# Nearly Minimax Optimal Regret for Learning Infinite-horizon Average-reward MDPs with Linear Function Approximation

**Yue Wu**
UCLA

**Dongruo Zhou**
UCLA

**Quanquan Gu**
UCLA

## Abstract

We study reinforcement learning in an infinite-horizon average-reward setting with linear function approximation for linear mixture Markov decision processes (MDPs), where the transition probability function of the underlying MDP admits a linear form over a feature mapping of the current state, action, and next state. We propose a new algorithm UCRL2-VTR, which can be seen as an extension of the UCRL2 algorithm with linear function approximation. We show that UCRL2-VTR with Bernstein-type bonus can achieve a regret of $\widetilde{O}(d\sqrt{DT})$, where $d$ is the dimension of the feature mapping, $T$ is the horizon, and $D$ is the diameter of the MDP. We also prove a matching lower bound $\widetilde{\Omega}(d\sqrt{DT})$, which suggests that the proposed UCRL2-VTR is minimax optimal up to logarithmic factors. To the best of our knowledge, our algorithm is the first nearly minimax optimal RL algorithm with function approximation in the infinite-horizon average-reward setting.

## 1 INTRODUCTION

One of the major goals of reinforcement learning (RL) is to maximize the expected accumulated reward within a certain environment, which is often represented by a *Markov Decision Process (MDP)*. To achieve this goal, the agent interacts with the environment sequentially under the guidance of certain *policy*, receives the reward returned by the environment, and updates the policy. There are several MDP settings such as finite-horizon episodic MDPs, infinite-horizon discounted MDPs, and infinite-horizon average-reward MDPs (See Sutton and Barto (2018) for a detailed introduction). Among them, the infinite-horizon average-reward MDP attracts a lot of attention, because it not only imposes the least constraints on the underlying MDP structure but also serves as a proper environment for many real-world decision-making problems that care more about the long-term return, such as factory optimization, and product delivery (Proper and Tadepalli, 2006), and automated trading in the financial markets. This is also the focus of this work.

A series of previous work (Bartlett and Tewari, 2009; Jaksch et al., 2010; Fruit et al., 2018b; Talebi and Maillard, 2018; Zhang and Ji, 2019) has proved both upper and lower regret bounds for learning infinite-horizon average-reward MDP in the tabular setting, where the number of states and actions are finite. Specifically, Jaksch et al. (2010) first proposed UCRL2 algorithm which achieves $\widetilde{O}(DS\sqrt{AT})$, where $S$ is the number of states, $A$ is the number of actions, and $D$ is the diameter of MDP. Jaksch et al. (2010) also proved that no algorithm can achieve a regret bound lower than $\Omega(\sqrt{DSAT})$. Zhang and Ji (2019) proposed a nearly minimax optimal algorithm EBF with $\widetilde{O}(\sqrt{DSAT})$ regret, which matches the lower regret up to logarithmic factors. However, these regret bounds depend on the cardinalities of state and action spaces (i.e., $S$ and $A$), which prevent the application of these algorithms to real-world RL with large state and action spaces.

To overcome the curse of large state space, *function approximation* has been used to design practically successful algorithms (Singh et al., 1995; Mnih et al., 2015; Bertsekas, 2018). However, most existing studies on learning infinite-horizon average-reward MDPs are limited to tabular MDPs, with only a few exceptions (Abbasi-Yadkori et al., 2019a,b; Hao et al., 2020; Wei et al., 2020a). More specifically, Abbasi-Yadkori et al. (2019a,b); Hao et al. (2020) studied RL with function approximation for infinite-horizon average-reward MDPs under strong assumptions such as uniformly-mixing and uniformly excited feature, and proved sub-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

linear regrets. More recently, Wei et al. (2020a) considered the linear MDP (Yang and Wang, 2019a; Jin et al., 2019) where the transition probability and reward function can be represented as linear functions over given feature mapping defined on the state and action pair, and proposed two algorithms, FOPO and OLSVI.FH. FOPO achieves $\widetilde{O}(\sqrt{d^3T})$ regret but relies on solving a fixed-point equation at each iteration, which is computationally inefficient. OLSVI.FH, on the other hand, is computationally efficient, but can only achieve $\widetilde{O}(T^{3/4})$ regret. Wei et al. (2021) also proposed another algorithm MDP-EXP2, which relies on very strong assumptions including the uniform-mixing and uniformly-excited-features assumptions to achieve $\widetilde{O}(\sqrt{T})$ regret. It remains an open question that if and how the minimax optimality of learning infinite-horizon average-reward MDPs can be achieved with linear function approximation.

**Main contribution.** In this paper, we resolve the above open question by proving nearly matching upper and lower regret bounds for a class of infinite-horizon MDPs called *linear mixture/kernel MDP* (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b) in the average-reward setting. Specifically, we propose a UCRL2-VTR algorithm based on the principle of "Optimism-in-Face-of-Uncertainty (OFU)". At the core of our algorithm is a variant of value-targeted regression (Jia et al., 2020) which estimates the unknown transition probability by least-squares over the expectations of bias functions, along with an optimistic exploration. We consider both Hoeffding-type bonus and Bernstein-type bonus for exploration. We show that UCRL2-VTR with Hoeffding-type bonus achieves a regret bound of $\widetilde{O}(dD\sqrt{T})$, and UCRL2-VTR with Bernstein-type bonus can improve the regret to be $\widetilde{O}(d\sqrt{DT})$, where $T$ is the time horizon, $D$ is the diameter of the MDP, and $d$ is the dimension of the feature mapping. We also prove a $\widetilde{\Omega}(d\sqrt{DT})$ lower bound on the regret of any algorithms under a given linear mixture MDP. The improved upper bound and the lower bound match each other up to the logarithmic factors. To the best of our knowledge, this is the first RL algorithm with linear function approximation that achieves a nearly-minimax optimal regret bound under the infinite-horizon average-reward MDP setting.

**Technical novelty.** Compared with many recent works on RL with linear function approximation (Abbasi-Yadkori et al., 2011; Azar et al., 2017; Jin et al., 2019; Ayoub et al., 2020; Zhou et al., 2021a) which study either the episodic MDP or discounted MDP settings, our work focuses on a quite different setting called the average-reward setting. Therefore, most algorithms and analyses are not directly applicable or extendable to our setting. For example, the average-reward setting uses a different notion of regret compared with those in the episodic setting or discounted settings. Thus, to bound the regret, a new regret decomposition is required, which does not appear in the aforementioned works. Besides the new regret decomposition, we also develop some other novel techniques. To bound the regret of the Bernstein-type algorithm, we construct a different variance estimator based on the centered value function $w_k$, instead of the standard value function. Another technical contribution we make is that we prove a new, non-trivial total variance lemma for the average-reward setting with linear function approximation, which plays a pivotal role in achieving minimax optimality.

**Notation.** We use lower case letters to denote scalars, lower and upper case bold letters to denote vectors and matrices. We use $\|\cdot\|$ to indicate Euclidean norm, and for a semi-positive definite matrix $\mathbf{\Sigma}$ and any vector $\mathbf{x}$, $\|\mathbf{x}\|_{\mathbf{\Sigma}} := \|\mathbf{\Sigma}^{1/2}\mathbf{x}\| = \sqrt{\mathbf{x}^\top \mathbf{\Sigma}\mathbf{x}}$. For a real value $x$ and an interval $[a, b]$, we use $[x]_{[a,b]}$ to indicate the projected value from $x$ onto $[a, b]$. We also use the standard $O$ and $\Omega$ notations. We say $a_n = O(b_n)$ if and only if $\exists C > 0, N > 0, \forall n > N, a_n \leq Cb_n$; $a_n = \Omega(b_n)$ if $a_n \geq Cb_n$. The notation $\widetilde{O}$ is used to hide logarithmic factors. For a random variable $X$, we say $X$ is $R$-sub-Gaussian if $\mathbb{E}[X] = 0$ and for any $s \in \mathbb{R}$, $\mathbb{E}[e^{sX}] \leq e^{R^2 s^2/2}$.

## 2 RELATED WORK

**Infinite-horizon average-reward tabular MDPs.** Learning infinite-horizon average-reward tabular MDPs has been thoroughly studied in the past decade. Bartlett and Tewari (2009); Jaksch et al. (2010) are among the first works providing $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. Recently, many algorithms were proposed to provide tighter regret bounds under various assumptions. Ouyang et al. (2017) proposed a PSRL that achieves the same regret as Jaksch et al. (2010). Agrawal and Jia (2017) proposed an algorithm using Thompson-sampling with an $\widetilde{\mathcal{O}}(D\sqrt{SAT})$ regret. Fruit et al. (2018b) proposed a SCAL algorithm with an $\widetilde{\mathcal{O}}(\mathrm{sp}(h^*)\sqrt{\Gamma SAT})$ regret, where sp is the span operator , $h^*$ is the optimal bias function for weakly-communicating MDP and $\Gamma$ is the number of next states. Fruit et al. (2018a) proposed a TUCRL algorithm for weakly-communicating MDPs. Talebi and Maillard (2018) proposed a KL-UCRL algorithm with an $\widetilde{\mathcal{O}}(\sqrt{S\mathbb{V}T})$ regret where $\mathbb{V}$ is the summation of variances of the bias function with respect to all state-action pairs. Zhang and Ji (2019) proposed an EBF algorithm with the near-optimal $\widetilde{\mathcal{O}}(\sqrt{\mathrm{sp}(h^*)SAT})$ regret. Fruit et al. (2020) proposed a UCRL2B algorithm with an $\widetilde{\mathcal{O}}(\Gamma\sqrt{SDAT})$ regret.

Ortner (2020) proposed an OSP algorithm for ergodic MDPs with an $\widetilde{\mathcal{O}}(\sqrt{t_{\text{mix}}SAT})$ regret, where $t_{\text{mix}}$ is the mixing time parameter. Wei et al. (2020a) proposed two model-free algorithms: optimistic Q-learning algorithm with an $\widetilde{\mathcal{O}}(\text{sp}(h^*)(SA)^{1/3}T^{2/3})$ regret, and MDP-OOMD for ergodic MDPs with an $\widetilde{\mathcal{O}}(\sqrt{t_{\text{mix}}^3\rho AT})$ regret, where $\rho$ is some distribution mismatch coefficient. Our algorithm is inspired by the algorithm UCRL2 proposed by Jaksch et al. (2010), which maintains a confidence set for the transition model and uses Extended Value Iteration (EVI) to obtain an optimistic model and a near-optimal policy under this model. Our work is also closely related to Fruit et al. (2020), which employs the empirical Bernstein inequality to provide tighter regret bound.

**RL with linear function approximation.** Provable RL with linear function approximation has received increasing interest in recent years. Jiang et al. (2017) proposed the low Bellman rank assumption and designed an OLIVE algorithm that achieves low sample complexity. As a special case of low Bellman rank MDPs, the *linear MDP* class (Yang and Wang, 2019a; Jin et al., 2019) assumes the transition probability and the reward function are linear functions Another related class of MDPs is the *linear mixture MDP* (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b), which assumes the transition probability is a linear combination of the feature mappings over the state-action-next-state triplet. Representative works include Yang and Wang (2019b); Modi et al. (2019); Jia et al. (2020); Zhou et al. (2021b); Cai et al. (2020); He et al. (2022). Note that while linear MDPs and linear mixture MDPs share some common subset, one cannot be covered by the other (Zhou et al., 2021b). Our algorithm and upper bound results are for linear mixture MDPs, while our lower bound holds for both MDPs. It is worth noting that for inhomogeneous episodic linear mixture MDPs, Zhou et al. (2021a) obtained matching upper and lower bounds of regret respectively, i.e., $\widetilde{O}(dH\sqrt{T})$ and $\Omega(dH\sqrt{T})$, where $H$ is the length of the episode. In comparison, we provide the first matching upper and lower bounds of regret in the infinite-horizon average-reward setting for linear mixture MDPs.

## 3 PRELIMINARIES

We denote a Markov Decision Process (MDP) by a tuple $M(\mathcal{S}, \mathcal{A}, r, \mathbb{P})$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathbb{P}(s'|s,a)$ is the transition probability function, and $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function. A deterministic stationary policy $\pi : \mathcal{S} \to \mathcal{A}$ maps a given state $s$ to a certain action $a$. In this work, we assume that the numbers of states and actions are finite, i.e., $|\mathcal{S}|, |\mathcal{A}| < \infty$ and the reward $r$ is known and deterministic. An algorithm $A$, which starts from some initial state $s \in \mathcal{S}$, will (stochastically) decide at each round $t$ what action $a_t$ to take. Given the state space $\mathcal{S}$, the action space $\mathcal{A}$, the algorithm starting from $s_0 = s$ actually induces a Markov process $\{s_t, a_t\}_t$, and it is natural to define the *undiscounted* reward as:

$$R(M, A, s, T) := \sum_{t=1}^{T} r(s_t, a_t) = \sum_{t=1}^{T} r_t.$$

We can define the expected average reward over time $T$ as $\mathbb{E}[R(M, A, s, T)]/T$ , and its limit

$$\rho(M, A, s) = \lim_{T\to\infty} \mathbb{E}[R(M, A, s, T)]/T,$$

is called the *average reward*.

In general, the learnability of an MDP depends on its transition structure. Following Jaksch et al. (2010), we define the *diameter* of an MDP as the expected steps taken from one state $s$ to another state $s'$ under the fastest stationary policy. Throughout the paper, the diameter is only used to provide an upper bound on the span of the value function.

**Definition 3.1.** Let $T(s'|M, \pi, s)$ be the random variable for the number of steps after which the state $s'$ is reached for the first time starting from $s$, under the MDP $M$ and policy $\pi$. Then the diameter $D(M)$ is defined as $D(M) = \max_{s,s'} \min_{\pi:\mathcal{S}\to\mathcal{A}} \mathbb{E}[T(s'|M, \pi, s)]$.

In this work, following Jaksch et al. (2010), we consider *communicating* MDPs (Puterman, 2014) which has a *finite* diameter. We assume a *known* upper bound $D$ on the diameters of all MDPs considered. One may question whether assuming a known upper bound is too restrictive. This will be further explained later in Remark 5.4.

For MDPs with finite diameter, the optimal average reward does not depend on the starting state. Therefore, we can define

$$\rho^*(M) = \rho^*(M, s) := \max_{\pi} \rho(M, \pi, s),$$

where $\pi$ is any stationary policy. Naturally, the regret is defined as

$$\text{Regret}(M, A, s, T) := T\rho^*(M) - R(M, A, s, T).$$

Besides the finite diameter assumption, another widely used assumption for infinite-horizon average-reward MDPs is the *finite span of optimal bias function (finite span)* assumption (Bartlett and Tewari, 2009; Zhang and Ji, 2019; Wei et al., 2020a), which assumes that there exists a $\rho^* \in \mathbb{R}$, $h^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that for any $s \in \mathcal{S}, a \in \mathcal{A}$, the following Bellman optimality equation holds:

$$\rho^* + h^*(s, a) = r(s, a) + \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}[\max_{a'\in\mathcal{A}} h^*(s', a')],$$

where the span of $\max_{a' \in \mathcal{A}} h^*(s, a')$ defined as $\max_{s,s' \in \mathcal{S}} |\max_{a' \in \mathcal{A}} h^*(s, a') - \max_{a' \in \mathcal{A}} h^*(s', a')|$ is finite, $h^*$ is the optimal state-action bias function. It can be verified that for any MDP with finite diameter, the span of the optimal bias function is also finite. We leave it as future work to extend our algorithm to deal with the finite span assumption.

**Linear mixture MDPs.** In many application scenarios where the state space $\mathcal{S}$ and action space $\mathcal{A}$ are intractably large, a certain structure of the transition kernel $\mathbb{P}(\cdot|s, a)$ will still enable efficient learning. One of such structures is *linear mixture MDP* (Modi et al., 2019; Zhou et al., 2021b; Jia et al., 2020; Ayoub et al., 2020), where the transition kernel can be represented by a linear combination of feature mappings.

**Definition 3.2.** $M(\mathcal{S}, \mathcal{A}, r, \mathbb{P})$ is called a $B$-bounded linear mixture MDP if there exist a *known* feature mapping $\boldsymbol{\phi}(s'|s, a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and an *unknown* vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$ with $\|\boldsymbol{\theta}^*\|_2 \leq B$, such that

- For any state-action-state triplet $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we have $\mathbb{P}(s'|s, a) = \langle \boldsymbol{\phi}(s'|s, a), \boldsymbol{\theta}^* \rangle$;
- For any bounded function $F : \mathcal{S} \to [0, 1]$ and any tuple $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have $\|\boldsymbol{\phi}_F(s, a)\|_2 \leq 1$, where $\boldsymbol{\phi}_F(s, a) = \sum_{s'} \boldsymbol{\phi}(s'|s, a) F(s') \in \mathbb{R}^d$.

We denote the linear mixture MDP by $M_{\boldsymbol{\theta}^*}$ for simplicity.

The motivation behind defining $\boldsymbol{\phi}_F$ is as follows: for any state action pair $s, a$, the expectation of $F(s')$ is

$$[\mathbb{P}F](s, a) := \mathbb{E}_{s' \sim \mathbb{P}}[F(s')] = \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, a) F(s') = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_F(s, a) \rangle,$$

which is a linear function of $\boldsymbol{\phi}_F(s, a)$. We will use this fact in both algorithm design and analysis.

It is noteworthy that tabular MDPs can be covered by linear mixture MDPs via setting the feature mapping $\boldsymbol{\phi}(s'|s, a)$ to be a one-hot vector for $s, a, s'$.

A recent work (Wei et al., 2020a) studied RL with linear function approximation under the *linear MDP* assumption, which assumes that there exists a known feature mapping $\boldsymbol{\Phi}(s, a)$ and an unknown mapping $\boldsymbol{\mu}(s')$ such that $\mathbb{P}(s'|s, a) = \langle \boldsymbol{\Phi}(s, a), \boldsymbol{\mu}(s') \rangle$. Both linear mixture MDPs and linear MDPs cover some common MDP, including tabular MDPs and bilinear MDPs (Yang and Wang, 2019b). However, in general, they are two different classes of MDPs, because their feature mappings are defined over different domains (Zhou et al., 2021b), and neither can cover the other.

In the rest of this paper, we also use the following shorthand to indicate the variance of the random variable $F(s')$ under distribution $P(\cdot|s, a)$:

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

and $[\mathbb{V}F]$ denotes the case when $P$ is the true transition probability $\mathbb{P}$ of the MDP.

# 4 ALGORITHMS

We are going to present two algorithms. The first algorithm is UCRL2-VTR (Algorithm 1), which extends the UCRL2 algorithm by Jaksch et al. (2010) from tabular MDPs to linear mixture MDPs. UCRL2-VTR includes two types of exploration strategies: Hoeffding-type bonus (OPTION I), and Bernstein-type bonus (OPTION II). The main difference between UCRL2-VTR and UCRL2 is the construction of confidence sets, due to the difference between the tabular MDP and the linear mixture MDP. The second algorithm is extended value iteration (EVI) (Algorithm 2), which serves as a subroutine of UCRL2-VTR to calculate the optimistic estimation of the value function.

## 4.1 UCRL2-VTR with Hoeffding-type bonus

We first present UCRL2-VTR with the **Hoeffding-type bonus (OPTION I)**. The learning process can be divided into several episodes indexed by $k$ ($t_k \leq t < t_{k+1}$). At the beginning of each episode, we call the subroutine Extended Value Iteration (EVI) (Jaksch et al., 2010) under a given confidence set $\widehat{\mathcal{C}}_{t_k}$, which contains the true parameter $\boldsymbol{\theta}^*$ with high probability. Within each episode, we follow the induced policy $\pi_{t_k}$ and use the new observation to obtain a better confidence set.

**How EVI works?** In Line 4 of Algorithm 2, it performs one step of extended value iteration, which takes maximum over both the action set $\mathcal{A}$ and the set of plausible transition models $\mathcal{C} \cap \mathcal{B}$. EVI terminates when the difference between two consecutive iterations is small enough (Line 5).

Generally speaking, EVI outputs the optimal value function corresponding to a near-optimal MDP among all plausible MDPs contained in the confidence set $\widehat{\mathcal{C}}_{t_k} \cap \mathcal{B}$, which is similar to its counterpart for tabular MDPs in Jaksch et al. (2010). The main difference from the EVI for tabular MDPs is that we need to restrict the parameter in the set $\mathcal{B}$ as well, which admits all parameters $\boldsymbol{\theta}$ that can induce a transition probability, i.e., $\mathcal{B} = \cap_{s,a} \mathcal{B}_{s,a}$ with $\mathcal{B}_{s,a} = \big\{\boldsymbol{\theta} : \langle \boldsymbol{\phi}(\cdot|s, a), \boldsymbol{\theta} \rangle \text{is a probability function}\big\}$. It is easy to show that $\mathcal{B}$ is a convex set. For some special feature mapping $\boldsymbol{\phi}$, $\mathcal{B}$ can be a very simple convex set. For

**Algorithm 1** Upper-Confidence Reinforcement Learning with Value Targeted Regression (UCRL2-VTR)

**Require:** Regularization parameter $\lambda$, upper bound $B$ of $\|\boldsymbol{\theta}^*\|_2$, precision of extended value iteration rounds $\epsilon$

1: Receive $s_1$, Set $k \leftarrow 0$, $t_0 \leftarrow 1$
2: **OPTION 1 (Hoeffding-type Bonus):** Set $\widehat{\boldsymbol{\Sigma}}_1 \leftarrow \lambda\mathbf{I}$, $\widehat{\mathbf{b}}_0 \leftarrow \mathbf{0}$
**OPTION 2 (Bernstein-type Bonus):** Set $\widehat{\boldsymbol{\Sigma}}_1, \widetilde{\boldsymbol{\Sigma}}_1 \leftarrow \lambda\mathbf{I}$, $\widehat{\mathbf{b}}_1, \widetilde{\mathbf{b}}_1, \widehat{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_1 \leftarrow \mathbf{0}$
3: Set $\pi_0(\cdot|s) \leftarrow \text{uniform}(\mathcal{A}), \forall s \in \mathcal{S}$
4: **for** $t = 1, 2, \ldots$ **do**
5: **if** $\det(\widehat{\boldsymbol{\Sigma}}_t) \leq 2\det(\widehat{\boldsymbol{\Sigma}}_{t_k})$ **then**
6: $\pi_t \leftarrow \pi_{t-1}$ { Keep the policy unchanged}
7: **else**
8: $k \leftarrow k+1$, $t_k \leftarrow t$ { Starting a new episode $k$}
9: Set $\widehat{\mathcal{C}}_t$ as (4.4) (**OPTION 1**) or (4.5) (**OPTION 2**)
10: Set $u_k(s) \leftarrow \text{EVI}(\widehat{\mathcal{C}}_t, \epsilon)$
11: Denote $w_k(s) = u_k(s) - \big(\max u_k(\cdot) - \min u_k(\cdot)\big)/2$
12: Set $\pi_t(s)$ as (4.2) { Compute new policy}
13: **end if**
14: Take action $a_t = \pi_t(s_t)$, receive $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$
15: **OPTION 1 (Hoeffding-type Bonus):**
16: Set $\widehat{\boldsymbol{\Sigma}}_{t+1} \leftarrow \widehat{\boldsymbol{\Sigma}}_t + \boldsymbol{\phi}_{w_k}(s_t, a_t)\boldsymbol{\phi}_{w_k}(s_t, a_t)^\top$, $\widehat{\mathbf{b}}_{t+1} \leftarrow \widehat{\mathbf{b}}_t + \boldsymbol{\phi}_{w_k}(s_t, a_t)w_k(s_{t+1})$
17: **OPTION 2 (Bernstein-type Bonus):**
18: Set $[\bar{\mathbb{V}}_t w_k]$ as in (4.7) and $E_t$ as in (4.8)
19: Set $\bar{\sigma}_t \leftarrow \sqrt{\max\{D^2/d, [\bar{\mathbb{V}}_t w_k](s_t, a_t) + E_t\}}$
20: Set $\widehat{\boldsymbol{\Sigma}}_{t+1} \leftarrow \widehat{\boldsymbol{\Sigma}}_t + \bar{\sigma}_t^{-2}\boldsymbol{\phi}_{w_k}(s_t, a_t)\boldsymbol{\phi}_{w_k}(s_t, a_t)^\top$, $\widehat{\mathbf{b}}_{t+1} \leftarrow \widehat{\mathbf{b}}_t + \bar{\sigma}_t^{-2}w_k(s_{t+1})\boldsymbol{\phi}_{w_k}(s_t, a_t)$
21: Set $\widetilde{\boldsymbol{\Sigma}}_{t+1} \leftarrow \widetilde{\boldsymbol{\Sigma}}_t + \boldsymbol{\phi}_{w_k^2}(s_t, a_t)\boldsymbol{\phi}_{w_k^2}(s_t, a_t)^\top$, $\widetilde{\mathbf{b}}_{t+1} \leftarrow \widetilde{\mathbf{b}}_t + w_k^2(s_{t+1})\boldsymbol{\phi}_{w_k^2}(s_t, a_t)$
22: Set $\widetilde{\boldsymbol{\theta}}_{t+1} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{t+1}^{-1}\widetilde{\mathbf{b}}_{t+1}$
23: Set $\widehat{\boldsymbol{\theta}}_{t+1} \leftarrow \widehat{\boldsymbol{\Sigma}}_{t+1}^{-1}\widehat{\mathbf{b}}_{t+1}$
24: **end for**

**Algorithm 2** Extended Value Iteration (EVI)

**Require:** A set $\mathcal{C}$, a desired accuracy level $\epsilon$

Set $u^{(0)}(s) \leftarrow 0, \forall s \in \mathcal{S}$
Set $i \leftarrow 0$
**repeat**
$\forall s \in \mathcal{S}$, set $u^{(i+1)}(s) \leftarrow \max_{a\in\mathcal{A}}\big\{r(s,a) + \max_{\boldsymbol{\theta}\in\mathcal{C}\cap\mathcal{B}}\{\langle\boldsymbol{\theta}, \boldsymbol{\phi}_{u^{(i)}}(s,a)\rangle\}\big\}$
**until** $\max_{s\in\mathcal{S}}\{u^{(i+1)}(s) - u^{(i)}(s)\} - \min_{s\in\mathcal{S}}\{u^{(i+1)}(s) - u^{(i)}(s)\} \leq \epsilon$
Return $u^{(i)}(s)$

instance, when $\boldsymbol{\phi}$ is a collection of $d$ transition probability functions, $\mathcal{B}$ is the $d$-dimension simplex (Modi et al., 2019). This will make the optimization in each step of EVI easy to solve. In detail, given the accuracy parameter $\epsilon$, $\text{EVI}(\widehat{\mathcal{C}}_t, \epsilon)$ outputs a value function $u^{(i)}$ satisfying

$$|u^{(i+1)}(s) - u^{(i)}(s) - \rho_k| \leq \epsilon, \tag{4.1}$$

where $\rho_k$ is the average reward under $\mathbb{P}_k$ and $\pi_{t_k}$, both of which are defined as follows

$$\begin{aligned}
\mathbb{P}_k(\cdot|s,a) &:= \langle\boldsymbol{\theta}_k(s,a), \boldsymbol{\phi}(\cdot|s,a)\rangle, \\
\boldsymbol{\theta}_k(s,a) &:= \operatorname*{argmax}_{\boldsymbol{\theta}\in\widehat{\mathcal{C}}_{t_k}\cap\mathcal{B}} \big\langle\boldsymbol{\theta}, \boldsymbol{\phi}_{u_k}(s,a)\big\rangle \\
\pi_{t_k}(s) &:= \operatorname*{argmax}_{a\in\mathcal{A}} \big\{r(s,a) + \big\langle\boldsymbol{\theta}_k(s,a), \boldsymbol{\phi}_{u^{(i)}}(s,a)\big\rangle\big\}.
\end{aligned} \tag{4.2}$$

When the context is clear, we will abuse the notation a little bit and use $\boldsymbol{\theta}_k$ to denote $\boldsymbol{\theta}_k(s_t, a_t)$ for different $t$. It is worth noting that $u(\cdot)$ is quite different from the traditional value function $V(\cdot)$ defined for episodic MDPs or discounted infinite-horizon MDPs. This requires a different analysis in the later proof.

Then the agent uses the greedy policy $\pi_{t_k}$ with respect to $u^{(i)}$ to select the actions in $k$-th episode. It is easy to see that centering $u^{(i)}(s)$ to $u^{(i)}(s) - (\max u^{(i)}(\cdot) - \min u^{(i)}(\cdot))/2$ does not change $\pi_{t_k}$. Therefore, we consider $w_k$ (in Line 11) as the centered version of $u_k$ in the later analysis.

An important observation for our later analysis, made by Jaksch et al. (2010), is that $u^{(i)}(s)$ computed in EVI (Algorithm 2) satisfies

$$\max_{s\in\mathcal{S}} u^{(i)}(s) - \min_{s\in\mathcal{S}} u^{(i)}(s) \leq D. \tag{4.3}$$

This is because $u^{(i)}(s)$ is the expected total $i$-step reward of an optimal non-stationary $i$-step policy starting from $s$. Suppose for $s$ and $s'$ we have $u^{(i)}(s) - D > u^{(i)}(s')$, then we can obtain a better $u^{(i)}(s')$ by adopting the following policy: first travel to $s$ as fast as possible (which takes at most $D$ steps in expectation), then following the optimal policy for $s$. Since the reward for each step belongs to $[0, 1]$, the new policy will gain at least $u^{(i)}(s) - D$, contradicting the optimality of $u^{(i)}(s')$. By (4.3), we also have the centered version $w_k$ satisfy $|w_k(s)| \le D/2$. It is safe and reasonable to consider $w_k$ instead of $u_k$ since we only care $\operatorname{argmax} u^{(i+1)}(\cdot)$ as the greedy policy.

**Convergence and efficiency of EVI.** Here we briefly discuss the convergence and computational efficiency of EVI. For the convergence, as we will show in the next section, the set of plausible MDPs induced by $\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{B}$ includes the true MDP $M_{\boldsymbol{\theta}^*}$ with high probability. Since we assume $M_{\boldsymbol{\theta}^*}$ is communicating in this work, according to Theorem 7 in Jaksch et al. (2010), EVI is guaranteed to converge. Here we present a sufficient condition under which EVI converges within logarithmic number of iterations. Define the quantity $\gamma(\mathcal{C})$ as follows:

$$\max_{\boldsymbol{\theta} \in \mathcal{C}, s, s' \in \mathcal{S}, a, a' \in \mathcal{A}} \left[1 - \sum_{j \in \mathcal{S}} \min\left\{\mathbb{P}_{\boldsymbol{\theta}}(j|s,a), \mathbb{P}_{\boldsymbol{\theta}}(j|s',a')\right\}\right],$$

and it is shown by Puterman (2014) (see Theorem 6.6.6.) that $\gamma(\mathcal{C})$ serves as a contraction coefficient:

$$\operatorname{span}(u^{(i+1)} - u^{(i)}) \le \gamma(\mathcal{C}) \cdot \operatorname{span}(u^{(i)} - u^{(i-1)}).$$

Therefore, as long as $\gamma(\mathcal{C}) < 1$, EVI will converge within logarithmic steps.

For the computation, suppose additionally, the feature mapping admits the form $[\boldsymbol{\phi}(s'|s,a)]_j = [\boldsymbol{\psi}(s')]_j \cdot [\boldsymbol{\mu}(s,a)]_j$ (Yang and Wang, 2019b), then we can use Monte Carlo integration to avoid the summation over the whole state space. In fact, we only need a few evaluations on each $[\boldsymbol{\psi}(s')]_j$ to obtain an accurate enough estimator $\widehat{\boldsymbol{\phi}}(s'|s,a)$ and perform the maximization over the estimated integration. This kind of feature mapping gives rise to a special case of linear mixture MDPs, namely bilinear MDPs (Yang and Wang, 2019b). Note that bilinear MDPs belong to both linear mixture MDPs and linear MDPs (Jin et al., 2019). More detailed discussions on the computational complexity of EVI can be found in Zhou et al. (2021b).

**Construction of $\widehat{\mathcal{C}}_t$.** Now we discuss how to construct the confidence set $\widehat{\mathcal{C}}_t$ at the end of each episode. We construct $\widehat{\mathcal{C}}_t$ as an ellipsoid centering at $\widehat{\boldsymbol{\theta}}_t$ with covariance matrix $\widehat{\boldsymbol{\Sigma}}_t$ defined in Line 16 of Algorithm 1. Moreover, we construct $\widehat{\boldsymbol{\theta}}_t$ as the minimizer to the ridge regression problem over contexts $\boldsymbol{\phi}_{w_k}(s_t, a_t)$ and targets $w_k(s_{t+1})$ with regularizer $\lambda\|\boldsymbol{\theta}\|_2^2$, whose closed-form solution is given in Line 23. The reason why we construct such a $\widehat{\boldsymbol{\theta}}_t$ is due to the following observation: the form of $\boldsymbol{\phi}_{w_k}(s_t, a_t)$ and $w_k(s_{t+1})$ fit in a linear bandits problem with stochastic reward. More specifically, by setting the action $\mathbf{x}_t = \boldsymbol{\phi}_{w_k}(s_t, a_t)$, the reward $y_t = w_k(s_{t+1})$, and the noise $\eta_t = w_k(s_{t+1}) - \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_{w_k}(s_t, a_t)\rangle$, we have

$$y_t = \langle \boldsymbol{\theta}^*, \mathbf{x}_t \rangle + \eta_t,$$

and $\mathbb{E}[\eta_t | \mathcal{F}_t] = 0$, $|\eta_t| \le D$, $\|\boldsymbol{\theta}^*\| \le B$, $\|\mathbf{x}_t\| \le D/2$.

This setting has been thoroughly studied in Abbasi-Yadkori et al. (2011). Define the confidence set as

$$\widehat{\mathcal{C}}_t = \left\{\boldsymbol{\theta} : \left\|\widehat{\boldsymbol{\Sigma}}_t^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t)\right\| \le \widehat{\beta}_t\right\}, \tag{4.4}$$

where $\widehat{\beta}_t$ is $\widehat{\beta}_t = D\sqrt{d \log\big((\lambda + tD^2)/(\delta\lambda)\big)} + \sqrt{\lambda} B$.

In the later section, we show that the true parameter $\boldsymbol{\theta}^*$ belongs to $\widehat{\mathcal{C}}_t$ with high probability. Therefore, $\widehat{\mathcal{C}}_t$ is a valid confidence set of $\boldsymbol{\theta}^*$, thus can be fed into the EVI procedure.

In summary, at each time step, UCRL2-VTR always takes action $a_t$ under the policy $\pi_t$, receives the next state $s_{t+1}$ and refines its confidence set with the new observation (Lines 14-23 of Algorithm 1). When it collects enough new observations for a better estimation than the previous one (Line 5), UCRL2-VTR calls the subroutine EVI with the tighter confidence set and obtains a better policy $\pi_t$ (Lines 8-12).

### 4.2 UCRL2-VTR with Bernstein-type bonus

UCRL2-VTR with **Bernstein-type bonus (OPTION 2)** is an improved variant of the basic version of UCRL2-VTR with Hoeffding-type bonus. The key difference is that here we are trying to utilize the variance information of the value functions to construct a tighter confidence set of $\boldsymbol{\theta}^*$. Recall the construction of $\widehat{\mathcal{C}}_t$ in Section 4.1, we set the center of the confidence set $\widehat{\boldsymbol{\theta}}_t$ as the solution to a ridge regression problem. The motivation is that UCRL2-VTR relies on the fact that the noise $\eta_t$ is $D$-bounded and therefore $D^2$-sub-Gaussian. However, the variance of $\eta_t$ is not necessarily that large. In fact, by the law of total variance (Azar et al., 2013), we know that on average, the variance of the noise is roughly on the order of $D$ rather than $D^2$. To enable the application of total variance, for $t_k < t \le t_{k+1}$, we set $\widehat{\boldsymbol{\theta}}_t$ as the solution to the following *weighted ridge regression problem*:

$$\operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{k} \sum_{j=t_k}^{t} \frac{[\langle \boldsymbol{\phi}_{w_i}(s_j, a_j), \boldsymbol{\theta}\rangle - w_i(s_{j+1})]^2}{\bar{\sigma}_j^2} + \lambda\|\boldsymbol{\theta}\|_2^2,$$

where $\bar{\sigma}_j^2$ (Line 19) is an estimation of the *variance* of $w_i(s_{j+1})$. Choosing the weights as the inverse of the variances can guarantee that the estimator has the lowest variance, similar to the best linear unbiased estimator (BLUE) estimator (Henderson, 1975) for linear regression with fixed design. After obtaining $\widehat{\boldsymbol{\theta}}_t$, similar to UCRL2-VTR with Hoeffding-type bonus, we construct the following confidence set which contains $\boldsymbol{\theta}^*$,

$$\widehat{\mathcal{C}}_t = \Big\{\boldsymbol{\theta} : \big\|\widehat{\boldsymbol{\Sigma}}_t^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t)\big\| \le \widehat{\beta}_t\Big\}, \tag{4.5}$$

where $\widehat{\boldsymbol{\Sigma}}_t$ is the covariance matrix of contexts $\boldsymbol{\phi}_{w_i}(s_j, a_j)$ weighted by $\bar{\sigma}_j^{-2}$, recursively defined in Line 20 of Algorithm 1; $\widehat{\beta}_t$ is defined as follows:

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

In our later analysis, we require $\bar{\sigma}_t^2$ to satisfy: (1) it upper bounds the true variance $[\mathbb{V}w_k](s_t, a_t)$ with high probability; and (2) it is strictly positive. To fulfil these two requirements, we first build a valid estimator $[\bar{\mathbb{V}}_t w_k](s_t, a_t)$ for the true variance $[\mathbb{V}w_k](s_t, a_t)$, based on the following fact:

$$\begin{aligned}[\mathbb{V}w_k](s_t, a_t) &= [\mathbb{P}w_k^2](s_t, a_t) - [\mathbb{P}w_k(s_t, a_t)]^2 \\ &= \langle \boldsymbol{\phi}_{w_k^2}(s_t, a_t), \boldsymbol{\theta}^*\rangle - \langle \boldsymbol{\phi}_{w_k}(s_t, a_t), \boldsymbol{\theta}^*\rangle^2,\end{aligned} \tag{4.6}$$

(4.6) suggests that the variance of $w_k$ can be regarded as the combination of two linear functions $\langle \boldsymbol{\phi}_{w_k^2}(s_t, a_t), \boldsymbol{\theta}^*\rangle$ and $\langle \boldsymbol{\phi}_{w_k}(s_t, a_t), \boldsymbol{\theta}^*\rangle$, with respect to different feature mappings. Therefore, we define our variance estimator $[\bar{\mathbb{V}}_t w_k](s_t, a_t)$ as follows

$$\begin{aligned}[\bar{\mathbb{V}}_t w_k](s_t, a_t) :=\ & \Big[\big\langle \boldsymbol{\phi}_{w_t^2}(s_t, a_t), \widetilde{\boldsymbol{\theta}}_t\big\rangle\Big]_{[0, D^2/4]} \\ & - \Big[\big\langle \boldsymbol{\phi}_{w_t}(s_t, a_t), \boldsymbol{\theta}_t\big\rangle\Big]^2_{[0, D/2]},\end{aligned} \tag{4.7}$$

where $\widetilde{\boldsymbol{\theta}}_t$ is another estimator for $\boldsymbol{\theta}^*$. Specifically, we choose $\widetilde{\boldsymbol{\theta}}_t$ as the solution to the ridge regression problem with contexts $\{\boldsymbol{\phi}_{w_i^2}(s_j, a_j)\}$ and targets $\{w_i^2(s_{j+1})\}$, whose closed-form solution is in Line 22 of Algorithm 1. Based on this initial variance estimator, we build our final variance estimator $\bar{\sigma}_t^2$ as in Line 19, where the correction term $E_t$ is defined as

$$\begin{aligned}E_t :=\ & \min\Big\{D^2/4, \widetilde{\beta}_t\Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_t^2}(s_t, a_t)\Big\|\Big\} \\ & + \min\Big\{D^2/4, D\check{\beta}_t\Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_t}(s_t, a_t)\Big\|\Big\},\end{aligned} \tag{4.8}$$

where $\widetilde{\boldsymbol{\Sigma}}_t$ is the covariance matrix of features $\boldsymbol{\phi}_{w_i^2}(s_j, a_j)$ recursively defined in Line 20, $\check{\beta}_t$ and $\widetilde{\beta}_t$ are defined respectively as follows

$$\begin{aligned}\check{\beta}_t :=\ & 8d\sqrt{\log(1+t/4\lambda)\log(4t^2/\delta)} \\ & + 4\sqrt{d}\log(4t^2/\delta) + \sqrt{\lambda}B, \\ \widetilde{\beta}_t :=\ & 2D^2\sqrt{d\log(1+tD^2/4d\lambda)\log(4t^2/\delta)} \\ & + D^2\log(4t^2/\delta) + \sqrt{\lambda}B.\end{aligned}$$

It can be verified that such a $\bar{\sigma}_t^2$ satisfies both conditions discussed above (i.e., larger than the true variance, strictly positive) simultaneously.

## 5 MAIN RESULTS

In this section, we present the regret analysis for Algorithm 1 with both Hoeffding-type exploration bonus (OPTION 1) and Bernstein-type bonus (OPTION 2).

**Theorem 5.1.** Setting $\lambda = 1/B^2, \epsilon = 1/\sqrt{T}$, then for any initial state $s_1$, with probability at least $1-2\delta$, the regret of Algorithm 1 with Hoeffding-type is bounded as follows:

$$\text{Regret}(T) = \widetilde{O}(dD\sqrt{T}),$$

where the $\widetilde{O}(\cdot)$ hides logarithmic terms of $d, D, T$, and $\delta^{-1}$.

Theorem 5.1 shows that the regret of Algorithm 1 only depends on the number of rounds $T$, the feature dimension $d$, and the diameter of the communicating MDP $D$. Therefore, Algorithm 1 is statistically efficient for linear mixture MDPs with a finite diameter but large state and action spaces.

**Remark 5.2.** The UCRL2 algorithm proposed in Jaksch et al. (2010) has an $\widetilde{O}(|\mathcal{S}|D\sqrt{|\mathcal{A}|T})$ regret bound for tabular MDPs with finite state and action spaces and diameter. As a comparison, our UCRL2-VTR enjoys a better upper bound $\widetilde{O}(dD\sqrt{T})$ when $d \le |\mathcal{S}|\sqrt{|\mathcal{A}|}$, which suggests that RL with linear function approximation can be more advantageous than vanilla RL algorithms when the underlying MDP has certain nice structures (Modi et al., 2019; Yang and Wang, 2019b).

**Theorem 5.3.** Set $\lambda = 1/B^2, \epsilon = 1/\sqrt{T}$, then for any initial state $s_1$, with probability at least $1-5\delta$, the regret of Algorithm 1 with Bernstein bonus is bounded as follows:

$$\text{Regret}(T) = \widetilde{O}\big(D\sqrt{dT} + d\sqrt{DT} + Dd^{7/4}T^{1/4}\big),$$

where the $\widetilde{O}(\cdot)$ hides logarithmic terms of $d, D, T, \delta^{-1}$.

Suppose $d \ge D$ and $T \ge D^2d^3$, then Theorem 5.3 suggests that by using the Bernstein-type exploration strategy, the regret bound of Algorithm 1 can be further improved by a factor of $\sqrt{D}$ to be $\widetilde{O}(d\sqrt{DT})$.

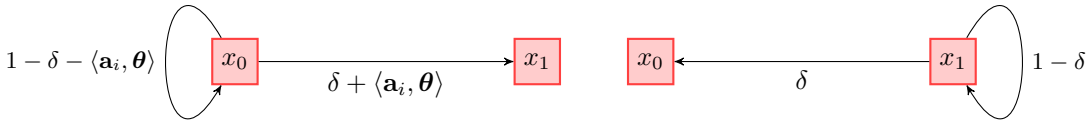


Figure 1: Illustration of the hard-to-learn linear mixture MDP considered in Theorem 5.5. The left figure demonstrates the state transition probability starting from $x_0$ with some specific action $\mathbf{a}_i$. The right figure demonstrates the state transition probability starting from $x_1$ with any action. For the detailed construction, see Supplementary Section C.1.

**Remark 5.4.** It can be shown that the algorithm and analysis can be easily modified to deal with an unknown bounded diameter $D$. More specifically, we can start from a small guess $D'$ and run the algorithm as if it is a valid upper bound. The guess $D'$ will be rejected once EVI returns a value function $u_k(s)$ with a span larger than $D'$ . Then we retry Line 9-11 with a doubled guess $2D'$. Meanwhile, if no violation happens and the algorithm ends, the proof will go through for $D'$ . In the worst case, a guess of at most $2D$ will ensure that EVI obtains a valid estimation, thus introducing a constant factor of up to 2.

**Comparison with FOPO** We would like to do a comparison between UCRL2-VTR in Algorithm 1 and the recently proposed FOPO by Wei et al. (2020a). While FOPO is originally proposed for linear MDPs and our algorithm is designed for linear mixture MDPs, since bilinear MDPs (Yang and Wang, 2019b) is a special class of both MDP classes, we can choose the bilinear MDPs as a common ground for comparison. More specifically, both UCRL2-VTR and FOPO focus on using linear function approximation to learn an infinite-horizon average-reward MDP, and both of them use the *optimism-in-the-face-of-uncertainty* (OFU) principle to learn the optimal value function among a class of plausible MDPs. However, FOPO adapts the Bellman optimality equation assumption and learns the optimal value function by solving a constrained nonconvex optimization problem, which is hard to solve even in the tabular MDP case. In sharp contrast, similar to Jaksch et al. (2010), UCRL2-VTR adapts the finite-diameter assumption and uses the EVI procedure to find the optimal value function, which is computationally efficient for bilinear MDPs (See Zhou et al. (2021b,a) for a detailed discussion). In the setting of bilinear MDPs, we have $\mathbb{P}(s'|s,a) = \sum_{j=1}^{d} \theta_j \psi_j(s') \mu_j(s,a)$. We can rewrite the transition probability as a linear mixture MDP ($\odot$ denotes the Hadamard product):

$$\mathbb{P}(s'|s,a) = \big\langle \boldsymbol{\theta}, \boldsymbol{\psi}(s) \odot \boldsymbol{\mu}(s,a) \big\rangle,$$

and as a linear MDP:

$$\mathbb{P}(s'|s,a) = \big\langle \boldsymbol{\theta} \odot \boldsymbol{\psi}(s), \boldsymbol{\mu}(s,a) \big\rangle.$$

Therefore, the regret bound of UCRL2-VTR and FOPO are $O(d\sqrt{DT})$ and $O(d^{3/2}\sqrt{\mathrm{span}(h^*)T})$, respectively. The diameter $D$ and the span $\mathrm{span}(h^*)$ are closely related and are often of the same scale. This immediately shows that UCRL2-VTR achieves a smaller regret and is statistically more efficient than FOPO for bilinear MDPs.

The following theorem presents a matching lower bound of regret for infinite-horizon average-reward linear mixture MDPs.

**Theorem 5.5.** Suppose $d \geq 2$, $T \geq 16d^2D/2025$ and $B > 1$. Then for any algorithm $A$, there exists a $B$-bounded MDP $M_{\boldsymbol{\theta}}$ such that

$$\mathbb{E}[\mathrm{Regret}(M_{\boldsymbol{\theta}}, A, s, T)] \geq d\sqrt{DT}/2025.$$

If we set $\delta = T^{-1}$ in Theorem 5.3, we can see the expected regret upper bound is of the order $\widetilde{O}(d\sqrt{DT})$, which differs from the lower bound only by logarithmic factors. The dependence on $d, D, T$ matches with each other and thus implies the upper bound cannot be improved.

**Remark 5.6.** An interesting observation is, Jaksch et al. (2010) proved in the tabular setting that for any algorithm, the regret is lower bounded by $\Omega(\sqrt{|\mathcal{S}||\mathcal{A}|DT})$. Since tabular MDPs can be regarded as a special case of linear mixture MDP with a $d = |\mathcal{S}|^2|\mathcal{A}|$ dimensional feature mapping, Jaksch et al. (2010)'s construction actually yields a slightly worse lower bound $\Omega(\sqrt{|\mathcal{S}|^2|\mathcal{A}|DT}) = \Omega(\sqrt{dDT})$, for the general linear mixture MDPs. And while our construction gains inspiration from Jaksch et al. (2010), $\Omega(d\sqrt{DT})$ is tighter than the induced lower bound by a factor of $d^{1/2}$. This also indicates that our MDP construction is nontrivial.

**Remark 5.7.** Our lower bound can also imply a lower bound for the *linear MDP* setting studied by Wei et al.

(2020a). By a similar construction of the hard-to-learn MDP instance, we can prove an $\Omega(d\sqrt{\mathrm{sp}(h^*)T})$ lower bound for learning linear MDPs. The detailed reasoning is deferred to the appendix. This suggests that there still exists a gap to be removed under the linear MDP setting considered by Wei et al. (2020a).

# 6 CONCLUSION AND FUTURE WORK

In this paper, we push the frontier of learning infinite-horizon average-reward Markov Decision Process with linear function approximation. We propose the first algorithm that achieves nearly minimax optimal regret. Our lower bound can also imply a lower bound for linear MDPs, which is of independent interest. Our current algorithms and results are limited to MDPs with finite diameter. In the future work, it is possible to relax this constraint, and extend our algorithms to deal with a milder assumption called finite span assumption (Bartlett and Tewari, 2009; Zhang and Ji, 2019; Wei et al., 2020a), while still achieving the minimax optimality.

### Acknowledgements

We thank the anonymous reviewers for their helpful comments. Part of this work was done when DZ and QG participated the Theory of Reinforcement Learning program at the Simons Institute for the Theory of Computing in Fall 2020. YW, DZ and QG are partially supported by the National Science Foundation CIF-1911168, CAREER Award 1906169, IIS-1904183 and AWS Machine Learning Research Award. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

### References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. (2019a). Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR.

Abbasi-Yadkori, Y., Lazic, N., Szepesvari, C., and Weisz, G. (2019b). Exploration-enhanced politex. *arXiv preprint arXiv:1908.10479*.

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

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

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77.

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

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349.

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

Bartlett, P. and Tewari, A. (2009). Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Uncertainty in Artificial Intelligence: Proceedings of the 25th Conference*, pages 35–42. AUAI Press.

Bertsekas, D. P. (2018). Feature-based aggregation and deep reinforcement learning: A survey and some new implementations. *IEEE/CAA Journal of Automatica Sinica*, 6(1):1–31.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2020). Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR.

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

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

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

Hao, B., Lazic, N., Abbasi-Yadkori, Y., Joulani, P., and Szepesvari, C. (2020). Provably efficient adaptive approximate policy iteration. *arXiv preprint arXiv:2002.03069*.

He, J., Zhou, D., and Gu, Q. (2022). Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In *International Conference on Artificial Intelligence and Statistics*.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447.

Jafarnia-Jahromi, M., Wei, C.-Y., Jain, R., and Luo, H. (2020). A model-free learning algorithm for infinite-horizon average-reward mdps with near-optimal regret. *arXiv preprint arXiv:2006.04354*.

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

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. (2020). Model-based reinforcement learning with value-targeted regression.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1704–1713. JMLR. org.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2019). Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

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

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

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

Proper, S. and Tadepalli, P. (2006). Scaling model-based average-reward reinforcement learning for product delivery. In *European Conference on Machine Learning*, pages 735–742. Springer.

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

Singh, S. P., Jaakkola, T., and Jordan, M. I. (1995). Reinforcement learning with soft state aggregation. In *Advances in neural information processing systems*, pages 361–368.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

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

Wei, C.-Y., Jafarnia-Jahromi, M., Luo, H., and Jain, R. (2020a). Learning infinite-horizon average-reward mdps with linear function approximation. *arXiv preprint arXiv:2007.11849*.

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

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. (2020b). Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International Conference on Machine Learning*, pages 10170–10180. PMLR.

Yang, L. and Wang, M. (2019a). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004.

Yang, L. F. and Wang, M. (2019b). Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*.

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

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

Zhou, D., He, J., and Gu, Q. (2021b). Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR.

# A Proof of Theorem 5.1

We use $K(T)-1$ to denote the value of the counter $k$ when Algorithm 1 finishes, and $t_{K(T)}=T+1$ for convenience. By these notation, the learning process from $t=1$ to $t=T$ can be divided into $K(T)$ episodes.

The following lemma, proved by Jaksch et al. (2010), states that EVI (Algorithm 2) always outputs a near-optimal policy and an optimistic model.

**Lemma A.1** (Theorem 7 and Equation (12) in Jaksch et al. 2010)**.** Stopping the extended value iteration when

$$\max_{s\in\mathcal{S}}\{u^{(i+1)}(s)-u^{(i)}(s)\}-\min_{s\in\mathcal{S}}\{u^{(i+1)}(s)-u^{(i)}(s)\}<\epsilon,$$

the greedy policy $\widetilde{\pi}$ with respect to $u^{(i)}$ is $\epsilon$-optimal, namely

$$\widetilde{\rho}:=\rho(\widetilde{M},\widetilde{\pi})\geq\max_{\pi,M\in\mathcal{M}}\rho(M,\pi)-\epsilon. \tag{A.1}$$

Here, $\widetilde{M}$ means the Markov Decision Process (MDP) determined by the parameterized transition probability, e.g. $\mathbb{P}_k(\cdot|s,a)=\langle\boldsymbol{\phi}(\cdot|s,a),\boldsymbol{\theta}_k(s,a)\rangle$. For each $M\in\mathcal{M}$, $M$ is an MDP with parameter from the confidence set. $\mathcal{M}$ is assumed to contain the true transition model.

Moreover, we have $\forall s\in\mathcal{S}$,

$$|u^{(i+1)}(s)-u^{(i)}(s)-\widetilde{\rho}|\leq\epsilon. \tag{A.2}$$

The next lemma describes that indeed, the confidence sets we constructed contain the true parameter with high probability.

**Lemma A.2.** With probability at least $1-\delta$, for all $0\leq k\leq K(T)-1$, we have $\boldsymbol{\theta}^*\in\widehat{\mathcal{C}}_{t_k}$.

*Proof.* See Section D.1. □

The number of episodes in our algorithm turns out can be bounded as follows:

**Lemma A.3.** We have $K(T)\leq d\log[(2\lambda+2TD^2)/\lambda]$.

*Proof.* See Section D.2. □

The rest lemmas either is standard concentration inequalities or is from the works regarding linear bandit problems.

**Lemma A.4** (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 $\mu$. Then with probability at least $1-\delta$, we have

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

**Lemma A.5** (Lemma 11 in Abbasi-Yadkori et al. 2011)**.** For any $\{\mathbf{x}_t\}_{t=1}^T\subset\mathbb{R}^d$ satisfying that $\|\mathbf{x}_t\|_2\leq L$, let $\mathbf{A}_0=\lambda\mathbf{I}$ and $\mathbf{A}_t=\mathbf{A}_0+\sum_{i=1}^{t-1}\mathbf{x}_i\mathbf{x}_i^\top$, then we have

$$\sum_{t=1}^{T}\min\{1,\|\mathbf{x}_t\|_{\mathbf{A}_{t-1}^{-1}}\}^2\leq 2d\log\frac{d\lambda+TL^2}{d\lambda}.$$

**Lemma A.6** (Lemma 12 in Abbasi-Yadkori et al. 2011)**.** Suppose $\mathbf{A},\mathbf{B}\in\mathbb{R}^{d\times d}$ are two positive definite matrices satisfying that $\mathbf{A}\succeq\mathbf{B}$, then for any $\mathbf{x}\in\mathbb{R}^d$, $\|\mathbf{x}\|_{\mathbf{A}}\leq\|\mathbf{x}\|_{\mathbf{B}}\cdot\sqrt{\det(\mathbf{A})/\det(\mathbf{B})}$.

*Proof of Theorem 5.1.* We first split the regret into each episode. Denote the regret in episode $k$ as $\Delta_k$, and we have

$$\begin{aligned}
\Delta_k &:= \sum_{t=t_k}^{t_{k+1}-1} [\rho^* - r(s_t, a_t)] \\
&\leq (t_{k+1} - t_k)\epsilon + \sum_{t=t_k}^{t_{k+1}-1} [\rho_k - r(s_t, a_t)] \\
&\leq 2(t_{k+1} - t_k)\epsilon + \sum_{t=t_k}^{t_{k+1}-1} \big[ \sum_{s' \in \mathcal{S}} \mathbb{P}_k(s'|s_t, a_t) u_k(s') - u_k(s_t) \big] \\
&= 2(t_{k+1} - t_k)\epsilon + \sum_{t=t_k}^{t_{k+1}-1} \big[ \sum_{s' \in \mathcal{S}} \mathbb{P}_k(s'|s_t, a_t) w_k(s') - w_k(s_t) \big] \\
&= 2(t_{k+1} - t_k)\epsilon + \sum_{t=t_k}^{t_{k+1}-1} \big[ [\mathbb{P}_k w_k](s_t, a_t) - w_k(s_t) \big].
\end{aligned}$$

The first inequality is due to the $\epsilon-$optimality of the EVI algorithm (Lemma A.1). The second inequality is due to (A.2) and substitute the iteration rule $u^{(i+1)}(s) \leftarrow \max_{a \in \mathcal{A}} \Big\{ r(s,a) + \max_{\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{B}} \Big\{ \big\langle \boldsymbol{\theta}, \boldsymbol{\phi}_{u^{(i)}}(s,a) \big\rangle \Big\} \Big\}$. Here, notice that we denote $\mathbb{P}_k(s'|s_t, a_t) = \langle \boldsymbol{\theta}_k(s_t, a_t), \boldsymbol{\phi}(s'|s_t, a_t) \rangle$ and $\boldsymbol{\theta}_k(s_t, a_t) = \operatorname{argmax}_{\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{B}} \Big\{ \big\langle \boldsymbol{\theta}, \boldsymbol{\phi}_{u^{(i)}}(s,a) \big\rangle \Big\}$. By the definition of $\pi_k$, $a_t$ achieves the outer maximum in the iteration rule of $u^{(i+1)}$. The second last equality is due to the fact that adding a bias to $u_k$ won't change the difference, as what has been done in Algorithm 1. So we subtract $(\max_s u_k(s) + \min_s u_k(s))/2$ from $u_k(s)$. The last equality is a shorthand. Notice that since the span of $u_k(s)$ is $D$, we have $|w_k(s)| \leq D/2$.

Summing over all episodes, we further have

$$\begin{aligned}
\sum_{k=0}^{K(T)-1} \Delta_k = 2T\epsilon &+ \underbrace{\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \big[ [\mathbb{P}_k w_k](s_t, a_t) - [\mathbb{P} w_k](s_t, a_t) \big]}_{I_1} \\
&+ \underbrace{\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \big[ [\mathbb{P} w_k](s_t, a_t) - w_k(s_t) \big]}_{I_2}.
\end{aligned}$$

The first term can be controlled following the idea of bounding the regret of linear bandit. We have that with probability $1 - \delta$,

$$\begin{aligned}
I_1 &= \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \big\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \boldsymbol{\phi}_{w_k}(s_t, a_t) \big\rangle \\
&\leq \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \big( \|\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_t} + \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_t} \big) \big\| \boldsymbol{\phi}_{w_k}(s_t, a_t) \big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}} \\
&\leq \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} 2\big( \|\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{t_k}} + \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{t_k}} \big) \big\| \boldsymbol{\phi}_{w_k}(s_t, a_t) \big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}} \\
&\leq \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} 4\widehat{\beta}_T \big\| \boldsymbol{\phi}_{w_k}(s_t, a_t) \big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}.
\end{aligned}$$

The first inequality is due to first applying Cauchy-Schwartz inequality and then the triangle inequality. The second inequality is due to Lemma A.6 and the fact that for $t_k \leq t < t_{k+1}$ $\det(\boldsymbol{\Sigma}_t) \leq \det(\boldsymbol{\Sigma}_{t_{k+1}}) \leq 2\det(\boldsymbol{\Sigma}_{t_k})$. The third inequality is due to Lemma A.2 and the fact that $\{\widehat{\beta}_t\}_t$ is increasing.

Meanwhile, for each term in $I_1$, we also have that due to the fact $|w_k(s)| \leq D/2$,

$$[\mathbb{P}_k w_k](s_t, a_t) - [\mathbb{P} w_k](s_t, a_t) \leq D.$$

Therefore, we have

$$\begin{aligned}
I_1 &\leq \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{D, 4\widehat{\beta}_T \big\|\boldsymbol{\phi}_{w_k}(s_t, a_t)\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\Big\} \\
&\leq 4\widehat{\beta}_T \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{1, \big\|\boldsymbol{\phi}_{w_k}(s_t, a_t)\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\Big\} \\
&\leq 4\widehat{\beta}_T \sqrt{T \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{1, \big\|\boldsymbol{\phi}_{w_k}(s_t, a_t)\big\|^2_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\Big\}} \\
&\leq 4\widehat{\beta}_T \sqrt{T \cdot 2d \log\bigg(\frac{d\lambda + TD^2}{d\lambda}\bigg)} \\
&\leq 6\widehat{\beta}_T \sqrt{dT \log\bigg(\frac{d\lambda + TD^2}{d\lambda}\bigg)}.
\end{aligned}$$

The second inequality is due to the fact $D \leq 4\widehat{\beta}_T$. The third is due to Cauchy-Schwartz inequality. The fourth is due to Lemma A.5.

The second term, can be controlled by the concentration of martingale. With probability $1 - \delta$,

$$\begin{aligned}
I_2 &= \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \big[[\mathbb{P} w_k](s_t, a_t) - w_k(s_t)\big] \\
&= \sum_{k=0}^{K(T)-1} \bigg[\sum_{t=t_k}^{t_{k+1}-1} \big([\mathbb{P} w_k](s_t, a_t) - w_k(s_{t+1})\big) - w_k(s_{t_k}) + w_k(s_{t_{k+1}})\bigg] \\
&\leq \sum_{k=0}^{K(T)-1} \bigg[\sum_{t=t_k}^{t_{k+1}-1} \big([\mathbb{P} w_k](s_t, a_t) - w_k(s_{t+1})\big)\bigg] + D \cdot K(T) \\
&\leq D\sqrt{2T \log(1/\delta)} + D \cdot K(T),
\end{aligned}$$

where the first inequality holds because $|w_k(s)| \leq D/2$; the second inequality is due to Lemma A.4.

Therefore, the total regret is bounded by

$$\text{Regret}(T) = \sum_{k=0}^{K(T)-1} \Delta_k \leq 2T\epsilon + 6\widehat{\beta}_T \sqrt{dT \log\bigg(\frac{\lambda + TD^2}{\lambda}\bigg)} + D\sqrt{2T \log(1/\delta)} + D \cdot K(T).$$

If we set

$$\widehat{\beta}_t = D\sqrt{d \log\bigg(\frac{\lambda + tD^2}{\delta\lambda}\bigg)} + \sqrt{\lambda} B,$$

and

$$\epsilon = \frac{1}{\sqrt{T}},$$

then by taking union bound we have with probability $1 - 2\delta$,

$$\begin{aligned}
\text{Regret}(T) &\leq 2\sqrt{T} + Dd\sqrt{T} \cdot \widetilde{O}(1) + B\sqrt{\lambda dT} \cdot \widetilde{O}(1) + D\sqrt{2T \log(1/\delta)} + Dd \log\bigg(\frac{2\lambda + 2dTD^2}{\lambda}\bigg) \\
&\leq \widetilde{O}(Dd\sqrt{T}),
\end{aligned}$$

where $\widetilde{O}(1)$ hides the log factor, the last inequality holds since we set $\lambda = 1/B^2$.

□

## B Proof of Theorem 5.3

Most part of the proof resembles that of Theorem 5.1. The additional part is essentially about the new concentration results from variance-aware linear bandit problem. As previously defined, we use $K(T)-1$ to denote the value of the counter $k$ when Algorithm 1 finishes, and $t_{K(T)} = T+1$ for convenience. By these notations, the learning process from $t=1$ to $t=T$ can be divided into $K(T)$ episodes.

The first lemma provides a better confidence set given the information of the noise's variance.

**Lemma B.1** (Bernstein inequality for vector-valued martingales (Zhou et al., 2021a))**.** Let $\{\mathcal{G}_t\}_{t=1}^{\infty}$ be a filtration, $\{\mathbf{x}_t, \eta_t\}_{t\geq 1}$ a stochastic process so that $\mathbf{x}_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{G}_{t+1}$-measurable. Fix $R, L, \sigma, \lambda > 0$, $\boldsymbol{\mu}^* \in \mathbb{R}^d$. For $t \geq 1$ let $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t \rangle + \eta_t$ and suppose that $\eta_t, \mathbf{x}_t$ also satisfy

$$|\eta_t| \leq R,\ \mathbb{E}[\eta_t|\mathcal{G}_t] = 0,\ \mathbb{E}[\eta_t^2|\mathcal{G}_t] \leq \sigma^2,\ \|\mathbf{x}_t\|_2 \leq L.$$

Then, for any $0 < \delta < 1$, with probability at least $1-\delta$ we have

$$\forall t > 0,\ \left\| \sum_{i=1}^{t} \mathbf{x}_i \eta_i \right\|_{\mathbf{Z}_t^{-1}} \leq \beta_t,\ \|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} \leq \beta_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2, \tag{B.1}$$

where for $t \geq 1$, $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1}\mathbf{b}_t$, $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^{t} \mathbf{x}_i\mathbf{x}_i^\top$, $\mathbf{b}_t = \sum_{i=1}^{t} y_i\mathbf{x}_i$ and

$$\beta_t = 8\sigma\sqrt{d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)} + 4R\log(4t^2/\delta)\,.$$

The number of episodes is bounded almost in the same way as in Lemma A.3:

**Lemma B.2.** Let $K(T)$ be as defined above. Then, $K(T) \leq 2d\log(1+Td/\lambda)$.

*Proof.* See Section D.3. □

The variance term is defined as

$$[\mathbb{V}w_k](s_t, a_t) := \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k^2(s')] - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')]^2.$$

The following lemma states that with high probability the estimated variance is close the the true variance.

**Lemma B.3.** With probability $1-3\delta$, we have for all $1 \leq t \leq T$,

$$\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_t \cap \mathcal{B}, \left|[\bar{\mathbb{V}}_t w_k](s_t, a_t) - [\mathbb{V}w_k](s_t, a_t)\right| \leq E_t.$$

We denote the event above by $\mathcal{E}_0$, and $\mathbb{P}(\mathcal{E}_0) \geq 1-3\delta$.

*Proof.* See Section D.4. □

Now, we define other events:

$$\mathcal{E}_1 := \left\{ \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \left[\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')^2] - w_k^2(s_{t+1})\right] \leq (D^2/4)\sqrt{2T\log(1/\delta)} \right\}$$

$$\mathcal{E}_2 := \left\{ \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \left[\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')] - w_k(s_{t+1})\right] \leq (D/2)\sqrt{2T\log(1/\delta)} \right\}$$

By the Azuma-Hoeffding inequality(Lemma A.4), we have $\mathbb{P}(\mathcal{E}_1) \geq 1-\delta$ and $\mathbb{P}(\mathcal{E}_2) \geq 1-\delta$.

The next lemma characterizes the total variance.

**Lemma B.4.** Under the events $\mathcal{E}_0$ and $\mathcal{E}_1$, we have

$$\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}[\mathbb{V}w_k](s_t,a_t) \le (D^2/4)\sqrt{2T\log(1/\delta)} + (K(T)+1)(D^2/4) + 2DT + D^2\widehat{\beta}_T\sqrt{T2d\log(1+T/\lambda)}.$$

*Proof.* See Section D.5. □

The following lemma serves as a wrapper of calculating the total estimation error.

**Lemma B.5.** Under the event $\mathcal{E}_0$, we have

$$\sum_{t=1}^{T} E_t \le \widetilde{\beta}_T\sqrt{2Td\log(1+TD^2/4d\lambda)} + D^2\check{\beta}_T\sqrt{2Td\log(1+T/\lambda)}.$$

*Proof.* See Section D.6. □

Now we are ready to show the regret upper bound.

*Proof.* We first follow the same procedure as Jaksch et al. (2010) did to decompose the regret and tackle each term respectively.

We have

$$\begin{aligned}
\text{Regret}(T) &:= \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}[\rho^* - r(s_t,a_t)] \\
&\le T\epsilon + \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}[\rho_k - r(s_t,a_t)] \\
&\le 2T\epsilon + \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[\mathbb{E}_{s'\sim\mathbb{P}_k(\cdot|s_t,a_t)}[u_k(s')] - u_k(s_t)\big] \\
&= 2T\epsilon + \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[\mathbb{E}_{s'\sim\mathbb{P}_k(\cdot|s_t,a_t)}[w_k(s')] - w_k(s_t)\big] \\
&= 2T\epsilon + \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}_k w_k](s_t,a_t) - w_k(s_t)\big].
\end{aligned}$$

The first inequality is due to the $\epsilon$-optimality of the EVI algorithm. The second inequality is due to (12) in Jaksch et al. (2010). The third inequality is due to the fact that add a bias to $u_t$ won't change the difference, as done in Algorithm 1. So we subtract $(\max_s u_t(s) + \min_s u_t(s))/2$ from $u_t(s)$. The last equality is a shorthand. It can be further decomposed into:

$$\begin{aligned}
\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}_k w_k](s_t,a_t) - w_k(s_t)\big] = &\underbrace{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}_k w_k](s_t,a_t) - [\mathbb{P}w_k](s_t,a_t)\big]}_{I_1} \\
&+ \underbrace{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}w_k](s_t,a_t) - w_k(s_t)\big]}_{I_2}.
\end{aligned}$$

We deal with the second term $I_2$ first:

The second term, can be controlled by the concentration of martingale. In fact, $\mathcal{E}_2$ defined above exactly characterizes the concentration. Under event $\mathcal{E}_2$, we have

$$\begin{aligned}
I_2 &= \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}w_k](s_t,a_t) - w_k(s_t)\big] \\
&= \sum_{k=0}^{K(T)-1}\left[\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}w_k](s_t,a_t) - w_k(s_{t+1})\big] - w_k(s_{t_k}) + w_k(s_{t_{k+1}})\right] \\
&\le \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}w_k](s_t,a_t) - w_k(s_{t+1})\big] + K(T)\cdot D \\
&\le D\sqrt{2T\log(1/\delta)} + K(T)\cdot D \\
&= \widetilde{O}(D\sqrt{T}) + \widetilde{O}(Dd),
\end{aligned}$$

where the first inequality holds since $|w_k(\cdot)| \le D/2$, the second one holds due to the definition of $\mathcal{E}_2$. For term $I_1$,

$$\begin{aligned}
I_1 &= \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big[[\mathbb{P}_k w_k](s_t,a_t) - [\mathbb{P}w_k](s_t,a_t)\big] \\
&= \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big\langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \boldsymbol{\phi}_{w_k}(s_t,a_t)\big\rangle \\
&\le \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big(\|\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_t} + \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_t}\big)\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}} \\
&\le 2\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\big(\|\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{t_k}} + \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{t_k}}\big)\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}} \\
&\le 4\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\widehat{\beta}_{t_k}\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}} \\
&\le 4\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\widehat{\beta}_t\bar{\sigma}_t\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}.
\end{aligned}$$

The first inequality is due to first applying Cauchy-Schwartz inequality and then the triangle inequality. The second is due to $\det(\widehat{\boldsymbol{\Sigma}}_t) \le 2\det(\widehat{\boldsymbol{\Sigma}}_{t_k})$ and Lemma A.6. The third is due to event $\mathcal{E}_0$. The last is due to the fact that $\{\widehat{\beta}_t\}_{t\ge 0}$ is increasing.

Meanwhile, for each term in $I_1$, we also have that due to $|w_k(s)| \le D/2$,

$$[\mathbb{P}_k w_k](s_t,a_t) - [\mathbb{P}w_k](s_t,a_t) \le D.$$

Therefore, we have

$$\begin{aligned}
I_1 &\le \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\min\left\{D, 4\widehat{\beta}_t\bar{\sigma}_t\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\} \\
&\le \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}4\widehat{\beta}_t\bar{\sigma}_t\min\left\{1, \big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\} \\
&\le 4\widehat{\beta}_T\sqrt{\underbrace{\sum_{t=1}^{T}(\bar{\sigma}_t)^2}_{J_1}}\sqrt{\underbrace{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\left\{1, \big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\big\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\}}_{J_2}}.
\end{aligned}$$

The second inequality is due to the fact $D \leq 4\widehat{\beta}_t\bar{\sigma}_t$. The third is due to Cauchy-Schwartz inequality.

Note that by Lemma A.5, it is clear that

$$J_2 \leq 2d\log(1+T/\lambda).$$

For term $J_1$,

$$\begin{aligned}
J_1 &= \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} \max\{D^2/d, [\bar{\mathbb{V}}_t w_k](s_t,a_t) + E_t\} \\
&\leq TD^2/d + \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} [\bar{\mathbb{V}}_t w_k](s_t,a_t) + \sum_{t=1}^{T} E_t \\
&\leq TD^2/d + \sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1} [\mathbb{V} w_k](s_t,a_t) + 2\sum_{t=1}^{T} E_t \\
&\leq TD^2/d + (D^2/4)\sqrt{2T\log(1/\delta)} + (K(T)+1)(D^2/4) + 2DT + D^2\widehat{\beta}_T\sqrt{T2d\log(1+T/\lambda)} \\
&\quad + \widetilde{\beta}_T\sqrt{2Td\log(1+TD^2/4d\lambda)} + D^2\check{\beta}_T\sqrt{2Td\log(1+T/\lambda)}.
\end{aligned}$$

The second inequality uses Lemma B.1. The third uses Lemma B.4 and Lemma B.5.

Now, based on Lemma B.2 we have $K(T) = \widetilde{O}(d)$. By definition, we have

$$\begin{aligned}
\widehat{\beta}_T &= \widetilde{O}(\sqrt{d}) \\
\check{\beta}_T &= \widetilde{O}(d) \\
\widetilde{\beta}_T &= \widetilde{O}(D^2\sqrt{d}),
\end{aligned}$$

if we set $\lambda = B^{-2}$.

This means we can express $I_1$ in Big-O notation term by term as:

$$\begin{aligned}
J_1 &= \widetilde{O}(TD^2/d) + \widetilde{O}(D^2\sqrt{T}) + \widetilde{O}(D^2 d) + \widetilde{O}(DT) + \widetilde{O}(D^2 d\sqrt{T}) + \widetilde{O}(D^2 d\sqrt{T}) + \widetilde{O}(D^2 d^{3/2}\sqrt{T}) \\
&= \widetilde{O}(TD^2/d) + \widetilde{O}(DT) + \widetilde{O}(D^2 d^{3/2}\sqrt{T}).
\end{aligned}$$

We have

$$\begin{aligned}
I_1 &= \widetilde{O}(\sqrt{d}) \cdot \sqrt{\widetilde{O}(TD^2/d) + \widetilde{O}(DT) + \widetilde{O}(D^2 d^{3/2}\sqrt{T})} \cdot \sqrt{\widetilde{O}(d)} \\
&= \widetilde{O}(D\sqrt{dT}) + \widetilde{O}(d\sqrt{DT}) + \widetilde{O}(Dd^{7/4}T^{1/4}).
\end{aligned}$$

Finally, by setting $\epsilon = 1/\sqrt{T}$, the regret is upper bounded as

$$\begin{aligned}
\text{Regret}(T) &= \mathcal{O}(\sqrt{T}) + \widetilde{O}(D\sqrt{dT}) + \widetilde{O}(d\sqrt{DT}) + \widetilde{O}(Dd^{7/4}T^{1/4}) + \widetilde{O}(D\sqrt{T}) + \widetilde{O}(Dd) \\
&= \widetilde{O}(D\sqrt{dT}) + \widetilde{O}(d\sqrt{DT}) + \widetilde{O}(Dd^{7/4}T^{1/4}).
\end{aligned}$$

As long as $d \geq D$ and $T \geq D^2 d^3$, we have

$$\text{Regret}(T) = \widetilde{O}(d\sqrt{DT}).$$

□

# C Proof of Theorem 5.5

## C.1 Construction of Hard-to-learn MDPs

Here we describe the construction of the hard-to-learn MDPs $M(\mathcal{S}, \mathcal{A}, r, \mathbb{P}_{\boldsymbol{\theta}})$ for our lower bound proof (illustrated in Figure 2). The state space $\mathcal{S}$ consists of two states $x_0, x_1$. The action space $\mathcal{A}$ consists of $2^{d-1}$ vectors $\mathbf{a} \in \mathbb{R}^{d-1}$

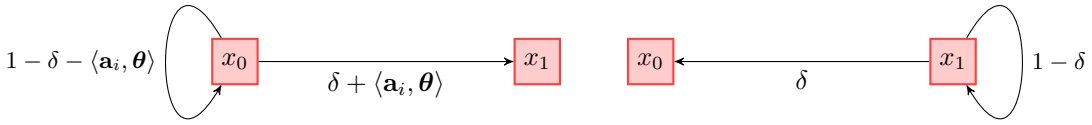


Figure 2: Illustration of the hard-to-learn linear mixture MDP considered in Theorem 5.5. The left figure demonstrates the state transition probability starting from $x_0$ with some specific action $\mathbf{a}_i$. The right figure demonstrates the state transition probability starting from $x_1$ with any action.

whose coordinates are 1 or $-1$. The reward function $r$ satisfies that $r(x_0, \mathbf{a}) = 0$ and $r(x_1, \mathbf{a}) = 1$ for any $\mathbf{a} \in \mathcal{A}$. The probability transition function $\mathbb{P}_{\boldsymbol{\theta}}$ is parameterized by a $(d-1)$-dimensional vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which is defined as

$$\mathbb{P}_{\boldsymbol{\theta}}(x_0|x_0, \mathbf{a}) = 1 - \delta - \langle \mathbf{a}, \boldsymbol{\theta}\rangle, \qquad \mathbb{P}_{\boldsymbol{\theta}}(x_1|x_0, \mathbf{a}) = \delta + \langle \mathbf{a}, \boldsymbol{\theta}\rangle,$$
$$\mathbb{P}_{\boldsymbol{\theta}}(x_0|x_1, \mathbf{a}) = \delta, \qquad \mathbb{P}_{\boldsymbol{\theta}}(x_1|x_1, \mathbf{a}) = 1 - \delta,$$
$$\boldsymbol{\Theta} = \{-\Delta/(d-1), \Delta/(d-1)\}^{d-1},$$

where $\delta$ and $\Delta$ are positive parameters that need to be determined in later proof. We set $\delta = 1/D$, and $\Delta$ as $\Delta = (1/45\sqrt{2\log 2/5})d/\sqrt{DT}$. It can be verified that $M$ is indeed a linear kernel MDP with the feature mapping $\boldsymbol{\phi}(s'|s, a)$ defined as follows:

$$\boldsymbol{\phi}(x_0|x_0, \mathbf{a}) = \begin{pmatrix} -\alpha\mathbf{a} \\ \beta(1-\delta) \end{pmatrix}, \boldsymbol{\phi}(x_1|x_0, \mathbf{a}) = \begin{pmatrix} \alpha\mathbf{a} \\ \beta\delta \end{pmatrix}, \boldsymbol{\phi}(x_0|x_1, \mathbf{a}) = \begin{pmatrix} \mathbf{0} \\ \beta\delta \end{pmatrix}, \boldsymbol{\phi}(x_1|x_1, \mathbf{a}) = \begin{pmatrix} \mathbf{0} \\ \beta(1-\delta) \end{pmatrix},$$

where $\alpha = \sqrt{\Delta/[(d-1)(1+\Delta)]}$, $\beta = \sqrt{1/(1+\Delta)}$, and the vector $\widetilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}^\top/\alpha, 1/\beta)^\top \in \mathbb{R}^d$. We can verify that $\boldsymbol{\phi}$ and $\widetilde{\boldsymbol{\theta}}$ satisfy the requirements of $B$-bounded linear mixture MDP. In detail, we have

$$\|\widetilde{\boldsymbol{\theta}}\|_2^2 = \frac{\|\boldsymbol{\theta}\|_2^2}{\alpha^2} + \frac{1}{\beta^2} = (1+\Delta)^2 \le B^2,$$

as long as $\Delta \le \sqrt{B} - 1$. In addition, for any function $F : \mathcal{S} \to [0, 1]$, we have

$$\|\boldsymbol{\phi}_F(x_0, \mathbf{a})\|_2^2 = \alpha^2\|\mathbf{a}\|_2^2(F(x_1) - F(x_0))^2 + (\beta(1-\delta)F(x_0) + \beta\delta F(x_1))^2 \le (d-1)\alpha^2 + \beta^2 = 1.$$

Therefore, our defined MDP is indeed a $B$-bounded linear mixture MDP.

**Remark C.1.** Similar to Zhou et al. (2021a), our lower bound can also imply a lower bound for a related MDP class called *linear MDPs* (Yang and Wang, 2019a; Jin et al., 2019), which assumes that $\mathbb{P}(s'|s, a) = \langle \boldsymbol{\psi}(s, a), \boldsymbol{\mu}(s')\rangle$ and $r(s, a) = \langle \boldsymbol{\psi}(s, a), \boldsymbol{\xi}\rangle$. We construct $\boldsymbol{\psi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\xi}$ as follows:

$$\boldsymbol{\psi}(s, a) = \begin{cases} (\alpha\mathbf{a}^\top, \beta, 0)^\top & s = x_0 \\ (0, 0, 1) & s = x_1 \end{cases}, \boldsymbol{\mu}(s') = \begin{cases} (-\boldsymbol{\theta}^\top/\alpha, (1-\delta)/\beta), \delta)^\top & s' = x_0 \\ (\boldsymbol{\theta}^\top/\alpha, \delta/\beta, 1-\delta)^\top & s' = x_1 \end{cases}, \boldsymbol{\xi} = (\mathbf{0}^\top, 1)^\top.$$

It can be verified that such a feature mapping $\boldsymbol{\phi}, \boldsymbol{\mu}$ and parameters $\boldsymbol{\xi}$ satisfy the requirements of a linear MDP with $(d+1)$-dimension feature mapping. Meanwhile, the MDP $\langle \boldsymbol{\psi}(s, a), \boldsymbol{\mu}(s')\rangle$ has exactly the same form as the linear mixture MDPs proposed in Theorem 5.5. Therefore, the lower bound in Theorem 5.5 can also be applied to infinite-horizon average-reward linear MDPs, which are studied by Wei et al. (2020a). This also suggests that there still exists a gap between the best upper bound (Wei et al., 2020a) and lower bound in the linear MDP setting.

### C.2 Proof of the Lower Bound in Theorem 5.5

Given the example we constructed above (shown in Figure 2), it is easy to see that the optimal policy is to choose the action $\mathbf{a}$ satisfying $\langle \mathbf{a}, \boldsymbol{\theta}\rangle = \Delta$, namely each coordinate of $\mathbf{a}$ has the same sign as $\boldsymbol{\theta}$'s.

Given the optimal policy, it is clear that the stationary distribution is

$$\boldsymbol{\mu} = \begin{bmatrix} \dfrac{\delta}{2\delta+\Delta} & \dfrac{\delta+\Delta}{2\delta+\Delta} \end{bmatrix},$$

and the optimal average reward is $\rho^* = (\delta+\Delta)/(2\delta+\Delta)$.

In the construction, we leave the two parameters $\delta$ and $\Delta$ unspecified. Now we set $\delta = 1/D$. From state $x_1$ to $x_0$, it is clear that any policy has only one action and the expected travel time is $1/\delta = D$. From state $x_0$ to $x_1$, there always exists an policy that chooses the action $\mathbf{a}$ that has the same sign coordinate-wise, and in that case the transition probability from $x_0$ to $x_1$ is $\delta+\Delta$, which indicates the expected travel time is smaller then $D$. From the argument above, we know the MDP has a diameter of $D$.

The choice of $\Delta$ is $\Delta = (1/45\sqrt{2\log 2/5})d/\sqrt{DT}$; the motivation will be revealed later in the proof.

In the following, we use $\text{Regret}_{\boldsymbol{\theta}}(T)$ to denote the regret $\text{Regret}(M_{\boldsymbol{\theta}}, A, s, T)$, where $A$ is a deterministic algorithm. As argued in Auer et al. (2002), it is sufficient to only consider deterministic policies. Let $\mathcal{P}_{\boldsymbol{\theta}}(\cdot)$ denote the distribution over $\mathcal{S}^T$, where $s_1 = x_0$, $s_{t+1} \sim \mathbb{P}_{\boldsymbol{\theta}}(\cdot|s_t, a_t)$, $a_t$ is decided by $A$. Let $\mathbb{E}_{\boldsymbol{\theta}}$ denote the expectation w.r.t. distribution $\mathcal{P}_{\boldsymbol{\theta}}$. Denote $N_1, N_0, N_0^{\mathbf{a}}$ as the random variables of the times state $x_1$ is visited, the times state $x_0$ is visited and the times state $x_0$ is visited and $\mathbf{a}$ is chosen. We further define $N_0^{\mathcal{V}}$ for some subset $\mathcal{V} \subset \mathcal{A}$ as the random variable of the times state $x_0$ is visited, and the action $\mathbf{a}$ belongs to $\mathcal{V}$.

**Lemma C.2.** Suppose $2\Delta < \delta$ and $(1-\delta)/\delta < T/5$, then for $\mathbb{E}_{\boldsymbol{\theta}} N_1$ and $\mathbb{E}_{\boldsymbol{\theta}} N_0$, we have

$$\mathbb{E}_{\boldsymbol{\theta}} N_1 \le \frac{T}{2} + \frac{1}{2\delta}\sum_{\mathbf{a}} \langle \mathbf{a}, \boldsymbol{\theta}\rangle \mathbb{E}_{\boldsymbol{\theta}} N_0^{\mathbf{a}}, \ \ \mathbb{E}_{\boldsymbol{\theta}} N_0 \le 4T/5.$$

*Proof.* See Section D.7. □

**Lemma C.3** (Pinsker's inequality, in Jaksch et al. (2010))**.** Denote $\mathbf{s} = \{s_1, \ldots, s_T\} \in \mathcal{S}^T$ as the observed states from step 1 to $T$. Then for any two distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ over $\mathcal{S}^T$ and any bounded function $f : \mathcal{S}^T \to [0, B]$, we have

$$\mathbb{E}_1 f(\mathbf{s}) - \mathbb{E}_2 f(\mathbf{s}) \le \sqrt{\log 2/2} B \sqrt{\text{KL}(\mathcal{P}_2 \| \mathcal{P}_1)},$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are expectations with respect to $\mathcal{P}_1$ and $\mathcal{P}_2$.

**Lemma C.4.** Suppose that $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ only differs from $j$-th coordinate, $2\Delta < \delta \le 1/3$. Then we have the following bound for the KL divergence between $\mathcal{P}_{\boldsymbol{\theta}}$ and $\mathcal{P}_{\boldsymbol{\theta}'}$:

$$\text{KL}(\mathcal{P}_{\boldsymbol{\theta}'} \| \mathcal{P}_{\boldsymbol{\theta}}) \le \frac{16\Delta^2}{(d-1)^2\delta} \mathbb{E}_{\boldsymbol{\theta}} N_0.$$

*Proof.* See Section D.8. □

*Proof of Theorem 5.5.* We have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}[\text{Regret}_{\boldsymbol{\theta}}(T)] &:= T\rho^* - \mathbb{E}_{\boldsymbol{\theta}}\bigg[\sum_{t=1}^{T} r(s_t, a_t)\bigg] \\ &= T\rho^* - \mathbb{E}_{\boldsymbol{\theta}}[N_1]. \end{aligned}$$

Averaging over all possible choice of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have

$$\frac{1}{|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}}[\text{Regret}_{\boldsymbol{\theta}}(T)] = T\rho^* - \frac{1}{|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}}[N_1].$$

Following Lemma C.2, we first have

$$\begin{aligned}\frac{1}{|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}}[N_1] &\le \frac{T}{2}+\frac{1}{2\delta|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}}\sum_{\mathbf{a}}\langle \mathbf{a},\boldsymbol{\theta}\rangle \mathbb{E}_{\boldsymbol{\theta}}N_0^{\mathbf{a}}\\ &= \frac{T}{2}+\frac{1}{2\delta|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}}\sum_{\mathbf{a}}\frac{\Delta}{d-1}\sum_{j=1}^{d-1}\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}\mathbb{E}_{\boldsymbol{\theta}}N_0^{\mathbf{a}}\\ &= \frac{T}{2}+\frac{1}{2\delta|\boldsymbol{\Theta}|}\frac{\Delta}{d-1}\sum_{j=1}^{d-1}\sum_{\boldsymbol{\theta}}\sum_{\mathbf{a}}\mathbb{E}_{\boldsymbol{\theta}}\big[\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\big]. \end{aligned} \tag{C.1}$$

For a fixed coordinate $j$, consider $\boldsymbol{\theta}'$ that only differs with $\boldsymbol{\theta}$ at its $j$-th coordinate. We have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}}\big[\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\big]+\mathbb{E}_{\boldsymbol{\theta}'}\big[\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}'_j)\}N_0^{\mathbf{a}}\big]\\ &=\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0^{\mathbf{a}}\big]+\mathbb{E}_{\boldsymbol{\theta}}\big[\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\big]-\mathbb{E}_{\boldsymbol{\theta}'}\big[\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\big],\end{aligned}$$

since $\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}'_j)\}=1-\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)\neq\mathrm{sign}(\boldsymbol{\theta}_j)\}$.

Summing the equation above over $\boldsymbol{\Theta}$ and $\mathcal{A}$, we have

$$\begin{aligned}&2\sum_{\boldsymbol{\theta}}\sum_{\mathbf{a}}\mathbb{E}_{\boldsymbol{\theta}}\big[\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\big]\\ &=\sum_{\boldsymbol{\theta}}\sum_{\mathbf{a}}\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0^{\mathbf{a}}\big]+\sum_{\boldsymbol{\theta}}\Big[\mathbb{E}_{\boldsymbol{\theta}}\Big[\sum_{\mathbf{a}}\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\Big]-\mathbb{E}_{\boldsymbol{\theta}'}\Big[\sum_{\mathbf{a}}\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}N_0^{\mathbf{a}}\Big]\Big]\\ &=\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0\big]+\sum_{\boldsymbol{\theta}}\Big[\mathbb{E}_{\boldsymbol{\theta}}\Big[N_0^{\mathcal{A}_j}\Big]-\mathbb{E}_{\boldsymbol{\theta}'}\Big[N_0^{\mathcal{A}_j}\Big]\Big]\\ &\le\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0\big]+\sum_{\boldsymbol{\theta}}\sqrt{\log 2/2}T\sqrt{\mathrm{KL}(\mathcal{P}_{\boldsymbol{\theta}'}\|\mathcal{P}_{\boldsymbol{\theta}})}\\ &\le\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0\big]+\sum_{\boldsymbol{\theta}}2\sqrt{2\log 2}\frac{T\Delta}{d\sqrt{\delta}}\sqrt{\mathbb{E}_{\boldsymbol{\theta}}[N_0]},\end{aligned} \tag{C.2}$$

where $\mathcal{A}_j$ is the set of all $\mathbf{a}$ which satisfy $\mathbb{1}\{\mathrm{sign}(\mathbf{a}_j)=\mathrm{sign}(\boldsymbol{\theta}_j)\}$. The first equality is by matching each $\boldsymbol{\theta}$ with $\boldsymbol{\theta}'$ that differs from $\boldsymbol{\theta}$ in its $j$-th coordinate, and moving $\sum_{\mathbf{a}}$ inside. The second equality applies the shorthand $\mathcal{A}_j$. The first inequality is due to Lemma C.3. The last is due to Lemma C.4.

Substituting (C.2) into (C.1), we have

$$\begin{aligned}\frac{1}{|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}}[N_1] &\le \frac{T}{2}+\frac{1}{4\delta|\boldsymbol{\Theta}|}\frac{\Delta}{d-1}\sum_{j=1}^{d-1}\sum_{\boldsymbol{\theta}}\Big[\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0\big]+2\sqrt{2\log 2}\frac{T\Delta}{d\sqrt{\delta}}\sqrt{\mathbb{E}_{\boldsymbol{\theta}}[N_0]}\Big]\\ &= \frac{T}{2}+\frac{\Delta}{4\delta|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}}\Big[\mathbb{E}_{\boldsymbol{\theta}'}\big[N_0\big]+2\sqrt{2\log 2}\frac{T\Delta}{d\sqrt{\delta}}\sqrt{\mathbb{E}_{\boldsymbol{\theta}}[N_0]}\Big]\\ &\le \frac{T}{2}+\frac{\Delta}{4\delta}\Big[\frac{4T}{5}+2\sqrt{2\log 2}\frac{T\Delta}{d\sqrt{\delta}}\frac{2\sqrt{T}}{\sqrt{5}}\Big]\\ &= \frac{T}{2}+\frac{\Delta T}{5\delta}+\sqrt{2\log 2/5}\frac{\Delta^2T^{3/2}}{d\delta^{3/2}},\end{aligned}$$

where the second inequality is due to Lemma C.2.

This further leads to

$$\begin{aligned}\frac{1}{|\Theta|}\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}}[\text{Regret}_{\boldsymbol{\theta}}(T)] &= T\rho^* - \frac{1}{|\Theta|}\sum_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\theta}}[N_1] \\ &\geq T\cdot\frac{\delta+\Delta}{2\delta+\Delta} - \frac{T}{2} - \frac{\Delta T}{5\delta} - \sqrt{2\log 2/5}\frac{\Delta^2 T^{3/2}}{d\delta^{3/2}} \\ &= \frac{\Delta(\delta-2\Delta)}{5\delta(4\delta+2\Delta)}\cdot T - \sqrt{2\log 2/5}\frac{\Delta^2 T^{3/2}}{d\delta^{3/2}} \\ &\geq \frac{2}{45\delta}\cdot\Delta T - \sqrt{2\log 2/5}\cdot\frac{\Delta^2 T^{3/2}}{d\delta^{3/2}} \\ &= \frac{1}{2025\sqrt{2\log 2/5}}\cdot d\sqrt{DT} \\ &> \frac{1}{2025}\cdot d\sqrt{DT},\end{aligned}$$

where the second inequality requires $0 < 4\Delta \leq \delta$; the last equality is due to the setting $\delta = D^{-1}$ and $\Delta = (1/45\sqrt{2\log 2/5})d/\sqrt{DT}$. This further requires that $T \geq 16d^2D/2025$. □

# D Proof of Supporting Lemmas

## D.1 Proof of Lemma A.2

*Proof of Lemma A.2.* Recall the definition of $\boldsymbol{\theta}_k$ in Algorithm 1, we have

$$\boldsymbol{\theta}_k = \Big(\lambda\mathbf{I} + \sum_{j=0}^{k-1}\sum_{i=t_j}^{t_{j+1}-1}\boldsymbol{\phi}_{w_j}(s_i,a_i)\boldsymbol{\phi}_{w_j}(s_i,a_i)^\top\Big)^{-1}\Big(\sum_{j=0}^{k-1}\sum_{i=t_j}^{t_{j+1}-1}\boldsymbol{\phi}_{w_j}(s_i,a_i)w(s_{i+1})\Big).$$

It is worth noting that for any $0 \leq j \leq k-1$ and $t_j \leq i \leq t_{j+1}-1$,

$$\begin{aligned}[\mathbb{P}w_j](s_i,a_i) &= \int_{s'}\mathbb{P}(s'|s_i,a_i)w_j(s_i,a_i)ds' \\ &= \int_{s'}\langle\boldsymbol{\phi}(s'|s_i,a_i),\boldsymbol{\theta}^*\rangle w_j(s')ds' \\ &= \Big\langle\int_{s'}\boldsymbol{\phi}(s'|s_i,a_i)w_j(s'),\boldsymbol{\theta}^*\Big\rangle \\ &= \langle\boldsymbol{\phi}_{w_j}(s_i,a_i),\boldsymbol{\theta}^*\rangle,\end{aligned} \tag{D.1}$$

thus $\{w_j(s_{i+1}) - \langle\boldsymbol{\phi}_{w_j}(s_i,a_i),\boldsymbol{\theta}^*\rangle\}$ forms a martingale difference sequence. Besides, since $|w(s)| \leq D/2$ for any $s$, then $w_j(s_{i+1}) - \langle\boldsymbol{\phi}_{w_j}(s_i,a_i),\boldsymbol{\theta}^*\rangle$ is a sequence of $D$-subgaussian random variables with zero means. Meanwhile, we have $\|\boldsymbol{\phi}_{w_j}(s_i,a_i)\|_2 \leq D$ and $\|\boldsymbol{\theta}^*\|_2 \leq B$ by Definition 3.2. By Theorem 2 in Abbasi-Yadkori et al. (2011), we have that with probability at least $1-\delta$, $\boldsymbol{\theta}^*$ belongs to the following set for all $1 \leq k \leq K$:

$$\Big\{\boldsymbol{\theta} : \Big\|\boldsymbol{\Sigma}_{t_k}^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_k)\Big\|_2 \leq D\sqrt{\log\Big(\frac{\lambda + t_k D^2}{\delta\lambda}\Big)} + \sqrt{\lambda}B\Big\}. \tag{D.2}$$

Finally, by the definition of $\widehat{\beta}_t$ and the fact that $\langle\boldsymbol{\theta}^*,\boldsymbol{\phi}(s'|s,a)\rangle = \mathbb{P}(s'|s,a)$ for all $(s,a)$, we draw the conclusion that $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{t_k}$ for $1 \leq k \leq K$. □

### D.2 Proof of Lemma A.3

*Proof of Lemma A.3.* For simplicity, we denote $K = K(T)$. Note that $\det(\mathbf{\Sigma}_0) = \lambda^d$. We further have

$$\begin{aligned}\|\mathbf{\Sigma}_T\|_2 &= \bigg\|\lambda\mathbf{I} + \sum_{k=0}^{K-1}\sum_{t=t_k}^{t_{k+1}-1}\boldsymbol{\phi}_{w_k}(s_t,a_t)\boldsymbol{\phi}_{w_k}(s_t,a_t)^\top\bigg\|_2 \\ &\le \lambda + \sum_{k=0}^{K-1}\sum_{t=t_k}^{t_{k+1}-1}\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\big\|_2^2 \\ &\le \lambda + TD^2, \end{aligned} \tag{D.3}$$

where the first inequality holds due to the triangle inequality, the second inequality holds due to the fact $w_k(s) \le D/2$ and Definition 3.2. (D.3) suggests that $\det(\mathbf{\Sigma}_T) \le (\lambda + TdD^2)^d$. Therefore, we have

$$(\lambda + TD^2)^d \ge \det(\mathbf{\Sigma}_T) \ge \det(\mathbf{\Sigma}_{t_{K-1}-1}) \ge 2^{K-1}\det(\mathbf{\Sigma}_{t_0-1}) = 2^{K-1}\lambda^d, \tag{D.4}$$

where the second inequality holds since $\mathbf{\Sigma}_T \succeq \mathbf{\Sigma}_{t_{K-1}-1}$, the third inequality holds due to the fact that $\det(\mathbf{\Sigma}_{t_k-1}) \ge 2\det(\mathbf{\Sigma}_{t_{k-1}-1})$ by the update rule in Algorithm 1. (D.4) suggests

$$K \le d\log\frac{2\lambda + 2TD^2}{\lambda}.$$

□

### D.3 Proof of Lemma B.2

*Proof of Lemma B.2.* For simplicity, we denote $K = K(T)$. Note that $\det(\widehat{\mathbf{\Sigma}}_1) = \lambda^d$. We further have

$$\|\widehat{\mathbf{\Sigma}}_{t_K}\|_2 = \bigg\|\lambda\mathbf{I} + \sum_{t=1}^{T}\boldsymbol{\phi}_{w_k}(s_t,a_t)\boldsymbol{\phi}_{w_k}(s_t,a_t)^\top/\bar{\sigma}_t^2\bigg\|_2 \le \lambda + \sum_{t=1}^{T}\big\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\big\|_2^2 \le \lambda + Td,$$

where the first inequality holds due to the triangle inequality, the second inequality holds because $w_k(s) \le D$ and $\bar{\sigma}_t \ge D/\sqrt{d}$. This suggests that $\det(\widehat{\mathbf{\Sigma}}_{t_K}) \le (\lambda + Td)^d$. Therefore, we have

$$(\lambda + Td)^d \ge \det(\mathbf{\Sigma}_{t_K}) \ge \det(\mathbf{\Sigma}_{t_{K-1}}) \ge 2^{K-1}\det(\mathbf{\Sigma}_{t_0}) = 2^{K-1}\lambda^d,$$

where the second inequality holds since $\mathbf{\Sigma}_T \succeq \mathbf{\Sigma}_{t_{K-1}-1}$, the third inequality holds due to the fact that $\det(\mathbf{\Sigma}_{t_k-1}) \ge 2\det(\mathbf{\Sigma}_{t_{k-1}-1})$ by the update rule in Algorithm 1 with OPTION 2. This suggests

$$K \le 2d\log(1 + dT/\lambda).$$

□

### D.4 Proof of Lemma B.3

*Proof of Lemma B.3.* In fact we are able to prove a stronger result:

$$\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_t \cap \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t \cap \mathcal{B},$$

where the two additional sets are defined as

$$\begin{aligned}\check{\mathcal{C}}_t &:= \Big\{\boldsymbol{\theta} : \big\|\check{\mathbf{\Sigma}}_t^{1/2}(\boldsymbol{\theta} - \check{\boldsymbol{\theta}}_t)\big\| \le \check{\beta}_t\Big\} \\ \widetilde{\mathcal{C}}_t &:= \Big\{\boldsymbol{\theta} : \big\|\widetilde{\mathbf{\Sigma}}_t^{1/2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_t)\big\| \le \widetilde{\beta}_t\Big\}.\end{aligned}$$

For any $1 \le t \le T$, we always have $k$ such that $t_k \le t < t_{k+1}$. We start with the following inequality:

$$\begin{aligned}
\big|[\bar{\mathbb{V}}_t w_k](s_t,a_t) - [\mathbb{V}w_k](s_t,a_t)\big| &= \Big| \min\Big\{D^2/4, \langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \widetilde{\boldsymbol{\theta}}_t\rangle\Big\} - \langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \boldsymbol{\theta}^*\rangle \\
&\quad + \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle^2 - \Big[\min\Big\{D/2, \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_t\rangle\Big\}\Big]^2\Big| \\
&\le \underbrace{\Big|\min\Big\{D^2/4, \langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \widetilde{\boldsymbol{\theta}}_t\rangle\Big\} - \langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \boldsymbol{\theta}^*\rangle\Big|}_{I_1} \\
&\quad + \underbrace{\Big|\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle^2 - \Big[\min\Big\{D/2, \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_t\rangle\Big\}\Big]^2\Big|}_{I_2},
\end{aligned}$$

where the inequality is by the triangle inequality.

For $I_1$, we have

$$\begin{aligned}
I_1 &\le \Big|\langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \widetilde{\boldsymbol{\theta}}_t\rangle - \langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \boldsymbol{\theta}^*\rangle\Big| \\
&= \Big|\langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\rangle\Big| \\
&\le \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k^2}(s_t,a_t)\Big\| \cdot \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{1/2}(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\Big\|,
\end{aligned}$$

where the first inequality is due to $\langle \boldsymbol{\phi}_{w_k^2}(s_t,a_t), \boldsymbol{\theta}^*\rangle = \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k^2(s')] \in [0, D^2/4]$, and the last inequality is due to Cauchy-Schwartz inequality. Also, it is clear $I_1 \le D^2/4$.

Similarly, for $I_2$, we have

$$\begin{aligned}
I_2 &= \Big|\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle + \min\Big\{D/2, \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_t\rangle\Big\}\Big| \\
&\quad \cdot \Big|\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle - \Big[\min\Big\{D/2, \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_t\rangle\Big\}\Big]\Big| \\
&\le D \cdot \Big|\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle - \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_t\rangle\Big| \\
&= D \cdot \Big|\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^* - \boldsymbol{\theta}_t\rangle\Big| \\
&\le D \cdot \Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k}(s_t,a_t)\Big\| \cdot \Big\|\widehat{\boldsymbol{\Sigma}}_t^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\Big\|,
\end{aligned}$$

where the first equality is by $a^2 - b^2 = (a+b)(a-b)$, and the following reasoning is the same as $I_1$. The only additional fact used in the first inequality is $\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle \in [0, D/2]$ and $\min\big\{D/2, \langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_t\rangle\big\} \in [0, D/2]$. Also, it is clear $I_2 \le D^2/4$.

The two terms combined together gives

$$\begin{aligned}
\big|[\bar{\mathbb{V}}_t w_k](s_t,a_t) - [\mathbb{V}w_k](s_t,a_t)\big| &\le \min\Big\{D^2/4, \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k^2}(s_t,a_t)\Big\| \cdot \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{1/2}(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\Big\|\Big\} \\
&\quad + \min\Big\{D^2/4, D \cdot \Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k}(s_t,a_t)\Big\| \cdot \Big\|\widehat{\boldsymbol{\Sigma}}_t^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\Big\|\Big\}. \qquad \text{(D.5)}
\end{aligned}$$

Now, we first show that with probability $1-\delta$, for all $t$, $\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t$. To show this, we apply Lemma B.1. By setting $\mathbf{x}_t = \bar{\sigma}_t^{-1}\boldsymbol{\phi}_{w_k}(s_t,a_t)$ and $\eta_t = \bar{\sigma}_t^{-1}w_k(s_{t+1}) - \bar{\sigma}_t^{-1}\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}^*\rangle$, $\mathcal{G}_t = \mathcal{F}_t$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}^*$, $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t\rangle + \eta_t$, $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^t \mathbf{x}_i\mathbf{x}_i^\top$, $\mathbf{b}_t = \sum_{i=1}^t \mathbf{x}_i y_i$ and $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1}\mathbf{b}_t$, we have $y_t = \bar{\sigma}_t^{-1}w_k(s_{t+1})$ and $\boldsymbol{\mu}_t = \widehat{\boldsymbol{\theta}}_t$. Moreover, we have

$$\|\mathbf{x}_t\|_2 \le \sqrt{d}/2, |\eta_t| \le \sqrt{d}, \mathbb{E}[\eta_t|\mathcal{G}_t] = 0, \mathbb{E}[\eta_t^2|\mathcal{G}_t] = d.$$

Therefore, by Lemma B.1, we have with probability $1-\delta$, for all $t \in [T]$,

$$\|\widehat{\boldsymbol{\Sigma}}_t^{1/2}(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\|_2 \le 8d\sqrt{\log(1+t/4\lambda)\log(4t^2/\delta)} + 4\sqrt{d}\log(4t^2/\delta) + \sqrt{\lambda}B = \check{\beta}_t.$$

This means that with probability $1-\delta$, for all $t$, $\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t$.

The same argument can be applied again, except that now we focus on the squared function $w_k^2$. This gives

$$\|\widetilde{\boldsymbol{\Sigma}}_t^{1/2}(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\|_2 \le 8(D^2/4)\sqrt{d\log(1+tD^2/4\lambda d\lambda)\log(4t^2/\delta)} + 4(D^2/4)\log(4t^2/\delta) + \sqrt{\lambda}B = \widetilde{\beta}_t.$$

This means that with probability $1-\delta$, for all $t$, $\boldsymbol{\theta}^* \in \widetilde{\mathcal{C}}_t$.

Now we show that $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_t$ with high probability. Let $\mathbf{x}_t = \bar{\sigma}_t^{-1}\boldsymbol{\phi}_{w_k}(s_t, a_t)$, and

$$\eta_t = \bar{\sigma}_t^{-1}\,\mathbb{1}\{\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\}\big[w_k(s_{t+1}) - \langle \boldsymbol{\phi}_{w_k}(s_t, a_t), \boldsymbol{\theta}^*\rangle\big].$$

In this case, it is clear that still we have $\mathbb{E}[\eta_t|\mathcal{G}_t] = 0$, $|\eta_t| \le \sqrt{d}$, $\|\mathbf{x}_t\|_2 \le \sqrt{d}$. Also,

$$\begin{aligned}
\mathbb{E}[\eta_t^2|\mathcal{G}_t] &= \bar{\sigma}_t^{-2}\,\mathbb{1}\{\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\}[\mathbb{V}w_t](s_t, a_t) \\
&\le \bar{\sigma}_t^{-2}\,\mathbb{1}\{\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\}\Bigg[[\bar{\mathbb{V}}_t w_t](s_t, a_t) \\
&\qquad + \min\left\{D^2/4, \left\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k^2}(s_t, a_t)\right\| \cdot \left\|\widetilde{\boldsymbol{\Sigma}}_t^{1/2}(\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\right\|\right\} \\
&\qquad + \min\left\{D^2/4, D \cdot \left\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k}(s_t, a_t)\right\| \cdot \left\|\widehat{\boldsymbol{\Sigma}}_t^{1/2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t)\right\|\right\}\Bigg] \\
&\le \bar{\sigma}_t^{-2}\Bigg[[\bar{\mathbb{V}}_t w_t](s_t, a_t) + \min\left\{D^2/4, \left\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k^2}(s_t, a_t)\right\|\widetilde{\beta}_t\right\} \\
&\qquad + \min\left\{D^2/4, D\check{\beta}_t \cdot \left\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\boldsymbol{\phi}_{w_k}(s_t, a_t)\right\|\right\}\Bigg] \\
&= 1,
\end{aligned}$$

where the first inequality is due to (D.5) and the second inequality is due to first, the event that $\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t$. The last equality is by the definition of $\bar{\sigma}_t$.

Again by Lemma B.1, we have that for all $t \in [T]$,

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} \le 8\sqrt{d\log(1+t/4\lambda)\log(4t^2/\delta)} + 4\sqrt{d}\log(4t^2/\delta) + \sqrt{\lambda}B = \widehat{\beta}_t.$$

Now, denote the event when $\{\forall t \in [T], \boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\}$ and the inequality above holds as $\mathcal{E}_0$. By union bound, we have $\mathbb{P}(\mathcal{E}_0) \ge 1 - 3\delta$.

It is clear that under $\mathcal{E}_0$, we have $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_t$ for all $t$ because under event $\mathcal{E}_0$,

$$\begin{aligned}
y_t &= \langle \bar{\sigma}_t^{-1}\boldsymbol{\phi}_{w_k}(s_t, a_t), \boldsymbol{\theta}^*\rangle + \bar{\sigma}_t^{-1}\,\mathbb{1}\{\boldsymbol{\theta}^* \in \check{\mathcal{C}}_t \cap \widetilde{\mathcal{C}}_t\}\big[w_k(s_{t+1}) - \langle \boldsymbol{\phi}_{w_k}(s_t, a_t), \boldsymbol{\theta}^*\rangle\big] \\
&= \bar{\sigma}_t^{-1} w_k(s_{t+1}),
\end{aligned}$$

so indeed we have $\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_{\widehat{\boldsymbol{\Sigma}}_t} \le \widehat{\beta}_t$.

Also, by the definition of $E_t$, it is clear that under event $\mathcal{E}_0$,

$$\big|[\bar{\mathbb{V}}_t w_k](s_t, a_t) - [\mathbb{V}w_k](s_t, a_t)\big| \le E_t.$$

□

### D.5 Proof of Lemma B.4

*Proof of Lemma B.4.* Part of the proof is inspired by Fruit et al. (2020). We will use $\mathbb{V}_P(w)$ to denote $\mathbb{E}_{s'\sim P(\cdot)}[w(s')^2] - (\mathbb{E}_{s'\sim P(\cdot)}[w(s')])^2$, namely the variance of the random variable $w(s')$ where $s' \sim P(\cdot)$. Some examples are

$$\begin{aligned}
\mathbb{V}_{\mathbb{P}(\cdot|s_t,a_t)}(w_k) &= \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')^2] - (\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])^2, \\
\mathbb{V}_{\mathbb{P}_k(\cdot|s_t,a_t)}(w_k) &= \mathbb{E}_{s'\sim\mathbb{P}_k(\cdot|s_t,a_t)}[w_k(s')^2] - (\mathbb{E}_{s'\sim\mathbb{P}_k(\cdot|s_t,a_t)}[w_k(s')])^2.
\end{aligned}$$

When the context is clear, we may also use short-hands like $\mathbb{E}_p[w(s')]$ to indicate expectation under $p(\cdot)$. The following decomposition is useful:

$$
\begin{aligned}
&\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\mathbb{V}_{\mathbb{P}(\cdot|s_t,a_t)}(w_k)\\
&=\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')^2]-(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])^2\\
&=\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\left[\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')^2]-w_k^2(s_{t+1})\right]\\
&\quad+\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\left[w_k^2(s_{t+1})-(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])^2\right]\\
&=\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\left[\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')^2]-w_k^2(s_{t+1})\right]\\
&\quad+\sum_{k=0}^{K(T)-1}\left[\sum_{t=t_k}^{t_{k+1}-1}\left[w_k^2(s_t)-(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])^2\right]+w_k^2(s_{t_{k+1}})-w_k^2(s_{t_k})\right]\\
&\le\underbrace{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\left[\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')^2]-w_k^2(s_{t+1})\right]}_{I_1}\\
&\quad+\underbrace{\sum_{k=0}^{K(T)-1}\sum_{t=t_k}^{t_{k+1}-1}\left[w_k^2(s_t)-(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])^2\right]}_{I_2}+K(T)\cdot D^2/4.
\end{aligned}
$$

For term $I_1$, since the event $\mathcal{E}_1$ holds, we have

$$I_1\le(D^2/4)\sqrt{2T\log(1/\delta)}.$$

For term $I_2$, we have

$$
\begin{aligned}
I_2&=\sum_{t=1}^{T}\left[w_k^2(s_t)-(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])^2\right]\\
&\le\sum_{t=1}^{T}\left|w_k(s_t)-\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')]\right|\cdot\left|w_k(s_t)+\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')]\right|\\
&\le D\sum_{t=1}^{T}\left|w_k(s_t)-(\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)}[w_k(s')])\right|.
\end{aligned}
$$

Note that, $w_k$, as the output of the Extended Value Iteration, satisfies the following condition(Lemma A.1):

$$|r(s_t,a_t)+\mathbb{E}_{s'\sim\mathbb{P}_k(\cdot|s_t,a_t)}[w_k(s')]-w_k(s_t)-\rho_k|\le\epsilon.$$

Therefore, we can further bound each term in $I_2$ as follows:

$$\begin{aligned}
\left|w_k(s_t) - \mathbb{E}_{\mathbb{P}}[w_k(s')]\right| &= \left|w_k(s_t) - \mathbb{E}_{\mathbb{P}_k}[w_k(s')] + \mathbb{E}_{\mathbb{P}_k}[w_k(s')] - \mathbb{E}_{\mathbb{P}}[w_k(s')]\right| \\
&\le |r(s_t,a_t) + \mathbb{E}_{\mathbb{P}_k}[w_k(s')] - w_k(s_t) - \rho_k| + |r(s_t,a_t) - \rho_k| \\
&\quad + \left|\mathbb{E}_{\mathbb{P}_k}[w_k(s')] - \mathbb{E}_{\mathbb{P}}[w_k(s')]\right| \\
&\le r_{\max} + r_{\max} + \left|\mathbb{E}_{\mathbb{P}_k}[w_k(s')] - \mathbb{E}_{\mathbb{P}}[w_k(s')]\right| \\
&= 2r_{\max} + \left|\langle \boldsymbol{\phi}_{w_k}(s_t,a_t), \boldsymbol{\theta}_k - \boldsymbol{\theta}^*\rangle\right| \\
&\le 2r_{\max} + \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\right\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}} \cdot \left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\right\|_{\widehat{\boldsymbol{\Sigma}}_t} \\
&\le 2r_{\max} + 2\widehat{\beta}_t \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\right\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}.
\end{aligned}$$

Here, the first inequality is due to triangle inequality. The second inequality is due to 1) the reward function( so should the average reward) should lie in $[0, r_{\max}]$ as assumed, and in this paper's setting actually $r_{\max} = 1$. The third inequality is due to Cauchy-Schwartz inequality. The last inequality is due to the assumption $\mathcal{E}_0$ holds. For the second equality, note that $\mathbb{E}_{\mathbb{P}}[w(s')] = \langle \boldsymbol{\phi}_w(s'|s_t,a_t), \boldsymbol{\theta}^*\rangle$.

Also, it is clear that $\left|\mathbb{E}_{\mathbb{P}_k}[w_k(s')] - \mathbb{E}_{\mathbb{P}}[w_k(s')]\right| \le D$. Therefore, term $I_2$ can be bounded as

$$\begin{aligned}
I_2 &\le D \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \left[2r_{\max} + \min\left\{D, \widehat{\beta}_t \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\right\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\}\right] \\
&= 2DT + D \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\left\{D, \widehat{\beta}_t \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)\right\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\} \\
&\le 2DT + D \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \widehat{\beta}_t \bar{\sigma}_t \min\left\{1, \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\right\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\} \\
&\le 2DT + D^2\widehat{\beta}_T \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\left\{1, \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\right\|_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\} \\
&\le 2DT + D^2\widehat{\beta}_T\sqrt{T}\sqrt{\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\left\{1, \left\|\boldsymbol{\phi}_{w_k}(s_t,a_t)/\bar{\sigma}_t\right\|^2_{\widehat{\boldsymbol{\Sigma}}_t^{-1}}\right\}} \\
&\le 2DT + D^2\widehat{\beta}_T\sqrt{T 2d\log(1+T/\lambda)}.
\end{aligned}$$

The second inequality holds because $\widehat{\beta}_t \ge \sqrt{d}$ and $\bar{\sigma}_t \ge D/\sqrt{d}$.The third inequality holds because $\widehat{\beta}_t \le \widehat{\beta}_T$ and $\bar{\sigma}_t \le D$. The fourth inequality is due to Cauchy-Schwartz inequality. The last inequality is from Lemma A.6.

Collecting $I_1$ and $I_2$ gives

$$\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} [\mathbb{V}w_k](s_t,a_t) \le (D^2/4)\sqrt{2T\log(1/\delta)} + (K(T)+1)(D^2/4) + 2DT + D^2\widehat{\beta}_T\sqrt{T2d\log(1+T/\lambda)},$$

given that $\mathcal{E}_0$ and $\mathcal{E}_1$ hold. Using big-O notation we have

$$\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} [\mathbb{V}w_k](s_t,a_t) = \widetilde{O}(DT) + \widetilde{O}(D^2d\sqrt{T}).$$

□

### D.6 Proof of Lemma B.5

*Proof of Lemma B.5.* Directly unroll the definition of $E_t$:

$$\sum_{t=1}^{T} E_t = \underbrace{\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{D^2/4, \widetilde{\beta}_t \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2} \boldsymbol{\phi}_{w_k^2}(s_t, a_t)\Big\|\Big\}}_{I_1} + \underbrace{\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{D^2/4, D\check{\beta}_t \Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2} \boldsymbol{\phi}_{w_k}(s_t, a_t)\Big\|\Big\}}_{I_2}.$$

For term $I_1$,

$$\begin{aligned} I_1 &\leq \widetilde{\beta}_T \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{1, \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2} \boldsymbol{\phi}_{w_k^2}(s_t, a_t)\Big\|\Big\} \\ &\leq \widetilde{\beta}_T \sqrt{T} \sqrt{\sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{1, \Big\|\widetilde{\boldsymbol{\Sigma}}_t^{-1/2} \boldsymbol{\phi}_{w_k^2}(s_t, a_t)\Big\|^2\Big\}} \\ &\leq \widetilde{\beta}_T \sqrt{2Td\log(1 + TD^2/4d\lambda)}, \end{aligned}$$

where the first inequality is due to $\widetilde{\beta}_t \leq \widetilde{\beta}_T$ and $\widetilde{\beta}_t \geq D^2/4$. The second inequality is due to Cauchy-Schwartz inequality. The third is due to Lemma A.5.

Similarly, for $I_2$, we have

$$\begin{aligned} I_2 &= \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{D^2/4, D\check{\beta}_t \bar{\sigma}_t \Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2} \boldsymbol{\phi}_{w_k}(s_t, a_t)/\bar{\sigma}_t\Big\|\Big\} \\ &\leq D^2 \check{\beta}_T \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2} \boldsymbol{\phi}_{w_k}(s_t, a_t)/\bar{\sigma}_t\Big\|\Big\} \\ &\leq D^2 \check{\beta}_T \sqrt{T \sum_{k=0}^{K(T)-1} \sum_{t=t_k}^{t_{k+1}-1} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2} \boldsymbol{\phi}_{w_k}(s_t, a_t)/\bar{\sigma}_t\Big\|^2\Big\}} \\ &\leq D^2 \check{\beta}_T \sqrt{2Td\log(1 + T/\lambda)}, \end{aligned}$$

where the first inequality is due to $\check{\beta}_t \bar{\sigma}_t \geq D$, $\check{\beta}_t \leq \check{\beta}_T$ and $\bar{\sigma}_t \leq D$(all can be verified by the definitions).

To summarize,

$$\sum_{t=1}^{T} E_t \leq \widetilde{\beta}_T \sqrt{2Td\log(1 + TD^2/4d\lambda)} + D^2 \check{\beta}_T \sqrt{2Td\log(1 + T/\lambda)}.$$

We can also conclude that

$$\sum_{t=1}^{T} E_t = \widetilde{O}(D^2 d^{3/2} \sqrt{T}).$$

□

### D.7 Proof of Lemma C.2

*Proof of Lemma C.2.* We have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}} N_1 &= \sum_{t=2}^{T} \mathcal{P}_{\boldsymbol{\theta}}(s_t = x_1) \\ &= \underbrace{\sum_{t=2}^{T} \mathcal{P}_{\boldsymbol{\theta}}(s_t = x_1 | s_{t-1} = x_1)\mathcal{P}_{\boldsymbol{\theta}}(s_{t-1} = x_1)}_{I_1} + \underbrace{\sum_{t=2}^{T} \mathcal{P}_{\boldsymbol{\theta}}(s_t = x_1, s_{t-1} = x_0)}_{I_2}. \end{aligned} \tag{D.6}$$

For $I_1$, since $\mathcal{P}_{\boldsymbol{\theta}}(s_t = x_1 | s_{t-1} = x_1) = 1 - \delta$ no matter which action is taken, thus we have

$$I_1 = (1-\delta)\sum_{t=2}^{T} \mathcal{P}_{\boldsymbol{\theta}}(s_{t-1} = x_1) = (1-\delta)\mathbb{E}_{\boldsymbol{\theta}} N_1 - (1-\delta)\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_1). \tag{D.7}$$

Next we bound $I_2$. We can further decompose $I_2$ as follows.

$$\begin{aligned} I_2 &= \sum_{t=2}^{T}\sum_{\mathbf{a}} \mathcal{P}_{\boldsymbol{\theta}}(s_t = x_1 | s_{t-1} = x_0, a_{t-1} = \mathbf{a})\mathcal{P}_{\boldsymbol{\theta}}(s_{t-1} = x_0, a_{t-1} = \mathbf{a}) \\ &= \sum_{t=2}^{T}\sum_{\mathbf{a}} (\delta + \langle \mathbf{a}, \boldsymbol{\theta} \rangle)\mathcal{P}_{\boldsymbol{\theta}}(s_{t-1} = x_0, a_{t-1} = \mathbf{a}) \\ &= \sum_{\mathbf{a}} (\delta + \langle \mathbf{a}, \boldsymbol{\theta} \rangle)\Big[\mathbb{E}_{\boldsymbol{\theta}} N_0^{\mathbf{a}} - \mathcal{P}_{\boldsymbol{\theta}}(s_T = x_0, a_T = \mathbf{a})\Big]. \end{aligned} \tag{D.8}$$

Substituting (D.7) and (D.8) into (D.6) and rearranging it, we have

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\theta}} N_1 \\ &= \sum_{\mathbf{a}} (1 + \langle \mathbf{a}, \boldsymbol{\theta} \rangle / \delta)\mathbb{E}_{\boldsymbol{\theta}} N_0^{\mathbf{a}} - \underbrace{\Big[\frac{1-\delta}{\delta}\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_1) + \sum_{\mathbf{a}} (1 + \langle \mathbf{a}, \boldsymbol{\theta} \rangle / \delta)\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_0, a_T = \mathbf{a})\Big]}_{\psi_{\boldsymbol{\theta}}} \\ &= \mathbb{E}_{\boldsymbol{\theta}} N_0 + \delta^{-1}\sum_{\mathbf{a}} \langle \mathbf{a}, \boldsymbol{\theta} \rangle \mathbb{E}_{\boldsymbol{\theta}} N_0^{\mathbf{a}} - \psi_{\boldsymbol{\theta}}, \end{aligned} \tag{D.9}$$

which suggests that

$$\mathbb{E}_{\boldsymbol{\theta}} N_1 \le T/2 + \delta^{-1}\sum_{\mathbf{a}} \langle \mathbf{a}, \boldsymbol{\theta} \rangle \mathbb{E}_{\boldsymbol{\theta}} N_0^{\mathbf{a}}/2. \tag{D.10}$$

We now bound $\mathbb{E}_{\boldsymbol{\theta}} N_0$. By (D.9), we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} N_1 &\ge \mathbb{E}_{\boldsymbol{\theta}} N_0 + \delta^{-1}\sum_{\mathbf{a}} \langle \mathbf{a}, \boldsymbol{\theta} \rangle \mathbb{E}_{\boldsymbol{\theta}} N_0^{\mathbf{a}} - \psi_{\boldsymbol{\theta}} \\ &\ge \mathbb{E}_{\boldsymbol{\theta}} N_0 - \frac{\Delta}{\delta}\mathbb{E}_{\boldsymbol{\theta}} N_0 - \frac{1-\delta}{\delta}\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_1) - \mathcal{P}_{\boldsymbol{\theta}}(s_T = x_0) - \frac{\Delta}{\delta}\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_0) \\ &= (1 - \Delta/\delta)\mathbb{E}_{\boldsymbol{\theta}} N_0 - (1-\delta)/\delta + \frac{1-\Delta}{\delta}\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_0) \\ &\ge (1 - \Delta/\delta)\mathbb{E}_{\boldsymbol{\theta}} N_0 - (1-\delta)/\delta, \end{aligned} \tag{D.11}$$

where the first inequality holds due to (D.9), the second inequality holds due to the fact that $\langle \mathbf{a}, \boldsymbol{\theta} \rangle \le \Delta$, the last inequality holds since $\mathcal{P}_{\boldsymbol{\theta}}(s_T = x_0) > 0$. (D.11) suggests that

$$\mathbb{E}_{\boldsymbol{\theta}} N_0 \le \frac{T + (1-\delta)/\delta}{2 - \Delta/\delta} \le \frac{4}{5}T,$$

where the last inequality holds due to the fact that $2\Delta \le \delta$ and $(1-\delta)/\delta < T/5$. □

### D.8 Proof of Lemma C.4

We need the following lemma:

**Lemma D.1** (Lemma 20 in Jaksch et al. (2010)). Suppose $0 \le \delta' \le 1/2$ and $\epsilon' \le 1 - 2\delta'$, then

$$\delta' \log \frac{\delta'}{\delta' + \epsilon'} + (1 - \delta') \log \frac{(1 - \delta')}{1 - \delta' - \epsilon'} \le \frac{2(\epsilon')^2}{\delta'}.$$

*Proof of Lemma C.4.* Let $\mathbf{s}_t$ denote $\{s_1, \ldots, s_t\}$. By the Markovnian property of MDP, we can first decompose the KL divergence as follows:

$$\mathrm{KL}(\mathcal{P}_{\boldsymbol{\theta}'} \| \mathcal{P}_{\boldsymbol{\theta}}) = \sum_{t=1}^{T-1} \mathrm{KL}\Big[\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1}|\mathbf{s}_t) \Big\| \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)\Big],$$

where the KL divergence between $\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1}|\mathbf{s}_t), \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)$ is defined as follows:

$$\mathrm{KL}\Big[\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1}|\mathbf{s}_t) \Big\| \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)\Big] = \sum_{\mathbf{s}_{t+1} \in \mathcal{S}^{t+1}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)}.$$

Now we further bound the above terms as follows:

$$\begin{aligned}
&\sum_{\mathbf{s}_{t+1} \in \mathcal{S}^{t+1}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)} \\
&= \sum_{\mathbf{s}_t \in \mathcal{S}^t} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}_t) \sum_{x \in \mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|\mathbf{s}_t) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1} = x|\mathbf{s}_t)} \\
&= \sum_{\mathbf{s}_{t-1} \in \mathcal{S}^{t-1}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}_{t-1}) \sum_{x' \in \mathcal{S}, \mathbf{a} \in \mathcal{A}} \mathcal{P}_{\boldsymbol{\theta}'}(s_t = x', a_t = \mathbf{a}|\mathbf{s}_{t-1}) \\
&\quad \cdot \sum_{x \in \mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|\mathbf{s}^{t-1}, s_t = x', a_t = \mathbf{a}) \underbrace{\log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|\mathbf{s}^{t-1}, s_t = x', a_t = \mathbf{a})}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1} = x|\mathbf{s}^{t-1}, s_t = x', a_t = \mathbf{a})}}_{I_1},
\end{aligned}$$

When $x' = x_1$, $\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|\mathbf{s}^{t-1}, s_t = x', a_t = \mathbf{a}) = \mathcal{P}_{\boldsymbol{\theta}}(s_{t+1} = x|\mathbf{s}^{t-1}, s_t = x', a_t = \mathbf{a})$ for all $\boldsymbol{\theta}', \boldsymbol{\theta}$ since the transition probability at $x_1$ is irrelevant to $\boldsymbol{\theta}$ due to the MDP we choose. That implies when $x' = x_1$, $I_1 = 0$. Therefore,

$$\begin{aligned}
&\sum_{\mathbf{s}_{t+1} \in \mathcal{S}^{t+1}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1}|\mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1}|\mathbf{s}_t)} \\
&= \sum_{\mathbf{s}^{t-1} \in \mathcal{S}^{t-1}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}^{t-1}) \sum_{\mathbf{a}} \mathcal{P}_{\boldsymbol{\theta}'}(s_t = x_0, a_t = \mathbf{a}|\mathbf{s}^{t-1}) \\
&\quad \cdot \sum_{x \in \mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|\mathbf{s}^{t-1}, s_t = x_0, a_t = \mathbf{a}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = s|\mathbf{s}^{t-1}, s_t = x_0, a_t = \mathbf{a})}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1} = s|\mathbf{s}^{t-1}, s_t = x_0, a_t = \mathbf{a})} \\
&= \sum_{\mathbf{a}} \mathcal{P}_{\boldsymbol{\theta}'}(s_t = x_0, a_t = \mathbf{a}) \\
&\quad \underbrace{\cdot \sum_{x \in \mathcal{S}} \mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = s|s_t = x_0, a_t = \mathbf{a}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} = x|s_t = x_0, a_t = \mathbf{a})}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1} = x|s_t = x_0, a_t = \mathbf{a})}}_{I_2}.
\end{aligned} \tag{D.12}$$

To bound $I_2$, due to the structure of the MDP, we know that $s_{t+1}$ follows the Bernoulli distribution over $x_0$ and $x_1$ with probability $1 - \delta - \langle \mathbf{a}, \boldsymbol{\theta}' \rangle$ and $\delta + \langle \mathbf{a}, \boldsymbol{\theta}' \rangle$, then we have

$$I_2 = (1 - \langle \boldsymbol{\theta}', \mathbf{a} \rangle - \delta) \log \frac{1 - \langle \boldsymbol{\theta}', \mathbf{a} \rangle - \delta}{1 - \langle \boldsymbol{\theta}, \mathbf{a} \rangle - \delta} + (\langle \boldsymbol{\theta}', \mathbf{a} \rangle + \delta) \log \frac{\langle \boldsymbol{\theta}', \mathbf{a} \rangle + \delta}{\langle \boldsymbol{\theta}, \mathbf{a} \rangle + \delta} \le \frac{2\langle \boldsymbol{\theta}' - \boldsymbol{\theta}, \mathbf{a} \rangle^2}{\langle \boldsymbol{\theta}', \mathbf{a} \rangle + \delta}, \tag{D.13}$$

where the inequality holds due to Lemma D.1 with $\delta' = \langle \boldsymbol{\theta}', \mathbf{a} \rangle + \delta$ and $\epsilon' = \langle \boldsymbol{\theta} - \boldsymbol{\theta}', \mathbf{a} \rangle$. It can be verified that

$$\delta' = \langle \boldsymbol{\theta}', \mathbf{a} \rangle + \delta \leq \Delta + \delta \leq 1/2, \tag{D.14}$$

where the first inequality holds due to the definition of $\boldsymbol{\theta}'$, the second inequality holds since $\Delta < \delta/2 \leq 1/6$. It can also be verified that

$$\epsilon' = \langle \boldsymbol{\theta} - \boldsymbol{\theta}', \mathbf{a} \rangle \leq 2\Delta \leq 1 - 2(\Delta + \delta) \leq 1 - 2\delta', \tag{D.15}$$

where the first inequality holds due to the definition of $\boldsymbol{\theta}', \boldsymbol{\theta}$, the second inequality holds since $\Delta < \delta/4 \leq 1/12$, the last inequality holds since $\delta' = \langle \boldsymbol{\theta}', \mathbf{a} \rangle + \delta \leq \Delta + \delta$ due to the definition of $\boldsymbol{\theta}'$. (D.14) and (D.15) suggest that we can apply Lemma D.1 onto (D.13). $I_2$ can be further bounded as follows:

$$I_2 \leq \frac{4\langle \boldsymbol{\theta}' - \boldsymbol{\theta}, \mathbf{a} \rangle^2}{\delta} = \frac{16\Delta^2}{(d-1)^2 \delta}, \tag{D.16}$$

where the inequality holds due to (D.13) and the fact that $\delta + \langle \boldsymbol{\theta}', \mathbf{a} \rangle \geq \delta - \Delta \geq \delta/2$. Substituting (D.16) into (D.12), taking summation from $t = 1$ to $T - 1$, we have

$$\begin{aligned}
\mathrm{KL}(\mathcal{P}_{\boldsymbol{\theta}'} \| \mathcal{P}_{\boldsymbol{\theta}}) &= \sum_{t=1}^{T-1} \sum_{\mathbf{s}_{t+1} \in \mathcal{S}^{t+1}} \mathcal{P}_{\boldsymbol{\theta}'}(\mathbf{s}_{t+1}) \log \frac{\mathcal{P}_{\boldsymbol{\theta}'}(s_{t+1} | \mathbf{s}_t)}{\mathcal{P}_{\boldsymbol{\theta}}(s_{t+1} | \mathbf{s}_t)} \\
&\leq \frac{16\Delta^2}{(d-1)^2 \delta} \sum_{t=1}^{T-1} \sum_{\mathbf{a}} \mathcal{P}_{\boldsymbol{\theta}'}(s_t = x_0, a_t = \mathbf{a}) \\
&= \frac{16\Delta^2}{(d-1)^2 \delta} \sum_{t=1}^{T-1} \mathcal{P}_{\boldsymbol{\theta}'}(s_t = x_0) \\
&\leq \frac{16\Delta^2}{(d-1)^2 \delta} \mathbb{E}_{\boldsymbol{\theta}'} N_0,
\end{aligned}$$

where the last inequality holds due to the definition of $N_0$. □

# E Experiments

In this section, we conduct experiments to empirically study the performance of the proposed algorithm.

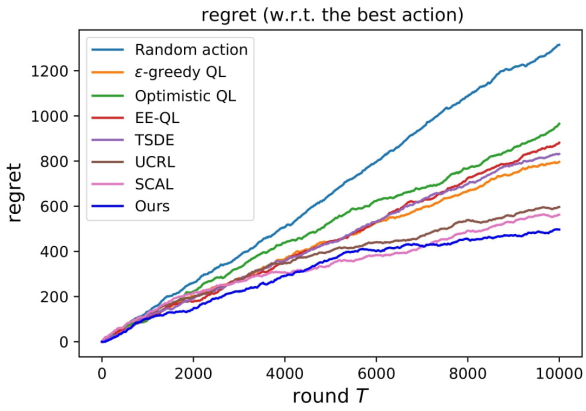


Figure 3: Regret comparison of different algorithms. UCRL2-VTR performs better than the tabular Q-learning by utilizing the given linear structure.

The MDP is constructed as described in Section C.1. We choose $d = 8$, and thus $|\mathcal{S}| = 2$ and $|\mathcal{A}| = 2^{d-1} = 128$.

We compare the following algorithms:

1. Randomly choose an action (Random action).
2. Q-learning with an $\epsilon$-greedy, uniformly random exploration ($\epsilon$-greedy QL).
3. Q-learning with a confidence bonus (Optimistic QL by Wei et al. (2020b)).
4. An Exploration Enhanced Q-learning algorithm (EE-QL by Jafarnia-Jahromi et al. (2020)).
5. A Thompson sampling-based algorithm (TSDE by Ouyang et al. (2017)).
6. A tabular model-based algorithm (UCRL by Jaksch et al. (2010)).
7. A tabular model-based algorithm that relies on the span of the MDP rather than the diameter (SCAL by Fruit et al. (2018b)).
8. Our algorithm with the Hoeffding bonus (Ours).

In our experiments, the parameters of each algorithm are tuned properly. For each algorithm, the experiment is replicated for 10 times and the averaged regret is plotted in Figure 3 for comparison. We can see that model-based algorithms (UCRL, SCAL, Ours) are generally better than the model-free ones (Q-learning algorithms and TSDE). Our proposed algorithm outperforms other model-based algorithms due to utilizing the linear structure of the underlying MDP.