2023) that employs the technique of approximating infinite-horizon average-reward RL by a discounted MDP in the tabular setting. We apply the discounted setting approximation idea to the linear MDP setting and design an optimistic value iteration algorithm that achieves $\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)\sqrt{d^3 T})$ regret without making the ergodicity assumption. A key component of our method is the use of value function clipping, coupled with a novel value iteration scheme to ensure efficient learning in this setting.

The rest of the paper is organized as follows. In Section 2, we formally define infinite-horizon average-reward setting and discounted setting. In Section 3, we introduce a simple value iteration based algorithm for the tabular setting that approximates the average-reward setting by the discounted setting. In Section 4, we adapt the algorithm design to the linear MDP setting, and addresses a unique challenge in this setting by proposing a novel algorithm structure.

1.1 Related Work

A comparison of our work to previous work on infinite-horizon average-reward linear MDPs is shown in Table 1. Entries highlighted in red indicate suboptimality compared to our algorithm. Our algorithm is the first to achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret with computation complexity polynomial in the parameters d, S, A, T without making the ergodicity assumption. Note that the uniform mixing assumption required for MDP-EXP2 (Wei et al. 2021) is a stronger assumption than the ergodicity assumption. Our regret matches that of FOPO (Wei et al. 2021), 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. In contrast, our algorithm's complexity is polynomial in d. However, it depends polynomially on the size of the state space S, whereas the complexity of FOPO does not depend on S. This implies our algorithm is an improvement over FOPO in the regime where $S \ll T^d$. We leave the problem of getting rid of the dependence on S for future work. Additional comparisons in the tabular setting, as well as a broader discussion of related work, can be found in Appendix E.

Approximation of Average-Reward Setting by Discounted Setting The technique of approximating the average-reward setting to the discounted setting is used by Jin et al. (2021), Wang et al. (2022), Zurek et al. (2023), and Wang et al. (2023) to solve the sample complexity problem of producing a nearly optimal policy given access to a simulator in the tabular setting. Wei et al. (2020) use the reduction for the online RL setting with tabular MDPs. They propose a Q-learning based algorithm, but has $\widetilde{\mathcal{O}}(T^{2/3})$ regret. Zhang et al. (2023) also use the reduction for the online RL setting with tabular MDPs. Their algorithm is also Q-learning based, but they improve the regret to $\widetilde{\mathcal{O}}(\sqrt{T})$ by introducing a novel method for estimating the span.

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. (2022) and Ayoub et al. (2020) design an optimism-based algorithm using a confidence set under the assumption that MDP is communicating. Chae et al. (2024) use the method of reducing the average-reward setting to the discounted setting, and achieve a nearly minimax optimal regret bound under a weaker assumption on the MDP.

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 $\mathrm{sp}(v) = \max_{s,s'} |v(s) - v(s')|$. Let $[Pv](s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)}[v(s')]$.

2.1 Infinite-Horizon Average-Reward MDPs

We consider a Markov decision process (MDP) (Puterman 2014), $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$ 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

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

Algorithm	Regret $\widetilde{\mathcal{O}}(\cdot)$	Assumption	Computation poly($\cdot$)
FOPO (Wei et al. 2021)	$\mathrm{sp}(v^*)\sqrt{d^3 T}$	Bellman optimality equation	T^d, A, d
OLSVI.FH (Wei et al. 2021)	$\sqrt{\mathrm{sp}(v^*)}(dT)^{\frac{3}{4}}$	Bellman optimality equation	T, A, d
LOOP (He et al. 2024)	$\sqrt{\mathrm{sp}(v^*)^3 d^3 T}$	Bellman optimality equation	T^d, A, d
MDP-EXP2 (Wei et al. 2021)	$d\sqrt{t_{\mathrm{mix}}^3 T}$	Uniform mixing	T, A, d
γ-**LSCVI-UCB (Ours)**	$\mathrm{sp}(v^*)\sqrt{d^3 T}$	Bellman optimality equation	T, S, A, d
Lower Bound (Wu et al. 2022)	$\Omega(d\sqrt{\mathrm{sp}(v^*)T})$		

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

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

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 $J^*(s_1) := J^{\pi^*}(s_1)$, the regret is

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

As discussed by Bartlett et al. (2009), 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. (2021) 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*

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

As discussed by Wei et al. (2021), the Bellman optimality equation assumption is a weaker assumption than the weakly communicating assumption, which in turn is weaker than ergodicity. Weakly communicating

assumption is another widely used assumption for the infinite-horizon average-reward setting that requires each pair of states in the set to be reachable from each other under some policy. Ergodicity assumption requires the Markov chain induced by any policy to be ergodic, i.e., irreducible and aperiodic. Uniform mixing assumption required for MDP-EXP2 proposed by Wei et al. (2021) is a stronger assumption than ergodicity that additionally assumes that the mixing time is uniformly bounded over all policies.

As shown by Wei et al. (2021), under Assumption A, the policy π^* that deterministically selects an action from $\operatorname{argmax}_a q^*(s, a)$ 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 (Bartlett et al. 2009; Wei et al. 2020), we assume $\mathrm{sp}(v^*)$ is known to the learner.

Remark 1. Instead of assuming exact knowledge of $\mathrm{sp}(v^*)$, we can assume that the learner knows an upper bound H. In this case, using H as an input to our algorithm instead of $\mathrm{sp}(v^*)$, the regret bound of our algorithm will scale with H rather than $\mathrm{sp}(v^*)$. Such a knowledge of an upper bound is a commonly made assumption (Bartlett et al. 2009). In practice, if one has a general sense of the diameter of the MDP, which is the expected number of steps needed to transition between any two states in the worst case, the diameter can serve as an upper bound H, since the diameter is guaranteed to be an upper bound of $\mathrm{sp}(v^*)$ when reward is bounded by 1. Relaxing the assumption of the knowledge of $\mathrm{sp}(v^*)$ or its upper bound is only achieved recently in the tabular setting (Boone et al. 2024). Extending this relaxation to the linear MDP setting remains an open question for 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* $t = 1, \ldots, T$ **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      $Q_{t+1}(s,a) \leftarrow (r(s,a) + \gamma[\widehat{P}_t V_t](s,a) + \beta/\sqrt{N_t(s,a)}) \wedge Q_t(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$
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     $V_{t+1}(s) \leftarrow \widetilde{V}_{t+1}(s) \wedge (\min_{s'} \widetilde{V}_{t+1}(s') + H), \quad \forall s \in \mathcal{S}$.

2.2 Infinite-Horizon Discounted Setting

The key idea of this paper, inspired by Zhang et al. (2023), 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 (Puterman 2014) 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. 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. (2020)). *For any $\gamma \in [0,1)$, the optimal value function V^* for the infinite-horizon discounted setting with discounting factor γ satisfies*

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

(ii) $|(1-\gamma)V^*(s) - J^*| \le (1-\gamma)sp(v^*)$ 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 (Azar et al. 2017), 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 $Q_t \ge Q^*$ 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, where the target span H is an input to the algorithm. Without clipping, the span of the value function V_t can be as large as $\frac{1}{1-\gamma}$, while with clipping, the span can only be as large as

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. Running the algorithm with the discounting factor set to $\gamma = 1 - 1/\sqrt{T}$ and the target span set to $H = 2 \cdot \mathrm{sp}(v^*)$ guarantees 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*

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

In the theorem above, the constant c in the definition of β is determined in Lemma 3 in the next subsection. 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 work 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. (2017). We leave the improvement to future work.

3.2 Analysis

In this section, we outline the proof of Theorem 2. We defer the complete proof to Appendix A. The key to the proof is the following concentration inequality.

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*

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

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

Without clipping, the span of V_t would be $\frac{1}{1-\gamma}$ instead of $2 \cdot \mathrm{sp}(v^*)$, making the bound of $[(\widehat{P}_t - P)V_t](s,a)$ scale with $\frac{1}{1-\gamma}$ instead of $\mathrm{sp}(v^*)$. Replacing the $\frac{1}{1-\gamma}$ factor by $\mathrm{sp}(v^*)$ by clipping is crucial to achieving $\widetilde{\mathcal{O}}(\sqrt{T})$ regret when tuning γ.

In Theorem 2, the bonus factor parameter β is chosen according to the concentration bound in the lemma above, ensuring that the concentration bound for $[(\widehat{P}_t - P)V_t](s,a)$ is $\beta/\sqrt{N_t(s,a)}$, which is the bonus term used by the algorithm. With this result, we can now establish the following optimism result.

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

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

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

The proof uses a standard induction argument (e.g. Lemma 18 in Azar et al. (2017)) to show $\widetilde{V}_t(s) \geq V^*(s)$. To establish that the clipped value function V_t, no larger than $\widetilde{V}_t$ by design, still satisfies $V_t(s) \geq V^*(s)$, we use $\mathrm{sp}(V^*) \leq 2 \cdot \mathrm{sp}(v^*)$ (Lemma 1), which guarantees the clipping operation does not reduce V_t below V^*.

Now, we show the regret bound under 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 $t = 2, \ldots, T$ that

$$r(s_t, a_t)$$
$$\geq Q_t(s_t, a_t) - \gamma[\widehat{P}_{t-1} V_{t-1}](s_t, a_t) - \beta/\sqrt{N_{t-1}(s_t, a_t)}$$
$$\geq V_t(s_t) - \gamma[PV_{t-1}](s_t, a_t) - 2\beta/\sqrt{N_{t-1}(s_t, a_t)}$$

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

$$R_T = \sum_{t=1}^{T}(J^* - r(s_t, a_t))$$
$$\leq \sum_{t=2}^{T}(J^* - V_t(s_t) + \gamma[PV_{t-1}](s_t, a_t)$$
$$+ 2\beta/\sqrt{N_{t-1}(s_t, a_t)}) + \mathcal{O}(1),$$

where the first inequality uses the fact that $J^* \leq 1$, which can be decomposed into

$$= \underbrace{\sum_{t=2}^{T}(J^* - (1-\gamma)V_t(s_t))}_{(a)}$$
$$+ \gamma\underbrace{\sum_{t=2}^{T}(V_{t-1}(s_{t+1}) - V_t(s_t))}_{(b)}$$
$$+ \gamma\underbrace{\sum_{t=2}^{T}(PV_{t-1}(s_t, a_t) - V_{t-1}(s_{t+1}))}_{(c)}$$
$$+ 2\beta\underbrace{\sum_{t=2}^{T}1/\sqrt{N_{t-1}(s_t, a_t)}}_{(d)} + \mathcal{O}(1).$$

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

Bounding Term (a) Using the optimism result (Lemma 4) that says $V_t(s) \geq V^*(s)$ for all $s \in \mathcal{S}$ and Lemma 1 that bounds $|J^* - (1-\gamma)V^*(s)|$ for all $s \in \mathcal{S}$,

we get

$$\sum_{t=2}^{T}(J^* - (1-\gamma)V_t(s_t)) \leq \sum_{t=2}^{T}(J^* - (1-\gamma)V^*(s_t))$$
$$\leq T(1-\gamma)\text{sp}(v^*).$$

Bounding Term (b) Note that for any $s \in \mathcal{S}$, the sequence $\{V_t(s)\}_{t=1}^{T}$ is monotonically decreasing due to Line 9-10 in Algorithm 1. Moreover, since $V_t(s) \in [0, \frac{1}{1-\gamma}]$ for all $t = 1, \ldots, T$, the total decrease in $V_t(s)$ from $t = 1$ to T is bounded above by $\frac{1}{1-\gamma}$. Hence,

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

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 next 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 $\text{sp}(v^*)$. Hence, by the Azuma-Hoeffding inequality, with probability at least $1 - \delta$, term (c) is bounded by $\text{sp}(v^*)\sqrt{2T\log(1/\delta)}$. 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) using a standard argument (Azar et al. (2017), Lemma 10 in Appendix A) by $\mathcal{O}(\beta\sqrt{SAT}) = \mathcal{O}(\text{sp}(v^*)\sqrt{S^2AT\log(SAT/\delta)})$.

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

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

Choosing $\gamma = 1 - 1/\sqrt{T}$, we get

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

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, which we formally define below. The tabular setting studied in the previous section has regret bound that scales polynomially with the size of the state space S. This is because, in the tabular setting, there is no structure in the state space that can be exploited to generalize to unseen states during learning. Therefore, when learning in a large state space with $S \gg T$, additional structural assumption are necessary. The linear MDP setting (Jin et al. 2020) is a widely used setting in the RL theory literature that allows for generalization to unseen states by introducing structure in the MDP through a low-dimensional state-action feature mapping. The additional assumption made in the linear MDP setting is as follows.

Assumption B (Linear MDP (Jin et al. 2020)). *We assume that the transition and the reward functions can be expressed as a linear function of a known d-dimensional feature map $\varphi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$ such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have*

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

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

As is commonly done in the literature on linear MDPs (Jin et al. 2020), we further assume, without loss of generality (see Wei et al. (2021) for justification), the following boundedness conditions:

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

As discussed by Jin et al. (2020), although the transition model P is linear in the d-dimensional feature mapping φ, P still has infinite degrees of freedom as the measure μ is unknown, making the estimation of the model P difficult. For sample efficient learning, we leverage the fact that $[Pv](s, a)$ is linear in $\varphi(s, a)$ for any function $v : \mathcal{S} \rightarrow \mathbb{R}$ so that $[Pv](s, a) = \langle\varphi(s, a), w_v^*\rangle$ for some w_v^*, since

$$[Pv](s, a) \coloneqq \int_{s' \in \mathcal{S}} v(s')P(ds'|s, a)$$
$$= \int_{s' \in \mathcal{S}} v(s')\langle\varphi(s, a), \mu(ds')\rangle$$
$$= \langle\varphi(s, a), \int_{s' \in \mathcal{S}} v(s')\mu(ds')\rangle.$$

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

Challenges Naively adapting the algorithm design and analysis for the tabular setting to the linear MDP setting would result in a regret bound that is 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, would lead to an exponential covering number for the function class of the value function, in either T or S (He et al. 2023). A major challenge in algorithm design and analysis is sidestepping these issues. We now present our algorithm for the linear MDP setting, which addresses these issues.

4.1 Algorithm

Our algorithm, called *discounted least-squares clipped value iteration with upper confidence bound* (γ-LSCVI-UCB), adapts LSVI-UCB (Jin et al. 2020) developed for the episodic setting to the discounted setting. We highlight key differences from LSVI-UCB below.

Clipping the Value Function We clip the value function estimates, as is done in the tabular setting in the previous section, to restrict the span (Line 12), which saves a factor of $1/(1-\gamma)$ in the regret bound.

Restricting the Range of Value Target When regressing $V_u^k(s')$ on $\varphi(s, a)$ 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 the target span H 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 $\boldsymbol{w}_u^k : \langle \varphi(\cdot, \cdot), \boldsymbol{w}_u^k \rangle + \min_{s'} \widetilde{V}_u^k(s')$ (Line 10).

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 (Zhang et al. 2023) and infinite-horizon discounted tabular MDPs (Liu et al. 2020; He et al. 2021). 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, forcing V_t to be 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.

Planning until the End of Horizon 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. This algorithm structure is reminiscent of the

value iteration based algorithms for the finite-horizon episodic setting (Azar et al. 2017; Jin et al. 2020), where Bellman operations are performed to generate action value functions at each time step in an episode of fixed length, then greedy actions with respect to those action value functions are taken for the entire episode. With the new algorithm structure, Q_{t-1} is now one Bellman operation ahead of Q_t, and the quantity of interest becomes $\sum_{t=1}^{T} V_{t+1}(s_{t+1}) - V_t(s_t)$, which can be bounded by telescoping sum.

Restarting when Information Doubles If we generate all T action value functions to be used at the initial time step by running approximate value iteration, and follow 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 restart the process of running value iterations for T steps every time a certain information measure of the collected data doubles. This allows us to incorporate the newly collected trajectory data into subsequent decision-making. We adopt the rarely-switching covariance matrix trick (Wang et al. 2021), which triggers a restart when the determinant of the empirical covariance matrix doubles (Line 5).

Our 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 (He et al. 2023) may improve our regret by a factor of $\sqrt{d}$. We leave this improvement to future work.

4.2 Computational Complexity

Our algorithm runs in episodes and since a new episode starts only when the determinant of the covariance matrix $\bar{\Lambda}_t$ doubles, there can be at most $\mathcal{O}(d \log_2 T)$ episodes (see Lemma 21). 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. Also, the algorithm computes $\boldsymbol{w}_{u+1}^k$, which requires $\mathcal{O}(d^2 + Td)$ operations. All other operations runs in $\mathcal{O}(d^2 + A)$ per value iteration. In total, the algorithm runs in $\mathcal{O}((\log_2 T)d^3 SAT^2)$. See Appendix C for detailed analysis.

The FOPO algorithm by Wei et al. (2021) that matches our regret bound under the same set of assumptions,

has a time complexity of $\mathcal{O}(T^d \log_2 T)$. Although our time complexity is an improvement over previous work in the sense that the time complexity is polynomial in problem parameters, it has linear dependency on S. The dependency on S arises from taking the minimum of value functions for clipping. We conjecture that this dependency can be eliminated by using an estimate of the minimum rather than computing the global minimum of value functions. For example, replacing $\min_{s'} V_u^k(s')$ with $\min_{s'} V^*(s')$ for clipping leads to the same regret bound (see Appendix B.3). A promising approach is to use $\min_{s' \in \{s_1, \ldots, s_t\}} \widetilde{V}_u^k(s')$, minimum over states visited so far, instead of the global minimum. However, as discussed in Appendix B.4, naively changing the clipping operation fails. We leave eliminating the dependency on S in the time complexity to future work.

4.3 Analysis

In this section, we outline the proof of the regret bound presented 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 indices k and for all vectors $\boldsymbol{\phi} \in \mathbb{R}^d$ where $\boldsymbol{w}_u^{k*} := \int (V_u^k(s) - \min_{s'} V_u^k(s'))d\boldsymbol{\mu}(s)$ 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

$$r(s,a) \geq Q_t^k(s,a) - \gamma(\langle \varphi(s,a), \boldsymbol{w}_{t+1}^k \rangle + \min_{s'} V_{t+1}^k(s')$$
$$- \beta \|\varphi(s,a)\|_{\Lambda_k^{-1}})$$
$$\geq Q_t^k(s,a) - \gamma[PV_{t+1}^k](s,a) - 4\beta \|\varphi(s,a)\|_{\bar{\Lambda}_t^{-1}}$$

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 20). Hence, we can bound the regret in episode k by

$$R^k = \sum_{t=t_k}^{t_{k+1}-1} (J^* - r(s_t, a_t))$$
$$\leq \sum_{t=t_k}^{t_{k+1}-1} (J^* - Q_t^k(s_t, a_t) + \gamma[PV_{t+1}^k](s_t, a_t)$$
$$+ 4\beta \|\varphi(s_t, a_t)\|_{\bar{\Lambda}_t^{-1}}),$$

which can be decomposed into

$$= \underbrace{\sum_{t=t_k}^{t_{k+1}-1} (J^* - (1-\gamma)V_{t+1}^k(s_{t+1}))}_{(a)}$$
$$+ \gamma \underbrace{\sum_{t=t_k}^{t_{k+1}-1} (V_{t+1}^k(s_{t+1}) - Q_t^k(s_t, a_t))}_{(b)}$$
$$+ \gamma \underbrace{\sum_{t=t_k}^{t_{k+1}-1} [PV_{t+1}^k](s_t, a_t) - V_{t+1}^k(s_{t+1}))}_{(c)}$$
$$+ 4\beta \underbrace{\sum_{t=t_k}^{t_{k+1}-1} \|\varphi(s_t, a_t)\|_{\bar{\Lambda}_t^{-1}}}_{(d)}$$

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 $V_u^k(s_{t+1}) \geq V^*(s_{t+1})$, and Lemma 1 that bounds $|J^* - (1-\gamma)V^*(s)|$ for all $s \in \mathcal{S}$. Term (d), summed over all episodes, can be bounded by $\mathcal{O}(\beta\sqrt{dT\log T})$ using Cauchy-Schwartz and Lemma 19. Term (c), summed over all episodes, is a sum of a martingale difference sequence, which can be bounded by $\mathcal{O}(\mathrm{sp}(v^*)\sqrt{T\log(1/\delta)})$ since $\mathrm{sp}(V_u^k) \leq 2 \cdot \mathrm{sp}(v^*)$ by the clipping step in Line 12.

Bounding Term (b) To bound term (b) note that

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

as long as the time step $t+1$ is in episode k, since the algorithm chooses a_{t+1} that maximizes $Q_{t+1}^k(s_{t+1}, \cdot)$. Hence,

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

where the second inequality uses telescoping sum and the fact that $Q_t^k \leq \frac{1}{1-\gamma}$. 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 21). Combining all the bounds, and using $\beta = \mathcal{O}(\mathrm{sp}(v^*)d\sqrt{\log(dT/\delta)})$, we get

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

Setting $\gamma = 1 - \sqrt{(\log T)/T}$, we get

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

which concludes the proof of Theorem 5.

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. Eliminating the dependence on the size of the state space in time complexity remains an open problem. Another promising direction for future work would be to extend these methods to the general function approximation setting.

References

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

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

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

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

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

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

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

Boone, Victor and Zihan Zhang (2024). "Achieving tractable minimax optimal regret in average reward mdps". In: *Advances in Neural Information Processing Systems*.

Chae, Woojin, Kihyuk Hong, Yufan Zhang, Ambuj Tewari, and Dabeen Lee (2024). "Learning Infinite-Horizon Average-Reward Linear Mixture MDPs of Bounded Span". In: *arXiv preprint arXiv:2410.14992*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes]

   (b) Complete proofs of all theoretical results. [Yes]

   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

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 (Abbasi-Yadkori et al. 2012)). *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*

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

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.

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

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

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

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

$\square$

We use $\mathcal{N}_\epsilon$ to denote the ϵ-covering number of $\mathcal{V}_{\text{tabular}}$ with respect to the distance $\text{dist}(V, V') = \|V - V'\|_\infty$. Using a grid of size ϵ, since functions in $\mathcal{V}_{\text{tabular}}$ has the range $[0, \tfrac{1}{1-\gamma}]$, it can be seen that $\log \mathcal{N}_\epsilon \leq S \log \frac{1}{\epsilon(1-\gamma)}$. Now, we prove a uniform concentration bound using a covering argument on $\mathcal{V}_{\text{tabular}}$.

Proof of Lemma 3. Note that $V_t \in \mathcal{V}_{\text{tabular}}$ for all t. Consider an ϵ-cover of $\mathcal{V}_{\text{tabular}}$. For any $V_t \in \mathcal{V}_{\text{tabular}}$, there exists $\widetilde{V}_t$ in the ϵ-cover such that $\sup_s |V_t(s) - \widetilde{V}_t(s)| \leq \epsilon$. Thus, we have

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

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

$$|[(\widehat{P}_t - P)V_t](s,a)| \leq C \cdot \text{sp}(v^*)\sqrt{\frac{\log(SAT\mathcal{N}_\epsilon/\delta)}{N_t(s,a)}} + 2\epsilon$$

$$\leq C \cdot \text{sp}(v^*)\sqrt{\frac{\log(SAT/\delta) + S\log(1/(\epsilon(1-\gamma)))}{N_t(s,a)}} + 2\epsilon.$$

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

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

$$Q_{t+1}(s,a) = (r(s,a) + \gamma[\widehat{P}_t V_t](s,a) + \beta/\sqrt{N_t(s,a)}) \wedge Q_t(s,a)$$
$$\geq (r(s,a) + \gamma[PV_t](s,a)) \wedge Q_t(s,a)$$
$$\geq (r(s,a) + \gamma[PV^*](s,a)) \wedge Q^*(s,a)$$
$$= Q^*(s,a)$$

where the first inequality is by the concentration inequality in Lemma 3 and our choice of β in Theorem 2, and the second inequality is by the induction hypotheses $V_t \geq V^*$ and $Q_t \geq Q^*$. The last equality is by the Bellman optimality equation.

Now, we show $V_{t+1}(s) \geq V^*(s)$. By the definition of $\widetilde{V}_{t+1}$ in Line 9 of Algorithm 1, we have

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

where the inequality is by the optimism of Q_{t+1} we just proved, and the induction hypothesis $V_t \geq V^*$.

Finally, by the definition of V_{t+1} in Line 10 of Algorithm 1, we have

$$V_{t+1}(s) = \widetilde{V}_{t+1}(s) \wedge (\min_{s'} \widetilde{V}_{t+1}(s') + \text{sp}(v^*)) \geq \widetilde{V}_{t+1}(s) \geq V^*(s),$$

which completes the proof by induction. $\qquad\square$

A.3 Omitted Proofs

Lemma 10 (Sum of Bonus Terms). *Consider running Algorithm 1. The sum of the bonus terms can be bounded by*

$$\sum_{t=1}^{T} \sqrt{1/N_{t-1}(s_t, a_t)} \leq 2\sqrt{SAT}.$$

Proof. Recall that $N_t(s,a) = 1 + \sum_{s=1}^{t} \mathbb{I}\{s_t = s, a_t = a\}$. For convenience, write $n_t(s,a) = \sum_{s=1}^{t} \mathbb{I}\{s_t = s, a_t = a\}$ such that $N_t(s,a) = 1 + n_t(s,a)$. Then,

$$
\begin{aligned}
\sum_{t=1}^{T} \sqrt{1/N_{t-1}(s_t, a_t)} &\leq \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{n=1}^{n_T(s,a)} \sqrt{1/n} \\
&\leq 2 \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sqrt{n_T(s,a)} \\
&\leq 2\sqrt{SAT}
\end{aligned}
$$

where the second inequality uses the identity $\sum_{n=1}^{N} 1/\sqrt{n} \leq 2\sqrt{N}$, and the last inequality is by Cauchy-Schwarz and the fact that $\sum_s \sum_a n_T(s,a) = T$. $\qquad\square$

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 11 (Concentration of vector-valued self-normalized processes (Abbasi-Yadkori et al. 2011)). *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 12. *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

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

Hence, $\boldsymbol{w}_V^* = \int_{\mathcal{S}} V(s') d\boldsymbol{\mu}(s')$ satisfies $[PV](s,a) = \langle \varphi(s,a), \boldsymbol{w}_V \rangle$ 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 $\boldsymbol{\mu}(\mathcal{S})$. $\qquad\square$

Lemma 13. *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 18.

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

The following self-normalized process bound is an adaptation of Lemma D.4 in Jin et al. (2020). 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 $sp(V) \leq B$. The following lemma is the strengthened version.

Lemma 14 (Adaptation of Lemma D.4 in Jin et al. (2020))**.** *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 15 (Adaptation of Lemma D.6 in Jin et al. (2020))**.** *Let $\mathcal{V}_{linear}$ be a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following parametric form*

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

where the parameters $(\boldsymbol{w}, \beta, v, \Lambda, M)$ satisfy $\|\boldsymbol{w}\| \leq L$, $\beta \in [0, B]$, $v \in [0, D]$, $M \geq 0$ and the minimum eigenvalue satisfies $\lambda_{min}(\Lambda) \geq \lambda$. Assume $\|\boldsymbol{\varphi}(s, a)\| \leq 1$ 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

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

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):

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.

$V_{T+1}^{(t)}(\cdot) \leftarrow \frac{1}{1-\gamma}, V_{T+1}^{(t)}(\cdot) \leftarrow \frac{1}{1-\gamma}$.

for $u = T, T-1, \ldots, 1$ **do**

$\quad \boldsymbol{w}_{u+1}^{(t)} \leftarrow \bar{\Lambda}_t^{-1} \sum_{\tau=1}^t \boldsymbol{\varphi}(s_\tau, a_\tau)(V_{u+1}^{(t)}(s_{\tau+1}) - \min_{s'} \widetilde{V}_{u+1}^{(t)}(s'))$.

$\quad \widetilde{Q}_u^{(t)}(\cdot, \cdot) \leftarrow \left(r(\cdot, \cdot) + \gamma(\langle \boldsymbol{\varphi}(\cdot, \cdot), \boldsymbol{w}_{u+1}^{(t)} \rangle + \min_{s'} \widetilde{V}_{u+1}^{(t)}(s') + \beta \|\boldsymbol{\varphi}(\cdot, \cdot)\|_{\bar{\Lambda}_t^{-1}}) \right) \wedge \frac{1}{1-\gamma}$.

$\quad \widetilde{V}_u^{(t)}(\cdot) \leftarrow \max_a \widetilde{Q}_u^{(t)}(\cdot, a)$.

$\quad V_u^{(t)}(\cdot) \leftarrow \widetilde{V}_u^{(t)}(\cdot) \wedge (\min_{s'} \widetilde{V}_u^{(t)}(s') + H)$.

Lemma 16 (Adaptation of Lemma B.3 in Jin et al. (2020)). *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*

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

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

Proof. For all $t = 1, \ldots, T$, by Lemma 13, we have $\|\boldsymbol{w}_t\|_2 \leq H\sqrt{dt/\lambda}$. Hence, by combining Lemma 15 and Lemma 14, 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

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

where we use the fact that $\tau_t \leq t$. 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 $1 - \delta$,

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

which concludes the proof. $\qquad \square$

Proof of Lemma 6. We prove under the event $\mathcal{E}$ defined in Lemma 16. 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 12 that $\|\boldsymbol{w}_u^{k*}\|_2 \leq H\sqrt{d}$. Hence,

the first term (a) in the display above can be bounded as

$$\langle \phi, \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau)[P\bar{V}_u^k](s_\tau, a_\tau)\rangle = \langle \phi, \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau)\varphi(s_\tau, a_\tau)^T \boldsymbol{w}_u^{k*}\rangle$$

$$= \langle \phi, \boldsymbol{w}_u^{k*}\rangle - \lambda\langle \phi, \Lambda_k^{-1}\boldsymbol{w}_u^{k*}\rangle$$

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

$$\leq \langle \phi, \boldsymbol{w}_u^{k*}\rangle + H\sqrt{\lambda d}\|\phi\|_{\Lambda_k^{-1}}$$

where the first inequality is by Cauchy-Schwartz and the second inequality is by Lemma 12. Under the event $\mathcal{E}$ defined in Lemma 16, the second term (b) can be bounded by

$$\langle \phi, \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau)(\bar{V}_u^k(s_{\tau+1}) - [P\bar{V}_u^k](s_\tau, a_\tau))$$

$$\leq \|\phi\|_{\Lambda_k^{-1}}\left\|\sum_{\tau=1}^{t_k-1} \varphi(s_\tau, a_\tau)(V_u^k(s_{\tau+1}) - [PV_u^k](s_\tau, a_\tau))\right\|_{\Lambda_k^{-1}}$$

$$\leq C \cdot Hd\sqrt{\log((c_\beta+1)dT/\delta)} \cdot \|\phi\|_{\Lambda_k^{-1}}.$$

Combining the two bounds and rearranging, we get

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

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

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

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 16. 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.

$$Q_u^k(s, a) = \left(r(s, a) + \gamma(\langle \varphi(s, a), \boldsymbol{w}_{u+1}^k\rangle + \min_{s'} V_{u+1}^k(s') + \beta\|\varphi(s, a)\|_{\Lambda_k^{-1}}\right) \wedge \frac{1}{1-\gamma}$$

$$\geq \left(r(s, a) + \gamma(\langle \varphi(s, a), \boldsymbol{w}_{u+1}^{k*}\rangle + \min_{s'} V_{u+1}^k(s'))\right) \wedge \frac{1}{1-\gamma}$$

$$= (r(s, a) + \gamma PV_{u+1}^k(s, a)) \wedge \frac{1}{1-\gamma}$$

$$\geq (r(s, a) + \gamma PV^*(s, a)) \wedge \frac{1}{1-\gamma}$$

$$= Q^*(s, a)$$

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

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

where the last equality is due to $\operatorname{sp}(V^*) \leq 2 \cdot \operatorname{sp}(v^*)$ 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. $\square$

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 17 (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 $H = 2 \cdot \operatorname{sp}(v^*)$ 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 16. 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.

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

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

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

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

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

where the last equality is due to $\operatorname{sp}(V^*) \leq 2 \cdot \operatorname{sp}(v^*)$ 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.4 Difficulty of Bounding Regret with Computationally Efficient Clipping

For computational efficiency, consider using $\min_{s' \in \{s_1, \ldots, s_{t_k}\}} \widetilde{V}_u^k(s')$ for clipping instead of $\min_{s' \in \mathcal{S}} \widetilde{V}_u^k(s')$ when running value iteration in the beginning of episode k for generating value functions for episode k. Note that in the beginning of episode k, we only have only seen states $s_1, \ldots, s_{t_k}$. A natural clipping operation for enforcing the span to be bounded by H is

$$V_u^k(\cdot) = (\widetilde{V}_u^k(\cdot) \wedge (\min_{s' \in \{s_1, \ldots, s_{t_k}\}} \widetilde{V}_u^k(s') + H)) \vee \min_{s' \in \{s_1, \ldots, s_{t_k}\}} \widetilde{V}_u^k(s'),$$

which is equivalent to clipping the value $\widetilde{V}_u^k(\cdot)$ to the interval $[\min_{s' \in \{s_1, \ldots, s_{t_k}\}} \widetilde{V}_u^k(s'), \min_{s' \in \{s_1, \ldots, s_{t_k}\}} \widetilde{V}_u^k(s') + H]$. One difficulty in the analysis that arises from this change in clipping is when bounding the term $\sum_{t=t_k}^{t_{k+1}-1}(V_{t+1}^k(s_{t+1}) - Q_t^k(s_t, a_t))$. To illustrate the difficulty, recall one of the steps in the proof presented in this paper:

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

The inequality $V_{t+1}^k(s_{t+1}) \leq \widetilde{V}_{t+1}^k(s_{t+1})$ may no longer hold when using the new computationally efficient version of the minimum. Instead, we get a bound that looks like

$$V_{t+1}^k(s_{t+1}) = (\widetilde{V}_{t+1}^k(s_{t+1}) \wedge (\min_{s' \in \{s_1, \ldots, s_{t_k}\}} \widetilde{V}_{t+1}^k(s') + H)) \vee \min_{s' \in \{s_1, \ldots, s_{t_k}\}} \widetilde{V}_{t+1}^k(s')$$
$$\leq \widetilde{V}_{t+1}^k(s_{t+1}) + (\min_{s' \in \{s_1, \ldots, s_{t_k}\}} \widetilde{V}_{t+1}^k(s') - \widetilde{V}_{t+1}^k(s_{t+1}))_+$$

where $(\cdot)_+ = \max\{0, \cdot\}$. It is unclear how to bound the sum of $(\min_{s' \in \{s_1, \ldots, s_{t_k}\}} \widetilde{V}_{t+1}^k(s') - \widetilde{V}_{t+1}^k(s_{t+1}))_+$ over $t = 1, \ldots, T$. We conjecture that additional algorithmic technique is required to proceed with the bound.

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.

Algorithm 3: FOPO

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

Initialize: $\Lambda_1 = \lambda I$

1 Receive initial state s_1.

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

3 $\quad$ Solve the following optimization problem to get w_t:

$$\max_{w_t, b_t \in \mathbb{R}^d, J_t \in \mathbb{R}} \quad J_t$$

$$\text{subject to} \quad w_t = \Lambda_t^{-1} \sum_{\tau=1}^{t-1} (\varphi(s_\tau, a_\tau)(r(s_\tau, a_\tau) - J_t + \max_a \langle \varphi(s_{\tau+1}, a), w_t \rangle) + b_t)$$

$$\|b_t\|_{\Lambda_t} \leq \beta$$

$$\|w_t\| \leq (2 + \text{sp}(v^*))\sqrt{d}$$

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 (Wei et al. 2021)

In this section, we provide time complexity analysis of the FOPO algorithm (Wei et al. 2021). The algorithm is shown in Algorithm 3.

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:

$$G_w(\Delta) = \{\Delta \cdot (k_1, \ldots, k_d) : \pm k_1, \ldots, \pm k_d \in [\lfloor (2 + \text{sp}(v^*))/\Delta \rfloor]\}$$

$$G_b(\Delta) = \{\Delta \cdot (k_1, \ldots, k_d) : \pm k_1, \ldots, \pm k_d \in [\lfloor T\beta/(\sqrt{d}\Delta) \rfloor]\}$$

$$G_J(\Delta) = \{\Delta k : k \in [\lfloor 1/\Delta \rfloor]\}.$$

The grids are designed such that the constraints $\|w\|_2 \leq (2 + \text{sp}(v^*))\sqrt{d}$, $\|b\|_{\Lambda_t} \leq \beta$ 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 $\mathcal{B}_d((2 + \text{sp}(v^*))\sqrt{d})$, $\mathcal{B}_d(T\beta)$ 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.,

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

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,

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

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

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

Inspecting the proof in Appendix C.1 in Wei et al. (2021), it can be seen that the additional regret incurred by approximating the solution of the fixed point problem is

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

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 (He et al. 2024)

The LOOP algorithm (He et al. 2024) solves the following optimization problem every episode:

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

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 18 (Lemma D.1 in Jin et al. (2020)). *Let* $\Lambda_t = \sum_{i=1}^{t} \phi_i \phi_i^T + \lambda I$ *where* $\phi_i \in \mathbb{R}^d$ *and* $\lambda > 0$. *Then,*

$$\sum_{i=1}^{t} \phi_i^T \Lambda_t^{-1} \phi_i \leq d.$$

Lemma 19 (Lemma 11 in Abbasi-Yadkori et al. (2011)). *Let $\{\phi_t\}_{t\geq 1}$ be a bounded sequence in $\mathbb{R}^d$ with $\|\phi_t\|_2 \leq 1$ for all $t \geq 1$. Let $\Lambda_0 = I$ and $\Lambda_t = \sum_{i=1}^{t} \phi_i \phi_i^T + I$ for $t \geq 1$. Then,*

$$\sum_{i=1}^{t} \phi_i^T \Lambda_{i-1}^{-1} \phi_i \leq 2 \log \det(\Lambda_t) \leq 2d \log(1 + t).$$

Lemma 20 (Lemma 12 in Abbasi-Yadkori et al. (2011)). *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*

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

Lemma 21 (Bound on number of episodes). *The number of episodes K in Algorithm 2 is bounded by*

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

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.

$\square$

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. (2008) 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. (2009) 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. (2018) propose a model-based algorithm called SCAL, which is a computationally efficient version of REGAL. Zhang et al. (2019) 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. (2020) 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. (2023) introduce a Q-learning-based

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

Algorithm	Regret $\widetilde{\mathcal{O}}(\cdot)$	Assumption	Computation
UCRL2 (Auer et al. 2008)	$DS\sqrt{AT}$	Bounded diameter	Efficient
REGAL (Bartlett et al. 2009)	$\text{sp}(v^*)\sqrt{SAT}$	Weakly communicating	Inefficient
PSRL (Ouyang et al. 2017)	$\text{sp}(v^*)S\sqrt{AT}$	Weakly communicating	Efficient
OSP (Ortner 2020)	$\sqrt{t_{\text{mix}}SAT}$	Ergodic	Inefficient
SCAL (Fruit et al. 2018)	$\text{sp}(v^*)S\sqrt{AT}$	Weakly communicating	Efficient
UCRL2B (Fruit et al. 2020)	$S\sqrt{DAT}$	Bounded diameter	Efficient
EBF (Zhang et al. 2019)	$\sqrt{\text{sp}(v^*)SAT}$	Weakly communicating	Inefficient
γ-UCB-CVI (Ours)	$\text{sp}(v^*)S\sqrt{AT}$	Bellman optimality equation	Efficient
Optimistic Q-learning (Wei et al. 2020)	$\text{sp}(v^*)(SA)^{\frac{1}{3}}T^{\frac{2}{3}}$	Weakly communicating	Efficient
MDP-OOMD (Wei et al. 2020)	$\sqrt{t_{\text{mix}}^3\eta AT}$	Ergodic	Efficient
UCB-AVG (Zhang et al. 2023)	$\text{sp}(v^*)S^5A^2\sqrt{T}$	Weakly communicating	Efficient
Lower bound (Auer et al. 2008)	$\Omega(\sqrt{DSAT})$		

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. (2024) 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. (2024) 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 $\mathcal{O}(\sqrt{\text{sp}(v^*)^3d^3T})$, 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. (2023) 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 $\widetilde{\mathcal{O}}(\text{sp}(v^*)\sqrt{d^3T})$ 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. (2022) 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. (2021) study the linear MDP setting and propose a finite-horizon reduction that uses the LSVI-UCB (Jin et al. 2020). 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. (2020) 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. (2021) propose a model-based algorithm that adapts UCBVI (Azar et al. 2017) to the discounted setting and achieve a nearly minimax optimal regret bound. Ji et al. (2024) propose a model-free algorithm with nearly minimax optimal regret bound.