Learning Infinite-Horizon Average-Reward Linear Mixture MDPs of Bounded Span

Woojin Chae
KAIST

Kihyuk Hong
University of Michigan

Yufan Zhang
University of Michigan

Ambuj Tewari
University of Michigan

Dabeen Lee
KAIST

Abstract

This paper proposes a computationally tractable algorithm for learning infinite-horizon average-reward linear mixture Markov decision processes (MDPs) under the Bellman optimality condition. Our algorithm for linear mixture MDPs achieves a nearly minimax optimal regret upper bound of $\widetilde{\mathcal{O}}(d\sqrt{\mathrm{sp}(v^*)T})$ over T time steps where $\mathrm{sp}(v^*)$ is the span of the optimal bias function v^* and d is the dimension of the feature mapping. Our algorithm applies the recently developed technique of running value iteration on a discounted-reward MDP approximation with clipping by the span. We prove that the value iteration procedure, even with the clipping operation, converges. Moreover, we show that the associated variance term due to random transitions can be bounded even under clipping. Combined with the weighted ridge regression-based parameter estimation scheme, this leads to the nearly minimax optimal regret guarantee.

1 INTRODUCTION

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

deavors to expand our theoretical understanding of RL with function approximation.

As a first step toward establishing theoretical foundations, linear function approximation frameworks have received significant attention. The works on linear function approximation can be categorized based on how linearity is assumed on the structure of the underlying Markov decision process (MDP). There are largely four settings: MDPs with a low Bellman rank (Jiang et al., 2017), linear MDPs (Yang and Wang, 2019; Jin et al., 2020), linear mixture MDPs (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021c), and MDPs with a low inherent Bellman error (Zanette et al., 2020b).

Among them, notable progress has been made for linear mixture MDPs where the underlying transition kernel and the reward function are assumed to be parameterized as a linear function of some given feature mappings over state-action pairs or state-action-state triplets. For the finite-horizon setting, Zhou et al. (2021a) developed a nearly minimax optimal algorithm. For the infinite-horizon discounted-reward case, Zhou et al. (2021c) established a regret lower bound. Shortly after this, Zhou et al. (2021b) announced an algorithm with a regret upper bound matching the lower bound up to logarithmic factors. For the infinite-horizon average-reward regime, Wu et al. (2022) showed a regret lower bound of $\Omega(d\sqrt{DT})$ from a communicating MDP instance with diameter D where d is the dimension of the feature map and T is the horizon. Moreover, they designed an algorithm that achieves a regret upper bound of $\widetilde{\mathcal{O}}(d\sqrt{DT})$, establishing near minimax optimality for the class of communicating MDPs.

For the infinite-horizon average-reward setting, however, the class of communicating MDPs is perhaps not the most general set of MDPs for which a learning algorithm can guarantee a sublinear regret (Bartlett and

Tewari, 2009). Although the diameter captures the number of steps needed to recover from a bad state to a good state, the actual regret incurred while recovering is better represented by the span $\mathrm{sp}(v^*)$ of the optimal bias function v^* (Fruit et al., 2018). While the diameter is an upper bound on the span, it can be arbitrarily larger than the span, and in fact, a weakly communicating MDP can have a finite span but an infinite diameter (Bartlett and Tewari, 2009).

The recent framework of He et al. (2024), LOOP, can be applied to infinite-horizon average-reward linear mixture MDPs of bounded span. Although LOOP guarantees a sublinear regret upper bound that depends on the span, LOOP is hardly practical as it relies on solving a complex constrained optimization problem. This motivates the following question.

Does there exist a computationally efficient, nearly minimax optimal algorithm for learning infinite-horizon average-reward linear mixture MDPs of bounded span?

This paper answers the question affirmatively. Let us summarize our contributions in Table 1 and as follows.

- We propose a computationally efficient algorithm, upper-confidence linear kernel reinforcement learning with clipping (UCLK-C; Algorithm 1) that achieves a regret upper bound of $\widetilde{\mathcal{O}}(d\sqrt{\mathrm{sp}(v^*)T})$.

- We deduce a regret lower bound of $\Omega(d\sqrt{\mathrm{sp}(v^*)T})$ by refining the regret lower bound analysis of Wu et al. (2022). This shows that UCLK-C is nearly minimax optimal.

- UCLK-C runs with a novel value iteration scheme by applying the clipping operation within discounted extended value iteration. The clipping operation due to Hong et al. (2025) is for controlling the span of intermediate value functions, which is crucial to provide a bounded regret for MDPs of bounded span. The clipping operation is much simpler to implement than constrained optimization-based frameworks to control the span. To run discounted extended value iteration, we approximate a given average-reward MDP by a discounted-reward MDP, as in UCLK due to Zhou et al. (2021c).

- We prove that for linear mixture MDPs, the discounted extended value iteration converges even with clipping. Moreover, we show that the associated variance term due to random transitions can be bounded even under the clipping

operation. Combined with the variance-aware weighted ridge regression-based parameter estimation scheme due to Wu et al. (2022), we deduce our nearly minimax optimal regret upper bound.

The idea of approximating an average-reward MDP by a discounted-reward MDP has been adopted for the tabular case (Wei et al., 2020; Zhang and Xie, 2023) and used for learning linear MDPs (Hong et al., 2025). Clipping an optimistic value function estimator to control its size is already a common practice when designing an algorithm for finite-horizon and infinite-horizon discounted-reward MDPs. However, the clipping operation in our algorithm sets the threshold in a different way to control the span of value functions, not their sizes, and it was first introduced by Hong et al. (2025). For linear MDPs, due to the clipping operation, convergence of value iteration is not guaranteed. In contrast, for linear mixture MDPs, we establish convergence of value iteration even with clipping.

2 RELATED WORK

Reinforcement Learning with Linear Function Approximation Recently, there has been remarkable progress in reinforcement learning frameworks with linear function approximation (Jiang et al., 2017; Yang and Wang, 2019, 2020; Jin et al., 2020; Wang et al., 2021; Modi et al., 2020; Dann et al., 2018; Du et al., 2021; Sun et al., 2019; Zanette et al., 2020a,b; Cai et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Weisz et al., 2021; Zhou et al., 2021c,a; He et al., 2021; Zhou and Gu, 2022; Hu et al., 2022; He et al., 2023; Agarwal et al., 2023; Hong et al., 2025). These works develop frameworks for MDP classes with certain linear structures. Among them, the most relevant to this paper are linear and linear mixture MDPs. Linear and linear mixture MDPs assume that the transition probability and the reward function are linear in some feature mappings over state-action pairs or state-action-state triplets. Although the two classes are closely related, one cannot be covered by the other (Zhou et al., 2021c). For learning infinite-horizon average-reward linear MDPs, Wei et al. (2021) developed several algorithms, including FOPO. FOPO achieves the best-known regret upper bound, but it needs to solve a fixed-point equation at each iteration, making the algorithm intractable. Recently, Hong et al. (2025) proposed a provably efficient algorithm that achieves the best-known regret upper bound for the setting. For learning infinite-horizon average-reward linear mixture MDPs, Wu et al. (2022) developed an algorithm that is shown to be minimax optimal for the communicating case. He et al. (2024) proposed an algorithm for RL with general function approximation, LOOP, incorporating

Table 1: Summary of Our Results on Regret Upper and Lower Bounds for Learning Linear Mixture MDPs

Setting	Regret Upper Bound	Regret Lower Bound
Communicating (bounded diameter)	$\widetilde{\mathcal{O}}\left(d\sqrt{DT}\right)$ (Wu et al., 2022)	$\Omega\left(d\sqrt{DT}\right)$ (Wu et al., 2022)
Bellman optimality (bounded span)	$\widetilde{\mathcal{O}}\left(d\sqrt{\mathrm{sp}(v^*)T}\right)$ (Theorem 1)	$\Omega\left(d\sqrt{\mathrm{sp}(v^*)T}\right)$ (Theorem 2)

linear and linear mixture MDPs as subclasses.

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

3 PRELIMINARIES

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

Infinite-Horizon Average-Reward MDP We consider an MDP given by $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathbb{P}(\cdot \mid s, a)$ specifies the transition probability function for state s with taking action a, and $r(s, a) \in [0, 1]$ is the reward from action a at state s. A (stochastic) stationary policy is given as a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ where $\Delta(\mathcal{A})$ is the set of probability measures on $\mathcal{A}$, and we use notation

$\pi(a \mid s)$ for the probability of taking action a at state s under policy π. When π is a deterministic policy, we write that $a = \pi(s)$ with abuse of notation where a is the action with $\pi(a \mid s) = 1$. At each time step t, an algorithm takes action a_t at given state s_t, after which it observes the next state s_{t+1} drawn from distribution $\mathbb{P}(\cdot \mid s_t, a_t)$. Then the cumulative reward over T steps is $\sum_{t=1}^T r(s_t, a_t)$. Then the (long-term) average reward is given by $\liminf_{T \to \infty} \mathbb{E}[\sum_{t=1}^T r(s_t, a_t)]/T$, which can be maximized by a deterministic stationary policy (See Puterman, 2014). We denote by $J^\pi(s) = \liminf_{T \to \infty} \mathbb{E}[\sum_{t=1}^T r(s_t, a_t) \mid s_1 = s]/T$ the average reward of a stationary policy π starting from initial state s.

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

$$
\begin{aligned}
J^* + q^*(s, a) &= r(s, a) + \mathbb{E}_{s' \sim \mathbb{P}(\cdot \mid s, a)}\left[v^*(s')\right], \\
v^*(s) &= \max_{a \in \mathcal{A}} q^*(s, a).
\end{aligned}
\tag{1}
$$

Under the Bellman optimality condition, the optimal average reward $J^*(s) := \max_\pi J^\pi(s)$ is invariant with the initial state s, and $J^*(s) = J^*$ for any $s \in \mathcal{S}$ (Bartlett and Tewari, 2009). Moreover, the class of weakly communicating MDPs satisfies the condition (See Puterman, 2014). There indeed exist other general classes of MDPs with which the condition holds (Hernandez-Lerma, 2012, Section 3.3). For any function $h : \mathcal{S} \to \mathbb{R}$, we define its span as $\mathrm{sp}(h) := \max_{s \in \mathcal{S}} h(s) - \min_{s \in \mathcal{S}} h(s)$. Then following the literature on infinite-horizon average-reward RL (Auer et al., 2008), we consider the regret function $\mathrm{Regret}(T) = T \cdot J^* - \sum_{t=1}^T r(s_t, a_t)$ to assess the performance of an algorithm.

Discounted-Reward MDP We also consider the discounted cumulative reward of a stationary policy π given by $V^\pi(s) = \mathbb{E}[\sum_{t=1}^\infty \gamma^{t-1} r(s_t, a_t) \mid s_1 = s]$ where s is the initial state and $\gamma \in (0, 1)$ is a discount factor. Similarly, we consider $Q^\pi(s, a) = \mathbb{E}[\sum_{t=1}^\infty \gamma^{t-1} r(s_t, a_t) \mid (s_1, a_1) = (s, a)]$. Then we define the optimal value function V^* and the optimal action-value function Q^* as $V^*(s) = \max_\pi V^\pi(s)$ and

$Q^*(s,a) = \max_\pi Q^\pi(s,a)$ for $(s,a) \in \mathcal{S} \times \mathcal{A}$. It is known that there exists a deterministic stationary policy that gives rise to V^* and Q^* (See Puterman, 2014; Agarwal et al., 2021). Moreover, V^* and Q^* satisfy the following Bellman optimality equation.

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

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

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

Assumption 1. (Linear Mixture MDP, Zhou et al., 2021a; Wu et al., 2022) *There is a known feature map* $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ *such that for any* $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,

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

where $\theta^* \in \mathbb{R}^d$ *is an unknown vector.*

We assume that the reward function is deterministic and known to the decision-maker, but our results easily extend to the setting where the reward function is given by $r(s,a) = \langle \varphi(s,a), \theta^* \rangle$ for some feature map $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. For any function $F : \mathcal{S} \to \mathbb{R}$, we use the following shorthand notations.

$$[\mathbb{P}F](s,a) = \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}[F(s')]$$
$$[\mathbb{V}F](s,a) = [\mathbb{P}F^2](s,a) - ([\mathbb{P}F](s,a))^2$$

Defining $\phi_F(s,a) = \int_{s'} \phi(s,a,s')F(s')ds'$, we have

$$\langle \phi_F(s,a), \theta^* \rangle = \int \langle \phi(s,a,s'), \theta^* \rangle F(s')ds' = [\mathbb{P}F](s,a).$$

Therefore, we also have

$$[\mathbb{V}F](s,a) = \langle \phi_{F^2}(s,a), \theta^* \rangle - \langle \phi_F(s,a), \theta^* \rangle^2.$$

Following Wu et al. (2022), we assume that the scales of the parameters are bounded as follows.

Assumption 2. θ^* *satisfies* $\|\theta^*\|_2 \leq B_\theta$ *for some* $B_\theta \in \mathbb{R}$. *Moreover, for any* $H > 0$, $F : \mathcal{S} \to [0,H]$, *and* $(s,a) \in \mathcal{S} \times \mathcal{A}$, *it holds that* $\|\phi_F(s,a)\|_2 \leq H$.

4   THE PROPOSED ALGORITHM

In this section, we present our algorithm, `UCLK-C`, described in Algorithm 1. As common in algorithms for learning infinite-horizon average-reward MDPs such as

`UCRL2` (Auer et al., 2008) and `UCRL2-VTR` (Wu et al., 2022), `UCLK-C` also proceeds with episodes. Following `UCRL2-VTR`, when to start the next episode is determined based on the Gram matrix (line 14). Each episode of `UCLK-C` consists of two phases, the planning phase (lines 4–13) and the execution phase (lines 14–21). During the planning phase, we run extended value iteration for a discounted MDP with estimated parameters. Then, based on value functions deduced from the planning phase, we take and execute a greedy deterministic (non-stationary) policy for the execution phase. What follows provides a more detailed discussion of the components of `UCLK-C`.

Discounted Value Iteration   As in Hong et al. (2025), we apply value iteration on a discounted-reward approximation of the underlying MDP. For each episode k, we take a confidence ellipsoid $\mathcal{C}_k$ (line 4) over which we run extended value iteration (line 7). We make sure that any $\theta \in \mathcal{C}_k$ induces a probability distribution, i.e. $\mathcal{C}_k \subseteq \mathcal{B}$ where

$$\mathcal{B} = \left\{ \theta \in \mathbb{R}^d : \begin{array}{l} \left\langle \int \phi(s,a,s')ds', \theta \right\rangle = 1, \\ \langle \phi(s,a,s'), \theta \rangle \geq 0 \end{array} \forall (s,a,s') \right\}.$$

As a result, we get that $Q^{(n)}(s,a) \leq (1-\gamma)^{-1}$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and n. Note that we apply multiple rounds of value iteration. We will show that even with the clipping operation (line 9), the value iteration procedure stated in lines 6–10 converges.

Clipping Operation   In each round of value iteration, `UCLK-C` applies the clipping operation stated in line 9. Note that the value function $\widetilde{V}^{(n)}$ from line 8 does not necessarily have a bounded span. After the clipping operation, it is clear that the span of $V^{(n)}$ from line 9 becomes bounded above as $\mathrm{sp}(V^{(n)}) \leq H$. As a result, the re-centering step in line 12 guarantees that $W_k(s) \in [0,H]$ for any $s \in \mathcal{S}$, which is crucial to parameterize estimation errors as a function of H, not $(1-\gamma)^{-1}$ which is set as large as $O(\sqrt{T})$. Note that the Gram matrix update steps (lines 17–18) are with respect to W_k, not V_k. We choose any upper bound H on $2 \cdot \mathrm{sp}(v^*)$ where v^* is the optimal bias function from the Bellman optimality condition (1).

Variance-Aware Ridge Regression-Based Parameter Estimation   We closely follow the Bernstein inequality-based estimation scheme of Wu et al. (2022). The idea is to build confidence ellipsoids for θ^* based on a Bernstein-type concentration inequality for linear bandits. To be more precise, we apply the following lemma for vector-valued martingales.

Lemma 1. (Theorem 4.1, Zhou et al., 2021a) *Let* $\{\mathcal{G}_t\}_{t=1}^\infty$ *be a filtration,* $\{x_t, \eta_t\}_{t \geq 1}$ *a stochastic process*

Algorithm 1 Upper-Confidence Linear Kernel Reinforcement Learning with Clipping (UCLK-C)

1: **Input:** upper bound H of $2 \cdot \text{sp}(v^*)$, feature map $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, confidence level $\delta \in (0, 1)$, discount factor $\gamma \in [0, 1)$, number of rounds N, and parameters λ, B_θ
2: **Initialize:** $t \leftarrow 1$, $\widehat{\Sigma}_1, \widetilde{\Sigma}_1 \leftarrow \lambda I_d$, $\widehat{b}_1, \widetilde{b}_1, \widehat{\theta}_1, \widetilde{\theta}_1 \leftarrow 0$, and observe the initial state $s_1 \in \mathcal{S}$
3: **for** episodes $k = 1, 2, \ldots,$ **do**
4:     Set $t_k = t$ and $\mathcal{C}_k = \widehat{\mathcal{C}}_{t_k}$ given in (3)
5:     Initialize $Q^{(0)}(\cdot, \cdot) \leftarrow (1 - \gamma)^{-1}$ and $V^{(0)}(\cdot) \leftarrow (1 - \gamma)^{-1}$
6:     **for** rounds $n = 1, 2, \ldots, N$ **do**
7:         Set $Q^{(n)}(\cdot, \cdot) \leftarrow r(\cdot, \cdot) + \gamma \cdot \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(n-1)}}(\cdot, \cdot), \theta \rangle$
8:         Set $\widetilde{V}^{(n)}(\cdot) \leftarrow \max_{a \in \mathcal{A}} Q^{(n)}(\cdot, a)$
9:         Set $V^{(n)}(\cdot) \leftarrow \min\{\widetilde{V}^{(n)}(\cdot), \min_{s' \in \mathcal{S}} \widetilde{V}^{(n)}(s') + H\}$
10:    **end for**
11:    Set $Q_k(\cdot, \cdot) \leftarrow Q^{(N)}(\cdot, \cdot)$ and $V_k(\cdot) \leftarrow V^{(N)}(\cdot)$
12:    Set $W_k(\cdot) \leftarrow V_k(\cdot) - \min_{s' \in \mathcal{S}} V_k(s')$
13:    Take a deterministic policy π_k given by $\pi_k(\cdot) \in \text{argmax}_{a \in \mathcal{A}} Q_k(\cdot, a)$
14:    **while** $\det(\widehat{\Sigma}_t) \leq 2 \det(\widehat{\Sigma}_{t_k})$ **do**
15:       Take action $a_t \leftarrow \pi_k(s_t)$, receive reward $r(s_t, a_t)$ and next state $s_{t+1} \sim \mathbb{P}(\cdot | s_t, a_t)$
16:       Set $\bar{\sigma}_t \leftarrow \sqrt{\max\left\{H^2/d, [\bar{\mathbb{V}}_t W_k](s_t, a_t) + E_t\right\}}$ where $[\bar{\mathbb{V}}_t W_k](s_t, a_t)$ and E_t are given as in (4) and (5)
17:       Set $\widehat{\Sigma}_{t+1} \leftarrow \widehat{\Sigma}_t + \bar{\sigma}_t^{-2} \phi_{W_k}(s_t, a_t) \phi_{W_k}(s_t, a_t)^\top$ and $\widehat{b}_{t+1} \leftarrow \widehat{b}_t + \bar{\sigma}_t^{-2} W_k(s_{t+1}) \phi_{W_k}(s_t, a_t)$
18:       Set $\widetilde{\Sigma}_{t+1} \leftarrow \widetilde{\Sigma}_t + \phi_{W_k^2}(s_t, a_t) \phi_{W_k^2}(s_t, a_t)^\top$ and $\widetilde{b}_{t+1} \leftarrow \widetilde{b}_t + W_k^2(s_{t+1}) \phi_{W_k^2}(s_t, a_t)$
19:       Set $\widehat{\theta}_{t+1} \leftarrow \widehat{\Sigma}_{t+1}^{-1} \widehat{b}_{t+1}$ and $\widetilde{\theta}_{t+1} \leftarrow \widetilde{\Sigma}_{t+1}^{-1} \widetilde{b}_{t+1}$
20:       Set $t \leftarrow t + 1$
21:    **end while**
22: **end for**

---

such that $x_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable while $\eta_t \in \mathbb{R}$ is $\mathcal{G}_{t+1}$-measurable. For $t \geq 1$, let $y_t = \langle x_t, \mu^* \rangle + \eta_t$ where $|\eta_t| \leq R$, $\mathbb{E}[\eta_t \mid \mathcal{G}_t] = 0$, $\mathbb{E}[\eta_t^2 \mid \mathcal{G}_t] \leq \sigma^2$, and $\|x_t\|_2 \leq L$ for some fixed $R, L, \sigma, \lambda > 0$ and $\mu^* \in \mathbb{R}^d$. Then, for any $0 < \delta < 1$, it holds with probability at least $1 - \delta$ that for every $t \geq 1$,

$$\|\mu_t - \mu^*\|_{Z_t} \leq \beta_t + \sqrt{\lambda} \|\mu^*\|_2$$

where $\beta_t = 8\sigma\sqrt{d \log(1 + tL^2/(d\lambda)) \log(4t^2/\delta)} + 4R \log(4t^2/\delta)$, $\mu_t = Z_t^{-1} b_t$, $Z_t = \lambda I_d + \sum_{i=1}^t x_i x_i^\top$, and $b_t = \sum_{i=1}^t y_i x_i$.

Note that $\widehat{\theta}_{t+1}$ in line 19 corresponds to μ_t in Lemma 1 by setting $(x_j, y_j) = (\bar{\sigma}_j^{-1} \phi_{W_i}(s_j, a_j), \bar{\sigma}_j^{-1} W_i(s_{j+1}))$ for $j \in [t_i : t_{i+1} - 1]$ and $i \in [k]$. Then $\widehat{\theta}_{t_k}$ is the solution of the following weighted ridge regression.

$$\min_{\theta \in \mathbb{R}^d} \lambda \|\theta\|_2^2 + \sum_{i=1}^{k-1} \sum_{j=t_i}^{t_{i+1}-1} \frac{(W_i(s_{j+1}) - \langle \phi_{W_i}(s_j, a_j), \theta \rangle)^2}{\bar{\sigma}_j^2}.$$

Here, $\bar{\sigma}_j^2$ is an estimator of the (conditional) variance of $W_i(s_{j+1})$, given by $[\mathbb{V} W_i](s_j, a_j)$.

We will choose the value of $\bar{\sigma}_t$ (line 16) for $t \in [t_k, t_{k+1} - 1]$ so that $|\eta_t| \leq \sqrt{d}$, $\mathbb{E}[\eta_t^2 \mid \mathcal{G}_t] \leq 1$, and $|x_t| \leq \sqrt{d}$, where $\eta_t = y_t - \langle x_t, \theta^* \rangle$ and $\mathcal{G}_t = \sigma(s_1, \ldots, s_t)$ is the σ-algebra generated by $s_1, \ldots, s_t$.

Then, based on Lemma 1, we may construct

$$\widehat{\mathcal{C}}_t = \left\{\theta \in \mathcal{B} : \|\theta - \widehat{\theta}_t\|_{\widehat{\Sigma}_t} \leq \widehat{\beta}_t\right\} \tag{3}$$

where we set

$$\begin{aligned}\widehat{\beta}_t = {} & 8\sqrt{d \log(1 + t/\lambda) \log(4t^2/\delta)} \\ & + 4\sqrt{d} \log(4t^2/\delta) + \sqrt{\lambda} B_\theta.\end{aligned}$$

We make sure that $\bar{\sigma}_t$ is strictly positive. Moreover, to guarantee $\mathbb{E}[\eta_t^2 \mid \mathcal{G}_t] \leq 1$, we take the right quantity for $\bar{\sigma}_t$ so that $\bar{\sigma}_t^2$ is an upper bound on the variance term given by

$$[\mathbb{V} W_k](s_t, a_t) = \langle \phi_{W_k^2}(s_t, a_t), \theta^* \rangle - \langle \phi_{W_k}(s_t, a_t), \theta^* \rangle^2.$$

To estimate this, we first take $[\bar{\mathbb{V}}_t W_k](s_t, a_t)$ given by

$$\begin{aligned}[\bar{\mathbb{V}}_t W_k](s_t, a_t) = {} & \left[\langle \phi_{W_k^2}(s_t, a_t), \widetilde{\theta}_t \rangle\right]_{[0, H^2]} \\ & - \left[\langle \phi_{W_k}(s_t, a_t), \widehat{\theta}_t \rangle\right]_{[0, H]}^2\end{aligned} \tag{4}$$

where $[x]_{[a,b]}$ denotes the projection of x onto the interval $[a, b]$ and $\widetilde{\theta}_t$ (line 19) corresponds to μ_{t-1} in Lemma 1 by setting $(x_j, y_j) = (\phi_{W_i^2}(s_j, a_j), W_i^2(s_{j+1}))$ for $j \in [t_i : t_{i+1} - 1]$ and $i \in [k]$. Note that $\widetilde{\theta}_t$ is the solution of the (unweighted) ridge regression problem

with contexts $\phi_{W_i^2}(s_j, a_j)$ and targets $W_i^2(s_{j+1})$ for $j \in [t_i : t_{i+1} - 1]$ and $i \in [k]$.

The last ingredient is to take an upper bound E_t on the error term $|[\mathbb{V}W_k](s_t, a_t) - [\bar{\mathbb{V}}_t W_k](s_t, a_t)|$. We take

$$
\begin{aligned}
E_t = & \min \left\{ H^2, 2H\check{\beta}_t \|\phi_{W_k}(s_t, a_t)\|_{\widehat{\Sigma}_t^{-1}} \right\} \\
& + \min \left\{ H^2, \widetilde{\beta}_t \|\phi_{W_k^2}(s_t, a_t)\|_{\widetilde{\Sigma}_t^{-1}} \right\}
\end{aligned}
\tag{5}
$$

where we set

$$
\begin{aligned}
\check{\beta}_t = & \, 8d\sqrt{\log(1 + t/\lambda) \log(4t^2/\delta)} \\
& + 4\sqrt{d} \log(4t^2/\delta) + \sqrt{\lambda} B_\theta, \\
\widetilde{\beta}_t = & \, 8H^2\sqrt{d \log(1 + tH^2/(d\lambda)) \log(4t^2/\delta)} \\
& + 4H^2 \log(4t^2/\delta) + \sqrt{\lambda} B_\theta.
\end{aligned}
$$

Here, we may observe from Lemma 1 that $\|\theta^* - \widehat{\theta}_t\|_{\widehat{\Sigma}_t} \leq \check{\beta}_t$ and $\|\theta^* - \widetilde{\theta}_t\|_{\widehat{\Sigma}_t} \leq \widetilde{\beta}_t$ with high probability, based on which we prove that E_t provides an upper bound on the error term.

Finally, we set the value of $\bar{\sigma}_t$ as

$$
\bar{\sigma}_t = \sqrt{\max \left\{ H^2/d, [\bar{\mathbb{V}}_t W_k](s_t, a_t) + E_t \right\}}.
$$

Note that $\bar{\sigma}_t$ is well-defined because the term inside the square root is strictly positive. Moreover, $\bar{\sigma}_t^2$ is greater than or equal to $[\bar{\mathbb{V}}_t W_k](s_t, a_t) + E_t$ which is an upper bound on the variance term $[\mathbb{V}W_k](s_t, a_t)$. To formalize this, we prove the following lemma.

Lemma 2. *With probability at least $1 - 3\delta$, it holds that for every $t \in [T]$,*

$$
|[\mathbb{V}W_k](s_t, a_t) - [\bar{\mathbb{V}}_t W_k](s_t, a_t)| \leq E_t, \quad \theta^* \in \widehat{\mathcal{C}}_t.
$$

Computational Efficiency Our algorithm UCLK-C runs in episodes with variable length and the number of episodes is $\mathcal{O}(d \log T)$. In each episode, we run $N = \Theta(\sqrt{HT/d} \log T)$ steps of value iterations to obtain a policy to execute in the episode. The computational bottleneck when running a step of value iteration is computing $\max_{\theta \in \mathcal{C}_k} \langle \phi_V(\cdot, \cdot), \theta \rangle$ where $\mathcal{C}_k$ is of the form $\{\theta \in \mathcal{B} : \|\theta - \widehat{\theta}\|_\Sigma \leq \beta\}$ and $\mathcal{B} = \{\theta \in \mathbb{R}^d : \langle \int \phi(s, a, s'') ds'', \theta \rangle = 1, \langle \phi(s, a, s'), \theta \rangle \geq 0, \forall(s, a, s')\}$. This optimization is also used by Zhou et al. (2021a); Wu et al. (2022), and as they discuss, using the fact that $\mathcal{C}_k$ is convex, we can solve the optimization problem efficiently. The bottleneck is computing $\phi_V(\cdot, \cdot) = \int \phi(\cdot, \cdot, s') V(s') ds$, which takes $\mathcal{O}(S)$ computations where S is the size of the state space, which is also incurred by Wu et al. (2022). Overall, our algorithm runs in time polynomial in the problem parameters d, S, A, T. This is an improvement over the general algorithm LOOP (He et al., 2024) reduced to the

linear mixture MDP setting, as they require solving an intractable max-min problem which requires time exponential in d when brute-forcing as discussed by Hong et al. (2025).

5 REGRET ANALYSIS OF UCLK-C

Let us state the following regret bound of UCLK-C for linear mixture MDPs.

Theorem 1. *Set $H \geq 2 \cdot \mathrm{sp}(v^*)$, $\gamma = 1 - \sqrt{d}/\sqrt{HT}$, $N \geq \sqrt{HT/d} \log(\sqrt{T}/d\sqrt{H})$, and $\lambda = 1/B_\theta^2$. Then UCLK-C guarantees with probability at least $1 - 5\delta$ that for any linear mixture MDP with any initial state,*

$$
\mathrm{Regret}(T) = \widetilde{\mathcal{O}} \left(d\sqrt{HT} + H\sqrt{dT} + d^{7/4} HT^{1/4} \right)
$$

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

By taking $H = 2 \cdot \mathrm{sp}(v^*)$, as a corollary of Theorem 1, we deduce that

$$
\mathrm{Regret}(T) = \widetilde{\mathcal{O}} \left(d\sqrt{\mathrm{sp}(v^*)T} \right)
$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in $T\mathrm{sp}(v^*)B_\theta/\delta$. The rest of this section gives a proof overview of Theorem 1.

Remark 1. *Our result requires the knowledge of an upper bound H of $2 \cdot sp(v^*)$. Such a knowledge of an upper bound is a commonly made assumption in the infinite-horizon average-reward setting (Bartlett and Tewari, 2009). A well-known upper bound of the span $sp(v^*)$ is the diameter of the MDP, which is defined to be the expected number of steps needed to transition between any two states in the worst case. Relaxing the assumption of the knowledge of $sp(v^*)$ or its upper bound has only been achieved recently in the tabular setting (Boone and Zhang, 2024). We leave relaxing the assumption in the linear mixture MDP setting as future work.*

Let us start by establishing convergence of the discounted extended value iteration procedure with clipping. For episode k, we consider the value functions $\widetilde{V}^{(n)}$ and $V^{(n)}$ for $n \in [N]$ (lines 7–9 of Algorithm 1). Recall that $V^{(n)}$ is obtained from $\widetilde{V}^{(n)}$ after applying the clipping operation. It turns out that the clipping operation is a contraction map.

Lemma 3. *For any $n \in [N]$, it holds that*

$$
\max_{s \in \mathcal{S}} (V^{(n-1)}(s) - V^{(n)}(s)) \leq \max_{s \in \mathcal{S}} (\widetilde{V}^{(n-1)}(s) - \widetilde{V}^{(n)}(s)).
$$

Remark 2. *The contraction property, crucial for achieving minimax regret bound, is possible only because we use extended value iteration, rather than an*

optimistic value iteration that adds bonus terms at each iteration. With optimistic value iteration, as in Hong et al. (2025), the difference in successive value functions is lower bounded by a bonus term, preventing the contraction argument from holding.

Recall that Q_k is the action-value function for the kth episode and that V_k denotes the value function for the kth episode, obtained from clipping $\widetilde{V}_k$ with $\widetilde{V}_k(s) = \max_{a \in \mathcal{A}} Q_k(s,a)$ for $s \in \mathcal{S}$ (line 11 of Algorithm 1). Based on Lemma 3, we prove the following lemma.

Lemma 4. *Suppose that $\theta^* \in \widehat{\mathcal{C}}_t$ for $t \in [T]$ where $\mathcal{C}_t$ is defined as in (3). Then for each episode k and $t_k \leq t < t_{k+1} - 1$, it holds that*

$$Q_k(s_t, a_t) \leq r(s_t, a_t) + \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V_k}(s_t, a_t), \theta \rangle + \gamma^N.$$

With Lemma 4, we may provide a decomposition of the regret function as follows. Let K_T denote the total number of distinct episodes over the horizon of T time steps. For simplicity, we assume that the last time step of the last episode and that time step $T + 1$ is the beginning of the $(K_T + 1)$th episode, i.e., $t_{K_T+1} = T + 1$. Then it follows from Lemma 4 that

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

$$\leq T\gamma^N + \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} (J^* - (1-\gamma)V_k(s_{t+1}))}_{I_1}$$

$$+ \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} (V_k(s_{t+1}) - Q_k(s_t, a_t))}_{I_2}$$

$$+ \gamma \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} (\langle \phi_{V_k}(s_t, a_t), \theta^* \rangle - V_k(s_{t+1}))}_{I_3}$$

$$+ \gamma \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \max_{\theta \in \mathcal{C}_k} \langle \phi_{V_k}(s_t, a_t), \theta - \theta^* \rangle}_{I_4}.$$

Regret Term I_1 Recall that V^* and Q^* are the optimal value function and the optimal action-value function for the discounted-reward setting with discount factor γ. The following lemma proves that V_k and Q_k are optimistic estimators of V^* and Q^*, respectively.

Lemma 5. *Suppose that $\theta^* \in \widehat{\mathcal{C}}_t$ for $t \in [T]$ where $\widehat{\mathcal{C}}_t$ is defined as in (3). Then for each episode k, $1/(1-\gamma) \geq V_k(s) \geq V^*(s)$ and $1/(1-\gamma) \geq Q_k(s,a) \geq Q^*(s,a)$.*

Lemma 5 implies that $J^* - (1-\gamma)V_k(s_{t+1}) \leq J^* - (1-\gamma)V^*(s_{t+1})$. This can be further bounded above based on the following lemma.

Lemma 6. *(Lemma 2, Wei et al., 2020) Let J^* and v^* be the optimal average reward and the optimal bias function given in (1), and let V^* be the optimal discounted value function given in (2) with discount factor $\gamma \in [0,1)$. Then it holds that*

$$\max_{s \in \mathcal{S}} |J^* - (1-\gamma)V^*(s)| \leq (1-\gamma)\text{sp}(v^*),$$

$$\text{sp}(V^*) \leq 2 \cdot \text{sp}(v^*).$$

Lemma 6 offers a tool to bridge an infinite-horizon average-reward MDP and its discounted-reward MDP approximation. In particular, we deduce that $I_1 \leq T(1-\gamma)\text{sp}(v^*) \leq d\sqrt{\text{sp}(v^*)T}$.

Regret Term I_2 Note that $V_k(s_{t+1}) \leq \widetilde{V}_k(s_{t+1}) = Q_k(s_{t+1}, a_{t+1})$ for $t \in [t_k : t_{k+1} - 2]$, which leads to a telescoping structure. To be precise, we have

$$I_2 \leq \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-2} (Q_k(s_{t+1}, a_{t+1}) - Q_k(s_t, a_t))$$

$$+ \sum_{k=1}^{K_T} \left(\frac{1}{1-\gamma} - Q_k(s_{t_{k+1}-1}, a_{t_{k+1}-1}) \right)$$

$$= -\sum_{k=1}^{K_T} Q_k(s_{t_k}, a_{t_k}) + \frac{K_T}{1-\gamma}.$$

The following lemma gives an upper bound on the number of episodes K_T.

Lemma 7. *If $\lambda = 1/B_\theta^2$, then $K(T) \leq 1 + d\log(1 + TH^2B_\theta^2/d)$.*

Then it follows form Lemma 7 that $I_2 = \widetilde{\mathcal{O}}(d\sqrt{HT})$ where $\widetilde{\mathcal{O}}(\cdot)$ hides a logarithmic factor in THB_θ.

Regret Term I_3 We first observe that

$$\langle \phi_{V_k}(s_t, a_t), \theta^* \rangle - \min_{s' \in \mathcal{S}} V_k(s') = \langle \phi_{W_k}(s_t, a_t), \theta^* \rangle$$

since $\langle \phi_{V_k}(s_t, a_t), \theta^* \rangle = [\mathbb{P}V_k](s_t, a_t)$. This implies that

$$I_3 = \gamma \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \underbrace{(\langle \phi_{W_k}(s_t, a_t), \theta^* \rangle - W_k(s_{t+1}))}_{\eta_t}.$$

Then $\{\eta_t\}_{t=1}^{T}$ is a martingale difference sequence. Moreover, $|\eta_t| \leq H$ as $W_k(s) \in [0, H]$ for any $s \in \mathcal{S}$. Therefore, applying the Azuma-Hoeffding inequality, we deduce the following upper bound on I_3.

Lemma 8. *It holds with probability at least $1 - \delta$ that $I_3 \leq H\sqrt{2T\log(1/\delta)}$.*

Regret Term I_4 Let us sketch the idea of how the term I_4 can be bounded while a more rigorous proof is given in the appendix. Any $\theta \in \mathcal{C}_k$ induces a probability distribution, which implies that $\langle \phi_{V_k}(s_t, a_t), \theta - \theta^* \rangle = \langle \phi_{W_k}(s_t, a_t), \theta - \theta^* \rangle \in [-H, H]$. Moreover, assuming that $\theta^* \in \mathcal{C}_k$ based on Lemma 2,

$$\langle \phi_{W_k}(s_t, a_t), \theta - \theta^* \rangle$$

$$\leq \|\phi_{W_k}(s_t, a_t)\|_{\widehat{\Sigma}_t^{-1}} \left(\|\theta - \widehat{\theta}_{t_k}\|_{\widehat{\Sigma}_t} + \|\widehat{\theta}_{t_k} - \theta^*\|_{\widehat{\Sigma}_t} \right)$$

$$\leq 2 \|\phi_{W_k}(s_t, a_t)\|_{\widehat{\Sigma}_t^{-1}} \left(\|\theta - \widehat{\theta}_{t_k}\|_{\widehat{\Sigma}_{t_k}} + \|\widehat{\theta}_{t_k} - \theta^*\|_{\widehat{\Sigma}_{t_k}} \right)$$

$$\leq 4 \widehat{\beta}_T \bar{\sigma}_t \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}}$$

where the first inequality is from the Cauchy-Schwarz inequality, the second one holds because time step t is in episode k and thus $\det(\widehat{\Sigma}_t) \leq 2\det(\widehat{\Sigma}_{t_k})$, and the third one follows from $\theta^*, \theta \in \mathcal{C}_k$ and $\widehat{\beta}_{t_k} \leq \widehat{\beta}_T$. Then we may argue that I_4 is less than or equal to

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ H, 4\widehat{\beta}_T \bar{\sigma}_t \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}} \right\}$$

$$\leq 4\widehat{\beta}_T \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \bar{\sigma}_t \min \left\{ 1, \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}} \right\}$$

$$\leq 4\widehat{\beta}_T \times$$

$$\underbrace{\sqrt{\sum_{t=1}^{T} \bar{\sigma}_t^2}}_{J_1} \underbrace{\sqrt{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ 1, \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}}^2 \right\}}}_{J_2}$$

where the first inequality holds because $H \leq 4\widehat{\beta}_T \bar{\sigma}_t$ and the second one is by the Cauchy-Schwarz inequality. Here, we deduce from (Lemma 11, Abbasi-yadkori et al., 2011) that $J_2 \leq \sqrt{2d \log(1 + T/\lambda)}$. To get an upper bound on J_1, we show the following two lemmas.

Lemma 9. *Suppose that the event of Lemma 2 holds and that $\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left([\mathbb{P}W_k^2](s_t, a_t) - W_k^2(s_{t+1}) \right) \leq H^2 \sqrt{2T \log(1/\delta)}$. Then it holds that*

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} [\mathbb{V}W_k](s_t, a_t) = \widetilde{\mathcal{O}} \left(HT + H^2 d\sqrt{T} \right)$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in $T/(\delta\lambda)$.

Lemma 9 gives an upper bound on the variance term due to random transitions. We remark that the bound holds true even under our extended value iteration scheme with discounting and clipping.

Lemma 10. *Suppose that the event of Lemma 2 holds. Then it holds that*

$$\sum_{t=1}^{T} E_t = \widetilde{\mathcal{O}} \left(d^{3/2} H^2 \sqrt{T} \right)$$

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

Combining the above results, it holds that

$$I_4 = \widetilde{\mathcal{O}} \left(d\sqrt{HT} + H\sqrt{dT} + d^{7/4} HT^{1/4} \right)$$

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

6 REGRET LOWER BOUND

We deduce our regret lower bound based on the same hard-to-learn MDP instance due to Wu et al. (2022), illustrated in Figure 1. We derive the result based on the observation that for the MDP instance, the span of the underlying optimal bias function coincides with the diameter up to a constant multiplicative factor.

Remark 3. *In Wu et al. (2022), the authors establish a lower bound in terms of the diameter D for the linear mixture MDP setting and the linear MDP setting, noting that their approach can be adapted to yield a lower bound in terms of the span $sp(v^*)$. Here, we provide a direct proof for such a lower bound in terms of the span $sp(v^*)$.*

There are two states x_0 and x_1 as in Figure 1. The action space is given by $\mathcal{A} = \{-1, 1\}^{d-1}$, and the reward function is given by $r(x_0, a) = 0$ and $r(x_1, a) = 1$ for any $a \in \mathcal{A}$. We set the transition core $\bar{\theta}$ as

$$\bar{\theta} = \left(\frac{\theta}{\alpha}, \frac{1}{\beta} \right) \quad \text{where} \quad \theta \in \left\{ -\frac{\Delta}{d-1}, \frac{\Delta}{d-1} \right\}^{d-1},$$

with

$$\Delta = \frac{(d-1)}{45\sqrt{(2T \log 2)/(5\delta)}},$$

$\alpha = \sqrt{\Delta/((d-1)(1+\Delta))}$, and $\beta = \sqrt{1/(1+\Delta)}$. The feature vector is given by $\phi(x_0, a, x_0) = (-\alpha a, \beta(1-\delta))$, $\phi(x_0, a, x_1) = (\alpha a, \beta\delta)$, $\phi(x_1, a, x_0) = (0, \beta\delta)$, and $\phi(x_1, a, x_1) = (0, \beta(1-\delta))$.

As a higher stationary probability at state x_1 means a larger average reward, choosing the action a that satisfies $\langle a, \theta \rangle = \Delta$ is optimal. Hence, under the optimal policy, the stationary distribution is given by

$$(\mu^*(x_0), \mu^*(x_1)) = \left(\frac{\delta}{2\delta + \Delta}, \frac{\delta + \Delta}{2\delta + \Delta} \right),$$

and therefore, the optimal average reward is given by

$$J^* = \frac{\delta + \Delta}{2\delta + \Delta}.$$

Moreover, we may observe that

$$(v^*(x_0), v^*(x_1)) = \left(0, \frac{1}{2\delta + \Delta} \right),$$

$$(q^*(x_0, a), q^*(x_1, a)) = \left(\frac{\langle a, \theta \rangle - \Delta}{2\delta + \Delta}, \frac{1}{2\delta + \Delta} \right)$$
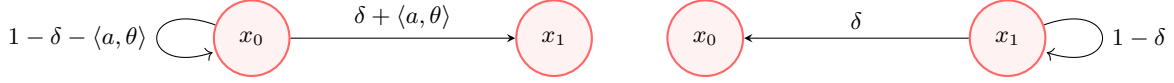
Figure 1: Illustration of the Hard-to-Learn Infinite-Horizon MDP Instance

satisfy the Bellman optimality condition (1). In particular, we have

$$\text{sp}(v^*) = \frac{1}{2\delta + \Delta}.$$

When $T \geq 16(d-1)^2\delta^{-1}/2025$, we have $\Delta \leq \delta/3$. Therefore, under this construction, we have

$$\frac{1}{3\delta} \leq \text{sp}(v^*) \leq \frac{1}{2\delta}.$$

For this MDP instance, based on (Theorem 5.5, Wu et al., 2022), we may deduce the following regret lower bound.

Theorem 2. *Suppose that $d \geq 2$ and $T \geq 16(d-1)^2\delta^{-1}/2025$. Then $\|\bar{\theta}\|_2 \leq 1 + \delta/3$ for any $\theta \in \{-\Delta/(d-1), \Delta/(d-1)\}^{d-1}$ and $\|\phi_F(x_0, a)\|_2, \|\phi_F(x_1, a)\|_2 \leq L$ for any $F : \mathcal{S} \to [0, L]$. Moreover, the span, $\text{sp}(v^*)$, of the optimal bias function for the MDP instance satisfies*

$$\frac{1}{3\delta} \leq \text{sp}(v^*) \leq \frac{1}{2\delta}.$$

Furthermore, for any algorithm, there exists $\theta \in \{-\Delta/(d-1), \Delta/(d-1)\}^{d-1}$ under which

$$\mathbb{E}\left[\text{Regret}(T)\right] \geq \frac{1}{2025}\sqrt{\frac{T}{\delta}} = \Omega\left(d\sqrt{\text{sp}(v^*)T}\right).$$

As pointed out in (Remark C.1, Wu et al., 2022), the regret lower bound in Theorem 2 also translates to a regret lower bound for learning infinite-horizon average-reward linear MDPs of bounded span.

7 CONCLUSION

This paper develops a provably efficient algorithm, `UCLK-C` for learning infinite-horizon average-reward linear mixture MDPs under the Bellman optimality condition. `UCLK-C` is the first algorithm that guarantees a nearly minimax optimal regret upper bound under the Bellman optimality condition. We establish this result based on our finding that discounted extended value iteration converges even with the additional clipping operation. We expect that this will be useful for infinite-horizon average-reward reinforcement under the Bellman optimality condition. We present some numerical results to test the computational performance of `UCLK-C` in the appendix.

Although we provide a nearly minimax optimal algorithm for linear mixture MDPs, there still exists a gap between the best-known regret upper and lower bounds for linear MDPs. To close the gap, one may attempt to extend the framework of this paper and other variance-aware parameter estimation schemes (e.g., He et al., 2023) to the linear MDP setting. We propose this as an open problem.

Acknowledgments

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. (2019). POLITEX: Regret bounds for policy iteration using expert prediction. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3692–3702. PMLR.

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. (2021). Reinforcement learning: Theory and algorithms.

Agarwal, A., Jin, Y., and Zhang, T. (2023). Voql: Towards optimal regret in model-free rl with nonlinear function approximation. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 987–1063. PMLR.

Agrawal, S. and Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Informa-*

tion Processing Systems, volume 30. Curran Associates, Inc.

Auer, P., Jaksch, T., and Ortner, R. (2008). Near-optimal regret bounds for reinforcement learning. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 463–474. PMLR.

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

Boone, V. and Zhang, Z. (2024). Achieving tractable minimax optimal regret in average reward mdps.

Bourel, H., Maillard, O., and Talebi, M. S. (2020). Tightening exploration in upper confidence reinforcement learning. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1056–1066. PMLR.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2020). Provably efficient exploration in policy optimization. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1283–1294. PMLR.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2018). On oracle-efficient pac rl with rich observations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. (2021). Bilinear classes: A structural framework for provable generalization in rl. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2826–2836. PMLR.

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

Fruit, R., Pirotta, M., and Lazaric, A. (2020). Improved analysis of ucrl2 with empirical bernstein inequality.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. (2018). Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1578–1586. PMLR.

He, J., Zhao, H., Zhou, D., and Gu, Q. (2023). Nearly minimax optimal reinforcement learning for linear Markov decision processes. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12790–12822. PMLR.

He, J., Zhong, H., and Yang, Z. (2024). Sample-efficient learning of infinite-horizon average-reward MDPs with general function approximation. In *The Twelfth International Conference on Learning Representations*.

He, J., Zhou, D., and Gu, Q. (2021). Logarithmic regret for reinforcement learning with linear function approximation. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4171–4180. PMLR.

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

Hong, K., Chae, W., Zhang, Y., Lee, D., and Tewari, A. (2025). Reinforcement learning for infinite-horizon average-reward linear mdps via approximation by discounted-reward mdps. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*. PMLR.

Hu, P., Chen, Y., and Huang, L. (2022). Nearly minimax optimal reinforcement learning with linear function approximation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8971–9019. PMLR.

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. (2020). Model-based reinforcement learning with value-targeted regression. In Bayen, A. M., Jadbabaie, A., Pappas, G., Parrilo, P. A., Recht, B., Tomlin, C., and Zeilinger, M., editors, *Proceedings of the 2nd Conference on Learning for Dynamics*

and Control, volume 120 of *Proceedings of Machine Learning Research*, pages 666–686. PMLR.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR.

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

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

Modi, A., Jiang, N., Tewari, A., and Singh, S. (2020). Sample complexity of reinforcement learning using linearly combined model ensembles. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2010–2020. PMLR.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. (2017a). Learning unknown markov decision processes: A thompson sampling approach. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. (2017b). Learning unknown markov decision processes: A thompson sampling approach. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

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

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. (2019). Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2898–2933. PMLR.

Talebi, M. S. and Maillard, O.-A. (2018). Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In Janoos, F., Mohri, M., and Sridharan, K., editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 770–805. PMLR.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. (2021). Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*.

Wei, C.-Y., Jafarnia Jahromi, M., Luo, H., and Jain, R. (2021). Learning infinite-horizon average-reward mdps with linear function approximation. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3007–3015. PMLR.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. (2020). Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10170–10180. PMLR.

Weisz, G., Amortila, P., and Szepesvári, C. (2021). Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In Feldman, V., Ligett, K., and Sabato, S., editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 1237–1264. PMLR.

Wu, Y., Zhou, D., and Gu, Q. (2022). Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., edi-

tors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3883–3913. PMLR.

Yang, L. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6995–7004. PMLR.

Yang, L. and Wang, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10746–10756. PMLR.

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

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. (2020a). Frequentist regret bounds for randomized least-squares value iteration. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1954–1964. PMLR.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. (2020b). Learning near optimal policies with low inherent Bellman error. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10978–10989. PMLR.

Zhang, Z. and Ji, X. (2019). Regret minimization for reinforcement learning by evaluating the optimal bias function. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhang, Z. and Xie, Q. (2023). Sharper model-free reinforcement learning for average-reward markov decision processes. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5476–5477. PMLR.

Zhou, D. and Gu, Q. (2022). Computationally efficient horizon-free reinforcement learning for linear mixture mdps. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36337–36349. Curran Associates, Inc.

Zhou, D., Gu, Q., and Szepesvari, C. (2021a). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In Belkin, M. and Kpotufe, S., editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4532–4576. PMLR.

Zhou, D., Gu, Q., and Szepesvari, C. (2021b). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes, *arXiv:2012.08507*.

Zhou, D., He, J., and Gu, Q. (2021c). Provably efficient reinforcement learning for discounted mdps with feature mapping. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12793–12802. PMLR.

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. [Yes]

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, in the appendix]

    (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). [Yes, in the supplementary material]

    (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

    (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). [Yes]

    (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

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   CONFIDENCE ELLIPSOIDS

In this section, we prove Lemma 2. We start by analyzing the error term

$$\left| [\bar{\mathbb{V}}_t W_k](s_t, a_t) - [\mathbb{V} W_k](s_t, a_t) \right|.$$

Note that

$$\begin{aligned}
&\left| [\bar{\mathbb{V}}_t W_k](s_t, a_t) - [\mathbb{V} W_k](s_t, a_t) \right| \\
&= \left| \left[\langle \phi_{W_k^2}(s_t, a_t), \widetilde{\theta}_t \rangle \right]_{[0, H^2]} - \langle \phi_{W_k^2}(s_t, a_t), \theta^* \rangle + \langle \phi_{W_k}(s_t, a_t), \theta^* \rangle^2 - \left[\langle \phi_{W_k}(s_t, a_t), \widehat{\theta}_t \rangle \right]^2_{[0, H]} \right| \\
&\leq \underbrace{\left| \left[\langle \phi_{W_k^2}(s_t, a_t), \widetilde{\theta}_t \rangle \right]_{[0, H^2]} - \langle \phi_{W_k^2}(s_t, a_t), \theta^* \rangle \right|}_{(a)} + \underbrace{\left| \langle \phi_{W_k}(s_t, a_t), \theta^* \rangle^2 - \left[\langle \phi_{W_k}(s_t, a_t), \widehat{\theta}_t \rangle \right]^2_{[0, H]} \right|}_{(b)}.
\end{aligned}$$

Let us consider term (a) first. Since $\langle \phi_{W_k^2}(s_t, a_t), \theta^* \rangle \in [0, H^2]$, it follows that

$$(a) \leq \min \left\{ H^2, \left| \langle \phi_{W_k^2}(s_t, a_t), \widetilde{\theta}_t \rangle - \langle \phi_{W_k^2}(s_t, a_t), \theta^* \rangle \right| \right\}.$$

Moreover, we may apply the Cauchy-Schwarz inequality to the right-hand side, and we obtain

$$(a) \leq \min \left\{ H^2, \left\| \phi_{W_k^2}(s_t, a_t) \right\|_{\widetilde{\Sigma}_t^{-1}} \left\| \widetilde{\theta}_t - \theta^* \right\|_{\widetilde{\Sigma}_t} \right\}.$$

Next we consider term (b). Again, since $\langle \phi_{W_k}(s_t, a_t), \theta^* \rangle \in [0, H]$, we have that $(b) \leq H^2$. Furthermore,

$$\begin{aligned}
(b) &= \left| \langle \phi_{W_k}(s_t, a_t), \theta^* \rangle + \left[\langle \phi_{W_k}(s_t, a_t), \widehat{\theta}_t \rangle \right]_{[0, H]} \right| \cdot \left| \langle \phi_{W_k}(s_t, a_t), \theta^* \rangle - \left[\langle \phi_{W_k}(s_t, a_t), \widehat{\theta}_t \rangle \right]_{[0, H]} \right| \\
&\leq 2H \left| \langle \phi_{W_k}(s_t, a_t), \theta^* \rangle - \left[\langle \phi_{W_k}(s_t, a_t), \widehat{\theta}_t \rangle \right]_{[0, H]} \right| \\
&\leq 2H \left| \langle \phi_{W_k}(s_t, a_t), \theta^* \rangle - \langle \phi_{W_k}(s_t, a_t), \widehat{\theta}_t \rangle \right| \\
&\leq 2H \left\| \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}} \left\| \widehat{\theta}_t - \theta^* \right\|_{\widehat{\Sigma}_t}
\end{aligned}$$

where the first and second inequalities hold because $\langle \phi_{W_k}(s_t, a_t), \theta^* \rangle \in [0, H]$ while the third inequality is due to the Cauchy-Schwarz inequality. Then we deduce that

$$(b) \leq \min \left\{ H^2, 2H \left\| \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}} \left\| \widehat{\theta}_t - \theta^* \right\|_{\widehat{\Sigma}_t} \right\}.$$

Therefore, we deduce that

$$\begin{aligned}
\left| [\bar{\mathbb{V}}_t W_k](s_t, a_t) - [\mathbb{V} W_k](s_t, a_t) \right| &\leq \min \left\{ H^2, \left\| \phi_{W_k^2}(s_t, a_t) \right\|_{\widetilde{\Sigma}_t^{-1}} \left\| \widetilde{\theta}_t - \theta^* \right\|_{\widetilde{\Sigma}_t} \right\} \\
&\quad + \min \left\{ H^2, 2H \left\| \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}} \left\| \widehat{\theta}_t - \theta^* \right\|_{\widehat{\Sigma}_t} \right\}.
\end{aligned} \tag{6}$$

Let us consider the following two confidence ellipsoids for θ^*:

$$\check{\mathcal{C}}_t = \left\{ \theta \in \mathcal{B} : \left\| \theta - \widehat{\theta}_t \right\|_{\widehat{\Sigma}_t} \leq \check{\beta}_t \right\}, \quad \widetilde{\mathcal{C}}_t = \left\{ \theta \in \mathcal{B} : \left\| \theta - \widetilde{\theta}_t \right\|_{\widetilde{\Sigma}_t} \leq \widetilde{\beta}_t \right\}$$

where

$$\begin{aligned}
\check{\beta}_t &= 8d \sqrt{\log(1 + t/\lambda) \log(4t^2/\delta)} + 4\sqrt{d} \log(4t^2/\delta) + \sqrt{\lambda} B_\theta, \\
\widetilde{\beta}_t &= 8H^2 \sqrt{d \log(1 + tH^2/(d\lambda)) \log(4t^2/\delta)} + 4H^2 \log(4t^2/\delta) + \sqrt{\lambda} B_\theta.
\end{aligned}$$

First, we apply Lemma 1 to the linear bandit instance defined with

$$(x_t, y_t) = (\bar{\sigma}_t^{-1}\phi_{W_k}(s_t, a_t), \bar{\sigma}_t^{-1}W_k(s_{t+1})), \quad \eta_t = y_t - \langle x_t, \theta^* \rangle \tag{7}$$

for $t \in [t_k : t_{k+1} - 1]$ and $k \in [K_T]$. Then we deduce that

$$\widehat{\Sigma}_{t+1} = Z_t := \lambda I_d + \sum_{i=1}^{t} x_i x_i^\top, \quad \widehat{b}_{t+1} = b_t := \sum_{i=1}^{t} y_i x_i, \quad \widehat{\theta}_{t+1} = \mu_t := Z_t^{-1} b_t.$$

As $\mathcal{G}_t = \sigma(s_1, \ldots, s_t)$ is the σ-algebra generated by $s_1, \ldots, s_t$, we have $\mathbb{E}[\eta_t \mid \mathcal{G}_t] = 0$. Moreover, we deduce that

$$|\eta_t| \leq \frac{\sqrt{d}}{H} |W_k(s_{t+1}) - \langle \phi_{W_k}(s_t, a_t), \theta^* \rangle| \leq \sqrt{d}$$

because $\bar{\sigma}_t^2 \geq H^2/d$ and it follows from $W_k(s) \in [0, H]$ for any $s \in \mathcal{S}$ that

$$-H \leq -\langle \phi_{W_k}(s_t, a_t), \theta^* \rangle \leq W_k(s_{t+1}) - \langle \phi_{W_k}(s_t, a_t), \theta^* \rangle \leq W_k(s_{t+1}) \leq H.$$

This also implies that $\mathbb{E}[\eta^2 \mid \mathcal{G}_t] \leq d$ and that

$$\|x_t\|_2 \leq \frac{\sqrt{d}}{H} \|\phi_{W_k}(s_t, a_t)\|_2 \leq \sqrt{d}.$$

Then it follows from Lemma 1 that with probability at least $1 - \delta$, $\theta^* \in \check{\mathcal{C}}_t$ for all $t \in [T]$.

Next, we apply again Lemma 1 to the linear bandit instance defined with

$$(x_t, y_t) = (\phi_{W_k^2}(s_t, a_t), W_k^2(s_{t+1})), \quad \eta_t = y_t - \langle x_t, \theta^* \rangle \tag{8}$$

for $t \in [t_k : t_{k+1} - 1]$ and $k \in [K_T]$. Then we deduce that

$$\widetilde{\Sigma}_{t+1} = Z_t := \lambda I_d + \sum_{i=1}^{t} x_i x_i^\top, \quad \widetilde{b}_{t+1} = b_t := \sum_{i=1}^{t} y_i x_i, \quad \widetilde{\theta}_{t+1} = \mu_t := Z_t^{-1} b_t.$$

As $\mathcal{G}_t = \sigma(s_1, \ldots, s_t)$ is the σ-algebra generated by $s_1, \ldots, s_t$, we have $\mathbb{E}[\eta_t \mid \mathcal{G}_t] = 0$. Since $W_k^2(s) \in [0, H^2]$ for any $s \in \mathcal{S}$, we have $|\eta_t| \leq H^2$ and thus $\mathbb{E}[\eta_t^2 \mid \mathcal{G}_t] \leq H^2$. Moreover, $\|x_t\|_2 = \|\phi_{W_k^2}(s_t, a_t)\|_2 \leq H^2$. Then, by Lemma 1, it hold with probability at least $1 - \delta$ that $\theta^* \in \widehat{\mathcal{C}}_t$ for all $t \in [T]$.

Lastly, we apply Lemma 1 to the linear bandit instance defined with

$$\begin{aligned}
x_t &= \bar{\sigma}_t^{-1}\phi_{W_k}(s_t, a_t), \\
\eta_t &= \bar{\sigma}_t^{-1}\mathbf{1}\left\{\theta^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\right\}(W_k(s_{t+1}) - \langle \phi_{W_k}(s_t, a_t), \theta^* \rangle), \\
y_t &= \eta_t + \langle x_t, \theta^* \rangle
\end{aligned} \tag{9}$$

where $\mathbf{1}\{\mathcal{E}\}$ denote the indicator function for a given event $\mathcal{E}$. Note that we still have $\mathbb{E}[\eta_t \mid \mathcal{G}_t] = 0$ where $\mathcal{G}_t = \sigma(s_1, \ldots, s_t)$ is the σ-algebra generated by $s_1, \ldots, s_t$ because $\mathbf{1}\left\{\theta^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\right\}$ is $\mathcal{G}_t$-measurable. Moreover, as $\mathbf{1}\{\cdot\} \leq 1$, we have $|\eta_t| \leq \sqrt{d}$ and $\|x_t\|_2 \leq \sqrt{d}$ as before. Let us consider $\mathbb{E}[\eta_t^2 \mid \mathcal{G}_t]$. We obtain that

$$\begin{aligned}
\mathbb{E}[\eta_t^2 \mid \mathcal{G}_t] &= \bar{\sigma}_t^{-2}\mathbf{1}\left\{\theta^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\right\}[\mathbb{V}W_k](s_t, a_t) \\
&\leq \bar{\sigma}_t^{-2}\mathbf{1}\left\{\theta^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\right\}\left([\bar{\mathbb{V}}_t W_k](s_t, a_t) + \min\left\{H^2, \left\|\phi_{W_k^2}(s_t, a_t)\right\|_{\widetilde{\Sigma}_t^{-1}}\left\|\widetilde{\theta}_t - \theta^*\right\|_{\widetilde{\Sigma}_t}\right\}\right) \\
&\quad + \bar{\sigma}_t^{-2}\mathbf{1}\left\{\theta^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\right\}\min\left\{H^2, 2H\|\phi_{W_k}(s_t, a_t)\|_{\widehat{\Sigma}_t^{-1}}\left\|\widehat{\theta}_t - \theta^*\right\|_{\widehat{\Sigma}_t}\right\} \\
&\leq \bar{\sigma}_t^{-2}\left([\bar{\mathbb{V}}_t W_k](s_t, a_t) + \min\left\{H^2, \widetilde{\beta}_t\left\|\phi_{W_k^2}(s_t, a_t)\right\|_{\widetilde{\Sigma}_t^{-1}}\right\} + \min\left\{H^2, 2H\check{\beta}_t\|\phi_{W_k}(s_t, a_t)\|_{\widehat{\Sigma}_t^{-1}}\right\}\right)
\end{aligned}$$

where the equality holds because $\mathbf{1}\left\{\theta^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\right\}$ is $\mathcal{G}_t$-measurable, the first inequality follows from (6), and the second inequality holds due to our construction of $\check{\mathcal{C}}_t$ and $\widetilde{\mathcal{C}}_t$. Hence, we deduce that

$$\mathbb{E}\left[\eta_t^2 \mid \mathcal{G}_t\right] \leq \bar{\sigma}_t^{-2}\left([\bar{\mathbb{V}}_t W_k](s_t, a_t) + E_t\right) \leq 1$$

where the first inequality is due to our choice of E_t and the second inequality holds because $\bar{\sigma}_t^2 \geq [\bar{\mathbb{V}}_t W_k](s_t, a_t) + E_t$.

Taking the union bound, with probability at least $1 - 3\delta$, the statement of Lemma 1 holds for each of the three linear bandit instances given by (7), (8), and (9). We denote by $\mathcal{E}_0$ this event. Note that under event $\mathcal{E}_0$, we have $\theta^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t$. In this case, we have that

$$|[\mathbb{V} W_k](s_t, a_t) - [\bar{\mathbb{V}}_t W_k](s_t, a_t)| \leq E_t.$$

Moreover, the outcome of Lemma 1 for the bandit instance (9) translates to the event that

$$\theta^* \in \widehat{\mathcal{C}}_t = \left\{\theta \in \mathcal{B} : \|\theta - \widehat{\theta}_t\|_{\widehat{\Sigma}_t} \leq \widehat{\beta}_t\right\}$$

where

$$\widehat{\beta}_t = 8\sqrt{d \log(1 + t/\lambda) \log(4t^2/\delta)} + 4\sqrt{d} \log(4t^2/\delta) + \sqrt{\lambda} B_\theta,$$

as required.

B CONVERGENCE OF VALUE ITERATION WITH CLIPPING

In this section, we provide the proofs of Lemmas 3 and 4.

B.1 Proof of Lemma 3: Clipping as a Contraction Map

Let $n \in [N]$, and let $s \in \mathcal{S}$. Then it is sufficient to argue that

$$V^{(n-1)}(s) - V^{(n)}(s) \leq \max_{s' \in \mathcal{S}}\left(\widetilde{V}^{(n-1)}(s') - \widetilde{V}^{(n)}(s')\right).$$

Recall that $V^{(i)}(s) = \min\{\widetilde{V}^{(i)}(s), \min_{s' \in \mathcal{S}} \widetilde{V}^{(i)}(s') + H\}$ for $i \in \{n-1, n\}$. Then there are two cases to consider depending on the value of $V^{(n)}(s)$.

Case I: $V^{(n)}(s) = \widetilde{V}^{(n)}(s)$. Note that $V^{(n-1)}(s) \leq \widetilde{V}^{(n-1)}(s)$. This in turn implies that

$$V^{(n-1)}(s) - V^{(n)}(s) \leq \widetilde{V}^{(n-1)}(s) - \widetilde{V}^{(n)}(s) \leq \max_{s' \in \mathcal{S}}\left(\widetilde{V}^{(n-1)}(s') - \widetilde{V}^{(n)}(s')\right).$$

Case II: $V^{(n)}(s) = \min_{s' \in \mathcal{S}} \widetilde{V}^{(n)}(s') + H$. Note that $V^{(n-1)}(s) \leq \min_{s' \in \mathcal{S}} \widetilde{V}^{(n-1)}(s') + H$. Then it follows that

$$V^{(n-1)}(s) - V^{(n)}(s) \leq \min_{s' \in \mathcal{S}} \widetilde{V}^{(n-1)}(s') + H - \min_{s' \in \mathcal{S}} \widetilde{V}^{(n)}(s') - H = \min_{s' \in \mathcal{S}} \widetilde{V}^{(n-1)}(s') - \min_{s' \in \mathcal{S}} \widetilde{V}^{(n)}(s').$$

Note that the right-most side satisfies

$$\begin{aligned}
&\min_{s' \in \mathcal{S}} \widetilde{V}^{(n-1)}(s') - \min_{s' \in \mathcal{S}} \widetilde{V}^{(n)}(s') \\
&= -\max_{s' \in \mathcal{S}}\left(-\widetilde{V}^{(n-1)}(s')\right) + \max_{s' \in \mathcal{S}}\left(-\widetilde{V}^{(n)}(s')\right) \\
&\leq -\max_{s' \in \mathcal{S}}\left(-\widetilde{V}^{(n-1)}(s')\right) + \max_{s' \in \mathcal{S}}\left(-\widetilde{V}^{(n-1)}(s')\right) + \max_{s' \in \mathcal{S}}\left(\widetilde{V}^{(n-1)}(s') - \widetilde{V}^{(n)}(s')\right) \\
&= \max_{s' \in \mathcal{S}}\left(\widetilde{V}^{(n-1)}(s') - \widetilde{V}^{(n)}(s')\right)
\end{aligned}$$

where the inequality holds because $\max_p\{f(p) + g(p)\} \leq \max_p\{f(p)\} + \max_p\{g(p)\}$, as required.

B.2 Proof of Lemma 4: Convergence of Discounted Extended Value Iteration with Clipping

We will first show the following lemma.

Lemma 11. *Let N be the number of rounds for discounted extended value iteration with clipping. Then for any episode k, it holds that $Q^{(N-1)}(s,a) - Q^{(N)}(s,a) \leq \gamma^{N-1}$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.*

Proof. We consider the discounted extended value iteration procedure with clipping for a fixed episode k. Note that for $n \geq 2$, we have

$$Q^{(n)}(s,a) = r(s,a) + \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(n-1)}}(s,a), \theta \rangle,$$

$$Q^{(n-1)}(s,a) = r(s,a) + \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(n-2)}}(s,a), \theta \rangle.$$

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

$$
\begin{aligned}
Q^{(n-1)}(s,a) - Q^{(n)}(s,a) &= \gamma \left(\max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(n-2)}}(s,a), \theta \rangle - \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(n-1)}}(s,a), \theta \rangle \right) \\
&\leq \gamma \max_{\theta \in \mathcal{C}_k} \left(\langle \phi_{V^{(n-2)}}(s,a), \theta \rangle - \langle \phi_{V^{(n-1)}}(s,a), \theta \rangle \right) \\
&= \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(n-2)} - V^{(n-1)}}(s,a), \theta \rangle
\end{aligned}
\tag{10}
$$

where the inequality holds because $\max_p \{f(p) + g(p)\} \leq \max_p \{f(p)\} + \max_p \{g(p)\}$. Recall that for any $\theta \in \mathcal{C}_k$ induces a probability distribution with $\phi(s,a,s')$ given by $\mathbb{P}_\theta(s' \mid s,a) = \langle \phi(s,a,s'), \theta \rangle$. Then it follows that

$$
\begin{aligned}
\gamma \langle \phi_{V^{(n-2)} - V^{(n-1)}}(s,a), \theta \rangle &= \gamma \mathbb{E}_{s' \sim \mathbb{P}_\theta(\cdot \mid s,a)} \left[V^{(n-2)}(s') - V^{(n-1)}(s') \right] \\
&\leq \gamma \max_{s' \in \mathcal{S}} \left(V^{(n-2)}(s') - V^{(n-1)}(s') \right) \\
&\leq \gamma \max_{s' \in \mathcal{S}} \left(\widetilde{V}^{(n-2)}(s') - \widetilde{V}^{(n-1)}(s') \right)
\end{aligned}
\tag{11}
$$

where the second inequality is due to Lemma 3. Combining (10) and (11), we have

$$
\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(Q^{(n-1)}(s,a) - Q^{(n)}(s,a) \right) \leq \gamma \max_{s \in \mathcal{S}} \left(\widetilde{V}^{(n-2)}(s) - \widetilde{V}^{(n-1)}(s) \right).
\tag{12}
$$

Here, the right-hand side of (12) can be further bounded as follows.

$$
\begin{aligned}
\gamma \max_{s \in \mathcal{S}} \left(\widetilde{V}^{(n-2)}(s) - \widetilde{V}^{(n-1)}(s) \right) &= \gamma \max_{s \in \mathcal{S}} \left(\max_{a \in \mathcal{A}} Q^{(n-2)}(s,a) - \max_{a \in \mathcal{A}} Q^{(n-1)}(s,a) \right) \\
&\leq \gamma \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(Q^{(n-2)}(s,a) - Q^{(n-1)}(s,a) \right)
\end{aligned}
\tag{13}
$$

where the inequality is due to $\max_{a'} \{f(a') + g(a')\} \leq \max_{a'} \{f(a')\} + \max_{a'} \{g(a')\}$ as before. Therefore, it follows that for any $n \geq 2$,

$$
\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(Q^{(n-1)}(s,a) - Q^{(n)}(s,a) \right) \leq \gamma \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(Q^{(n-2)}(s,a) - Q^{(n-1)}(s,a) \right).
$$

In particular, this implies that

$$
\begin{aligned}
\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(Q^{(N-1)}(s,a) - Q^{(N)}(s,a) \right) &\leq \gamma^{N-1} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(Q^{(0)}(s,a) - Q^{(1)}(s,a) \right) \\
&= \gamma^{N-1} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(\frac{1}{1-\gamma} - r(s,a) - \frac{\gamma}{1-\gamma} \right) \\
&\leq \gamma^{N-1}
\end{aligned}
$$

where the last inequality holds because $0 \leq r(s,a) \leq 1$. $\qquad \square$

Based on Lemma 11, we complete the proof of Lemma 4. Note that

$$
\begin{aligned}
Q^{(N)}(s_t, a_t) &= r(s_t, a_t) + \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(N-1)}}(s_t, a_t), \theta \rangle \\
&\leq r(s_t, a_t) + \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(N)}}(s_t, a_t), \theta \rangle + \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(N-1)} - V^{(N)}}(s_t, a_t), \theta \rangle \\
&\leq r(s_t, a_t) + \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(N)}}(s_t, a_t), \theta \rangle + \gamma \max_{s \in \mathcal{S}} \left(\widetilde{V}^{(N-1)}(s) - \widetilde{V}^{(N)}(s) \right) \\
&\leq r(s_t, a_t) + \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(N)}}(s_t, a_t), \theta \rangle + \gamma \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(Q^{(N-1)}(s, a) - Q^{(N)}(s, a) \right) \\
&\leq r(s_t, a_t) + \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(N)}}(s_t, a_t), \theta \rangle + \gamma^N
\end{aligned}
$$

where the first inequality applies the same argument as in (11), the second inequality follows the same argument as in (13), and the third inequality is due to Lemma 11. Since $Q^{(N)}$ equals Q_k and $V^{(N)}$ equals V_k, we have

$$
Q_k(s_t, a_t) \leq r(s_t, a_t) + \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V_k}(s_t, a_t), \theta \rangle + \gamma^N,
$$

as required.

C   REGRET ANALYSIS AND PROOFS

In this section, we prove Lemma 5 and Lemmas 7 to 10. Based on these results, in Appendix C.6, we complete the proof of Theorem 1.

C.1   Proof of Lemma 5: Optimistic Estimators for Value Functions

For a fixed episode k, we prove by induction on n that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$
\frac{1}{1 - \gamma} \geq V^{(n)}(s) \geq V^*(s), \quad \frac{1}{1 - \gamma} \geq Q^{(n)}(s, a) \geq Q^*(s, a).
$$

by induction on n. For $n = 0$, it is trivial that

$$
V^{(0)} = \frac{1}{1 - \gamma} \geq V^*(s), \quad Q^{(0)}(s, a) = \frac{1}{1 - \gamma} \geq Q^*(s, a)
$$

for every $(s, a) \in \mathcal{S} \times \mathcal{A}$. Next, we assume that for some $n \geq 0$, the inequalities

$$
\frac{1}{1 - \gamma} \geq V^{(n)}(s) \geq V^*(s) \quad \text{and} \quad \frac{1}{1 - \gamma} \geq Q^{(n)}(s, a) \geq Q^*(s, a)
$$

hold for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. First of all, since any $\theta \in \mathcal{C}_k$ induces a probability distribution, we get

$$
Q^{(n+1)}(s, a) = r(s, a) + \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(n)}}(s, a), \theta \rangle \leq 1 + \frac{\gamma}{1 - \gamma} = \frac{1}{1 - \gamma}
$$

because $r(s, a) \leq 1$ and $V^{(n)}(s') \leq (1 - \gamma)^{-1}$ for any $s' \in \mathcal{S}$. Then

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

Next, we show that $Q^{(n+1)}(s, a) \geq Q^*(s, a)$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. Note that

$$
\begin{aligned}
Q^{(n+1)}(s, a) &= r(s, a) + \gamma \max_{\theta \in \mathcal{C}_k} \langle \phi_{V^{(n)}}(s, a), \theta \rangle \\
&\geq r(s, a) + \gamma \langle \phi_{V^{(n)}}(s, a), \theta^* \rangle \\
&= r(s, a) + \gamma [\mathbb{P} V^{(n)}](s, a) \\
&\geq r(s, a) + \gamma [\mathbb{P} V^*](s, a) \\
&= Q^*(s, a)
\end{aligned}
$$

where the first inequality holds because $\theta^* \in \mathcal{C}_k = \widehat{\mathcal{C}}_{t_k}$, the second inequality is by the induction hypothesis, and the last equality is by the Bellman optimality condition (2).

Let us also consider $V^{(n+1)}$. Note that

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

where the inequality holds because $Q^{(n+1)}(s,a) \geq Q^*(s,a)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. This in turn implies that $\widetilde{V}^{(n+1)}(s) \geq V^*(s)$ for any $s \in \mathcal{S}$. This further implies that

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

where the first inequality comes from our observation that $\widetilde{V}^{(n+1)}(s) \geq V^*(s)$ for any $s \in \mathcal{S}$ while the second equality holds because $\mathrm{sp}(V^*) \leq 2 \cdot \mathrm{sp}(v^*) \leq H$, as supported by Lemma 6.

Since k was chosen arbitrarily, we conclude that in every episode, for all $n \in [N]$ and for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

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

as required.

C.2   Proof of Lemma 7: Upper Bound on the Number of Episodes

Note that $\det(\widehat{\Sigma}_1) = \lambda^d$ because $\widehat{\Sigma}_1 = \lambda I_d$. To upper bound $\det(\widehat{\Sigma}_{T+1})$, we apply the following lemma.

Lemma 12. (Lemma 10, Abbasi-yadkori et al., 2011) *For any $x_1, \ldots, x_T \in \mathbb{R}^d$ such that $\|x_t\|_2 \leq L$, let $A_1 = \lambda I_d$ and $A_{t+1} = \lambda I_d + \sum_{i=1}^t x_i x_i^\top$ for $t \geq 1$. Then*

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

Recall that

$$
\widehat{\Sigma}_{T+1} = \lambda I_d + \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \phi_{W_k}(s_t, a_t) \phi_{W_k}(s_t, a_t)^\top.
$$

Here, we have $\|\phi_{W_k}(s_t, a_t)\|_2 \leq H$ as $W_k(s) \in [0, H]$ for any $s \in \mathcal{S}$. By applying Lemma 12, we get

$$
\det(\widehat{\Sigma}_{T+1}) \leq \left(\lambda + \frac{TH^2}{d}\right)^d.
$$

As $\lambda = 1/B_\theta^2$, it follows that

$$
\frac{\det(\widehat{\Sigma}_{T+1})}{\det(\widehat{\Sigma}_1)} \leq \left(1 + \frac{TH^2}{d\lambda}\right)^d = \left(1 + \frac{TH^2 B_\theta^2}{d}\right)^d.
$$

Moreover, note that

$$
\det(\widehat{\Sigma}_{T+1}) \geq \det(\widehat{\Sigma}_T) \geq \det(\widehat{\Sigma}_{t_{K_T}}) \geq \cdots \geq 2^{K_T-1} \det(\widehat{\Sigma}_{t_1}),
$$

implying in turn that

$$
K_T \leq 1 + \log_2(\det(\widehat{\Sigma}_{T+1})/\det(\widehat{\Sigma}_{t_1})) \leq 1 + d\log_2(1 + TH^2 B_\theta^2/d),
$$

as required.

C.3 Proof of Lemma 8: Martingale Difference Sequence

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

$$\eta_t = \langle \phi_{V_k}(s_t, a_t), \theta^* \rangle - V_k(s_{t+1})$$

and $\mathcal{G}_t = \sigma(s_1, \ldots, s_t)$ is the σ-algebra generated by $s_1, \ldots, s_t$. for $t \in [t_k : t_{k+1} - 1]$. This is because η_t is $\mathcal{G}_{t+1}$-measurable, $\mathbb{E}[|\eta_t|] < \infty$, and $\mathbb{E}[\eta_t | \mathcal{G}_t] = 0$, which we will show in the following paragraphs. In fact, we have

$$\eta_t = \langle \phi_{V_k}(s_t, a_t), \theta^* \rangle - V_k(s_{t+1}) = \langle \phi_{W_k}(s_t, a_t), \theta^* \rangle - W_k(s_{t+1}),$$

which implies that

$$|\eta_t| \leq \mathrm{sp}(W_k) \leq H.$$

Lemma 13. (Azuma-Hoeffding inequality) *Let $\{X_k\}_{k=0}^{\infty}$ be a discrete-parameter real-valued martingale sequence such that for every $k \in \mathbb{N}$, the condition $|X_k - X_{k-1}| \leq \mu$ holds for some non-negative constant μ. Then with probability at least $1 - \delta$, we have*

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

Since $X_t = \sum_{n=1}^{t} \eta_t$ for $t \geq 1$ and X_0 give rise to a martingale sequence with $|\eta_t| \leq H$, it follows from Lemma 13 that

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

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

C.4 Proof of Lemma 9: Variance Term

Note that

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} [\mathbb{V}W_k](s_t, a_t)$$

$$= \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left([\mathbb{P}W_k^2](s_t, a_t) - ([\mathbb{P}W_k](s_t, a_t))^2\right)$$

$$= \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left([\mathbb{P}W_k^2](s_t, a_t) - W_k^2(s_{t+1})\right) + \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(W_k^2(s_{t+1}) - W_k^2(s_t)\right)$$

$$+ \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(W_k^2(s_t) - ([\mathbb{P}W_k](s_t, a_t))^2\right)$$

$$= \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left([\mathbb{P}W_k^2](s_t, a_t) - W_k^2(s_{t+1})\right) + \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(W_k^2(s_t) - ([\mathbb{P}W_k](s_t, a_t))^2\right)$$

$$+ \sum_{k=1}^{K_T} \left(W_k^2(s_{t_{k+1}}) - W_k^2(s_{t_k})\right)$$

$$\leq \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left([\mathbb{P}W_k^2](s_t, a_t) - W_k^2(s_{t+1})\right)}_{(i)} + \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(W_k^2(s_t) - ([\mathbb{P}W_k](s_t, a_t))^2\right)}_{(ii)} + H^2 K_T.$$

By the condition of the lemma, we have

$$(i) \leq H^2\sqrt{2T\log(1/\delta)}.$$

Let us consider the term (ii). Note that

$$(ii) = \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(W_k(s_t) - [\mathbb{P}W_k](s_t, a_t)\right)\left(W_k(s_t) + [\mathbb{P}W_k](s_t, a_t)\right).$$

Here, we have

$$
\begin{aligned}
W_k(s_t) &= V_k(s_t) - \min_{s' \in \mathcal{S}} V_k(s') \\
&\leq \widetilde{V}_k(s_t) - \min_{s' \in \mathcal{S}} V_k(s') \\
&= Q_k(s_t, a_t) - \min_{s' \in \mathcal{S}} V_k(s') \\
&\leq 1 + \gamma^N + \max_{\theta \in \mathcal{C}_k} \langle \phi_{V_k}(s_t, a_t), \theta \rangle - \min_{s' \in \mathcal{S}} V_k(s') \\
&\leq 2 + \max_{\theta \in \mathcal{C}_k} \langle \phi_{W_k}(s_t, a_t), \theta \rangle
\end{aligned}
$$

where the second inequality holds because $\gamma \leq 1$ and $r(s_t, a_t) \leq 1$ and the third inequality holds because $\gamma \leq 1$ and any $\theta \in \mathcal{C}_k$ induces a probability distribution. Then it follows that

$$
\begin{aligned}
(ii) &\leq 2 \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(W_k(s_t) + [\mathbb{P}W_k](s_t, a_t) \right) + \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left(\max_{\theta \in \mathcal{C}_k} \langle \phi_{W_k}(s_t, a_t), \theta - \theta^* \rangle \right) \left(W_k(s_t) + [\mathbb{P}W_k](s_t, a_t) \right) \\
&\leq 4HT + 2H \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left| \max_{\theta \in \mathcal{C}_k} \langle \phi_{W_k}(s_t, a_t), \theta - \theta^* \rangle \right|.
\end{aligned}
$$

Note that

$$
\left| \max_{\theta \in \mathcal{C}_k} \langle \phi_{W_k}(s_t, a_t), \theta - \theta^* \rangle \right| \leq \max_{\theta \in \mathcal{C}_k} | \langle \phi_{W_k}(s_t, a_t), \theta - \theta^* \rangle | \leq \max_{\theta \in \mathcal{C}_k} \| \phi_{W_k}(s_t, a_t) \|_{\widehat{\Sigma}_t^{-1}} \| \theta - \theta^* \|_{\widehat{\Sigma}_t}.
$$

To bound the right-most side, we need the following lemma.

Lemma 14. (Abbasi-yadkori et al., 2011, Lemma 12) *Let $A, B \in \mathbb{R}^{d \times d}$ be positive semidefinite matrices such that $A \succeq B$. Then for any $x \in \mathbb{R}^d$, we have $\|x\|_A \leq \|x\|_B \sqrt{\det(A)/\det(B)}$.*

By Lemma 14, we have $\|\theta - \theta^*\|_{\widehat{\Sigma}_t} \leq 2\|\theta - \theta^*\|_{\widehat{\Sigma}_{t_k}}$. Therefore,

$$
\left| \max_{\theta \in \mathcal{C}_k} \langle \phi_{W_k}(s_t, a_t), \theta - \theta^* \rangle \right| \leq \sqrt{2} \max_{\theta \in \mathcal{C}_k} \| \phi_{W_k}(s_t, a_t) \|_{\widehat{\Sigma}_t^{-1}} \left(\| \theta - \widehat{\theta}_{t_k} \|_{\widehat{\Sigma}_{t_k}} + \| \widehat{\theta}_{t_k} - \theta^* \|_{\widehat{\Sigma}_{t_k}} \right) \leq 2\sqrt{2} \widehat{\beta}_T \| \phi_{W_k}(s_t, a_t) \|_{\widehat{\Sigma}_t^{-1}}
$$

where the second inequality holds because $\theta, \theta^* \in \mathcal{C}_k$. Meanwhile, we already know that $|\langle \phi_{W_k}(s_t, a_t), \theta - \theta^* \rangle| \leq H$ for any θ that induces a probability distribution. Then it follows that

$$
\begin{aligned}
\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left| \max_{\theta \in \mathcal{C}_k} \langle \phi_{W_k}(s_t, a_t), \theta - \theta^* \rangle \right| &\leq \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ H, 2\sqrt{2} \widehat{\beta}_T \| \phi_{W_k}(s_t, a_t) \|_{\widehat{\Sigma}_t^{-1}} \right\} \\
&\leq \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} 2\sqrt{2} \widehat{\beta}_T \bar{\sigma}_t \min \left\{ 1, \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}} \right\} \\
&\leq 4\sqrt{2} H \widehat{\beta}_T \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ 1, \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}} \right\} \\
&\leq 4\sqrt{2} H \widehat{\beta}_T \sqrt{T} \sqrt{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ 1, \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}}^2 \right\}}
\end{aligned}
$$

where the second inequality is from $H \leq 4\widehat{\beta}_T \bar{\sigma}_t$, the third inequality holds because we have $[\bar{\mathbb{V}}_t W_k](s_t, a_t) \leq H^2$ and $E_t \leq 2H^2$, which implies that $\bar{\sigma}_t \leq \sqrt{\max\{H^2/d, 3H^2\}} \leq 2H$, and the last inequality is implied by the Cauchy-Schwarz inequality. To bound the right-most side, we need the following lemma.

Lemma 15. (Lemma 11, Abbasi-yadkori et al., 2011). *Suppose $x_1, \ldots, x_t \in \mathbb{R}^d$ and $\|x_s\|_2 \leq L$ for any $1 \leq s \leq t$. Let $V_t = \lambda I_d + \sum_{i=1}^{t} x_i x_i^\top$ for some $\lambda > 0$. Then*

$$
\sum_{i=1}^{t} \min \left\{ 1, \|x_i\|_{V_{i-1}^{-1}}^2 \right\} \leq 2d \log \left(1 + \frac{tL^2}{d\lambda} \right).
$$

Since $\left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_2 \leq \sqrt{d}$, Lemma 15 implies that

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ 1, \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}}^2 \right\} \leq 2d \log(1 + T/\lambda), \tag{14}$$

implying in turn that

$$(ii) \leq 4HT + 16H^2 \widehat{\beta}_T \sqrt{dT \log(1 + T/\lambda)}.$$

Consequently,

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} [\mathbb{V}W_k](s_t, a_t) \leq (i) + (ii) + H^2 K_T \leq 4HT + 16H^2 \widehat{\beta}_T \sqrt{dT \log(1 + T/\lambda)} + H^2 K_T.$$

Then it follows from our choice of $\widehat{\beta}_T$ and Lemma 7 that

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} [\mathbb{V}W_k](s_t, a_t) = \widetilde{\mathcal{O}} \left(HT + H^2 d \sqrt{T} \right)$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in $T/(\delta\lambda)$.

C.5 Proof of Lemma 10: Cumulative Error in Estimating the Variance

Note that

$$\sum_{t=1}^{T} E_t = \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ H^2, 2H\check{\beta}_t \| \phi_{W_k}(s_t, a_t) \|_{\widehat{\Sigma}_t^{-1}} \right\}}_{(I)} + \underbrace{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ H^2, \widetilde{\beta}_t \| \phi_{W_k^2}(s_t, a_t) \|_{\widetilde{\Sigma}_t^{-1}} \right\}}_{(II)}.$$

Term (I) can be bounded as follows:

$$(I) \leq \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} 2H\check{\beta}_t \bar{\sigma}_t \min \left\{ 1, \| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \|_{\widehat{\Sigma}_t^{-1}} \right\}$$

$$\leq 4H^2 \check{\beta}_T \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ 1, \| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \|_{\widehat{\Sigma}_t^{-1}} \right\}$$

$$\leq 4H^2 \check{\beta}_T \sqrt{T} \sqrt{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ 1, \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}}^2 \right\}}$$

$$\leq 8H^2 \check{\beta}_T \sqrt{dT \log(1 + T/\lambda)}$$

where the first inequality is due to $H \leq 2\check{\beta}_t \bar{\sigma}_t$, the second inequality holds because we have $[\bar{\mathbb{V}}_t W_k](s_t, a_t) \leq H^2$ and $E_t \leq 2H^2$, which implies that $\bar{\sigma}_t \leq \sqrt{\max\{H^2/d, 3H^2\}} \leq 2H$, the third inequality is due to the Cauchy-Schwarz inequality, and the last one follows from (14).

Term (II) can be bounded as follows:

$$(II) \leq \widetilde{\beta}_T \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ 1, \| \phi_{W_k^2}(s_t, a_t) \|_{\widetilde{\Sigma}_t^{-1}} \right\}$$

$$\leq \widetilde{\beta}_T \sqrt{T} \sqrt{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ 1, \| \phi_{W_k^2}(s_t, a_t) \|_{\widetilde{\Sigma}_t^{-1}}^2 \right\}}$$

$$\leq 2\widetilde{\beta}_T \sqrt{dT \log(1 + TH^2/\lambda)}$$

where the first inequality holds because $H^2 \leq \breve{\beta}_t \leq \widetilde{\beta}_T$, the second inequality is by the Cauchy-Schwarz inequality, and the third one follows from Lemma 15.

Therefore, it holds that

$$\sum_{t=1}^{T} E_t \leq 8H^2 \breve{\beta}_T \sqrt{dT \log(1 + T/\lambda)} + 2\widetilde{\beta}_T \sqrt{dT \log(1 + TH^2/\lambda)}.$$

Due to our choice of $\breve{\beta}_T$ and $\widetilde{\beta}_T$, we have

$$\sum_{t=1}^{T} E_t = \widetilde{\mathcal{O}}\left(d^{3/2} H^2 \sqrt{T}\right)$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in TH/λ, as required.

C.6 Completing the Proof of Theorem 1: Regret Upper Bound of `UCLK-C`

We first provide an upper bound on I_4. Let $\theta \in \mathcal{C}_k$. Since θ induces a probability distribution $\mathbb{P}_\theta$, we have

$$\begin{aligned}
\langle \phi_{V_k}(s_t, a_t), \theta - \theta^* \rangle &= \mathbb{E}_{s' \sim \mathbb{P}_\theta(\cdot|s_t, a_t)}[V_k(s')] - \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t, a_t)}[V_k(s')] \\
&= \mathbb{E}_{s' \sim \mathbb{P}_\theta(\cdot|s_t, a_t)}[V_k(s')] - \min_{s' \in \mathcal{S}} V_k(s') - \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t, a_t)}[V_k(s')] + \min_{s' \in \mathcal{S}} V_k(s') \\
&= \mathbb{E}_{s' \sim \mathbb{P}_\theta(\cdot|s_t, a_t)}[W_k(s')] - \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t, a_t)}[W_k(s')] \\
&= \langle \phi_{W_k}(s_t, a_t), \theta - \theta^* \rangle.
\end{aligned}$$

Moreover, assuming that $\theta^* \in \mathcal{C}_k$ based on Lemma 2, we have

$$\begin{aligned}
\langle \phi_{W_k}(s_t, a_t), \theta - \theta^* \rangle &\leq \|\phi_{W_k}(s_t, a_t)\|_{\widehat{\Sigma}_t^{-1}} \left(\|\theta - \widehat{\theta}_{t_k}\|_{\widehat{\Sigma}_t} + \|\widehat{\theta}_{t_k} - \theta^*\|_{\widehat{\Sigma}_t} \right) \\
&\leq 2 \|\phi_{W_k}(s_t, a_t)\|_{\widehat{\Sigma}_t^{-1}} \left(\|\theta - \widehat{\theta}_{t_k}\|_{\widehat{\Sigma}_{t_k}} + \|\widehat{\theta}_{t_k} - \theta^*\|_{\widehat{\Sigma}_{t_k}} \right) \\
&\leq 4\widehat{\beta}_T \bar{\sigma}_t \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}}
\end{aligned}$$

where the first inequality is from the Cauchy-Schwarz inequality, the second one holds by Lemma 14, and the third one follows from $\theta^*, \theta \in \mathcal{C}_k$ and $\widehat{\beta}_{t_k} \leq \widehat{\beta}_T$. We also know that

$$\langle \phi_{W_k}(s_t, a_t), \theta - \theta^* \rangle = \mathbb{E}_{s' \sim \mathbb{P}_\theta(\cdot|s_t, a_t)}[W_k(s')] - \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t, a_t)}[W_k(s')] \in [-H, H]$$

because $W_k(s) \in [0, H]$ for any $s \in \mathcal{S}$. Then it follows that

$$\begin{aligned}
I_4 &\leq \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ H, 4\widehat{\beta}_T \bar{\sigma}_t \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}} \right\} \\
&\leq 4\widehat{\beta}_T \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \bar{\sigma}_t \min \left\{ 1, \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}} \right\} \\
&\leq 4\widehat{\beta}_T \underbrace{\sqrt{\sum_{t=1}^{T} \bar{\sigma}_t^2}}_{J_1} \underbrace{\sqrt{\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \min \left\{ 1, \left\| \bar{\sigma}_t^{-1} \phi_{W_k}(s_t, a_t) \right\|_{\widehat{\Sigma}_t^{-1}}^2 \right\}}}_{J_2}
\end{aligned}$$

where the second inequality holds because $H \leq 4\beta_T \bar{\sigma}_t$ and the third one is by the Cauchy-Schwarz inequality.

Note that by the Azuma-Hoeffding inequality (Lemma 13),

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \left([\mathbb{P}W_k^2](s_t, a_t) - W_k^2(s_{t+1}) \right) \leq H^2 \sqrt{2T \log(1/\delta)} \tag{15}$$

holds with probability at least $1 - \delta$. By taking the union bound, all of the statement of Lemma 2, the statement of Lemma 8, and (15) hold with probability at least $1 - 5\delta$.

Now we suppose that all of the statement of Lemma 2, the statement of Lemma 8, and (15) hold. For the term J_1, note that

$$
\begin{aligned}
\sum_{t=1}^{T} \bar{\sigma}_t^2 &= \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \max\left\{ H^2/d, [\bar{\mathbb{V}}_t W_k](s_t, a_t) + E_t \right\} \\
&\leq \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \max\left\{ H^2/d, [\mathbb{V} W_k](s_t, a_t) + 2E_t \right\} \\
&\leq TH^2/d + \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} [\mathbb{V} W_k](s_t, a_t) + 2\sum_{t=1}^{T} E_t
\end{aligned}
$$

where the first inequality is from Lemma 2 and the second inequality holds because both H^2/d and $[\mathbb{V} W_k](s_t, a_t) + 2E_t$ are nonnegative. Then we obtain from Lemmas 9 and 10 that

$$
\sum_{t=1}^{T} \bar{\sigma}_t^2 = \widetilde{\mathcal{O}}\left(H^2 T/d + HT + H^2 d\sqrt{T} + d^{3/2} H^2 \sqrt{T} \right)
$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in $TH/(\delta\lambda)$. Moreover, we know that $J_2 \leq \sqrt{2d \log(1 + T/\lambda)}$ by (14). Therefore, we finally deduce that

$$
I_4 = \widetilde{\mathcal{O}}\left(\sqrt{d} \cdot \sqrt{H^2 T/d + HT + H^2 d\sqrt{T} + d^{3/2} H^2 \sqrt{T}} \cdot \sqrt{d} \right) = \widetilde{\mathcal{O}}\left(H\sqrt{dT} + d\sqrt{HT} + d^{7/4} HT^{1/4} \right)
$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in $TH/(\delta\lambda)$.

Recall that when $\lambda = 1/B_\theta^2$ we have

$$
\begin{aligned}
I_1 &\leq d\sqrt{\mathrm{sp}(v^*)T} = d\sqrt{HT}, \\
I_2 &= \widetilde{\mathcal{O}}\left(d\sqrt{HT} \right), \\
I_3 &\leq H\sqrt{2d \log(1/\delta)}.
\end{aligned}
$$

Lastly, setting γ and N as

$$
\gamma = 1 - \sqrt{\frac{d}{HT}} \quad \text{and} \quad N = \frac{1}{1-\gamma} \log\left(\frac{\sqrt{T}}{d\sqrt{H}} \right) = \sqrt{\frac{HT}{d}} \log\left(\frac{\sqrt{T}}{d\sqrt{H}} \right),
$$

we have

$$
N \geq \frac{\log\left(\sqrt{T}/d\sqrt{H} \right)}{\log(1/\gamma)},
$$

in which case we get $T\gamma^N \leq d\sqrt{HT}$.

Therefore, we conclude that with probability at least $1 - 5\delta$

$$
\mathrm{Regret}(T) = \widetilde{\mathcal{O}}\left(H\sqrt{dT} + d\sqrt{HT} + d^{7/4} HT^{1/4} \right)
$$

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

D Experiments

In this section, we numerically compare the performance of `UCLK-C` with `UCRL2-VTR` (Bernstein-type exploration, Wu et al. (2022)), `TSDE` (Ouyang et al., 2017b), and `UCRL2` (Auer et al., 2008), given random policy as a baseline.

We conduct experiments with the MDP instance introduced in Section 6. Given that `UCLK-C` and `UCRL2-VTR` are minimax optimal on this MDP instance, our comparison provides an intuitive understanding of how controlling the span of the value function can improve regret performance. In addition, it enables the comparison of RL algorithms with linear function approximation (`UCLK-C`, `UCRL2-VTR`) and tabular RL algorithms (`TSDE`, `UCRL2`), when deployed on the linearly structured MDP. The next paragraph details the experimental setup and results.
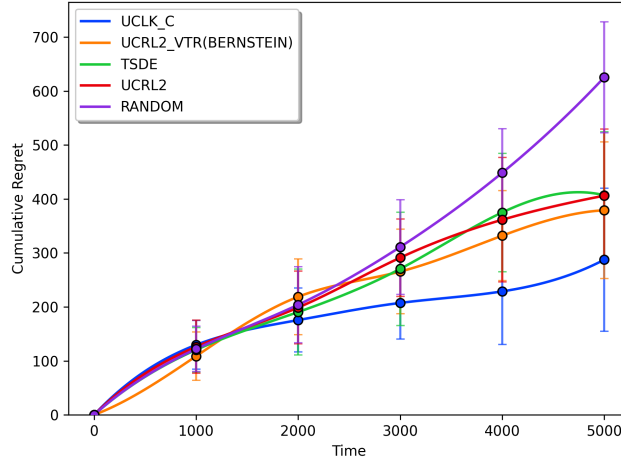

Figure 2: Regret comparison of `UCLK-C`, `UCRL2-VTR` (Bernstein-type), `TSDE`, `UCRL2`, and random policy

We choose $d = 8$ and $\delta = 1/120$, thereby the resulting MDP consists of two states ($|\mathcal{S}| = 2$) and 128 actions ($|\mathcal{A}| = 2^{d-1} = 128$) with diameter 120 ($D = 1/\delta = 120$). Under this setting, we compared the algorithms based on the average regret with respect to the best action, evaluated over 40 realizations for each time horizon in units of thousands. The average regret and its deviation are plotted in Figure 2.

The numerical experiment supports the theoretical result that our proposed algorithm `UCLK-C`, outperforms `UCRL2-VTR`, demonstrating the effectiveness of controlling the span for improving regret performance. In addition, `UCLK-C` outperforms the other tabular RL algorithms, `TSDE` and `UCRL2`, due to the linear structure of the MDP, which we expect may be even clearer in a larger state, action space.