Nearly Minimax Optimal Reinforcement Learning for Linear Mixture Markov Decision Processes

Dongruo Zhou DRZHOU@CS.UCLA.EDU
University of California, Los Angeles

Quanquan Gu QGU@CS.UCLA.EDU
University of California, Los Angeles

Csaba Szepesvári CSABA.SZEPESVARI@UALBERTA.CA
Deepmind and University of Alberta

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study reinforcement learning (RL) with linear function approximation where the underlying transition probability kernel of the Markov decision process (MDP) is a linear mixture model (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021) and the learning agent has access to either an integration or a sampling oracle of the individual basis kernels. For the fixed-horizon episodic setting with inhomogeneous transition kernels, we propose a new, computationally efficient algorithm that uses the basis kernels to approximate value functions. We show that the new algorithm, which we call UCRL-VTR$^+$, attains an $\widetilde{O}(dH\sqrt{T})$ regret where d is the number of basis kernels, H is the length of the episode and T is the number of interactions with the MDP. We also prove a matching lower bound $\Omega(dH\sqrt{T})$ for this setting, which shows that UCRL-VTR$^+$ is minimax optimal up to logarithmic factors. At the core of our results are (1) a weighted least squares estimator for the unknown transitional probability; and (2) a new Bernstein-type concentration inequality for self-normalized vector-valued martingales with bounded increments. Together, these new tools enable tight control of the Bellman error and lead to a nearly minimax regret. To the best of our knowledge, this is the first computationally efficient, nearly minimax optimal algorithm with an integration or a sampling oracle for RL with linear function approximation.

Keywords: Reinforcement learning, stochastic linear bandits, concentration inequality

1. Introduction

Improving the sample efficiency of reinforcement learning (RL) algorithms has been a central research question in the RL community. When there are finitely many states and actions and the value function is represented using "tables", the case known as "tabular RL", a number of breakthroughs during the past decade led to a thorough understanding of the limits of sample efficiency of RL. In particular, algorithms with nearly minimax optimal sample complexity have been discovered for the planning setting where a generative model is available (Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2020). Significant further work then led to nearly minimax optimal algorithms[1] for the more challenging online learning setting, where the results cover a wide variety of objectives, ranging from episodic Markov Decision Processes (MDPs) (Azar et al., 2017; Zanette and Brunskill, 2019; Zhang et al., 2020), through discounted MDPs (Lattimore and Hutter, 2012; Zhang et al.,

[1] In this paper, we say an algorithm is nearly minimax optimal if this algorithm attains a regret/sample complexity that matches the minimax lower bound up to logarithmic factors.

© 2021 D. Zhou, Q. Gu & C. Szepesvári.

2021b; He et al., 2020) to infinite horizon MDPs with the average reward criterion (Zhang and Ji, 2019; Tossou et al., 2019).

A classical approach to deal with such large MDPs is to assume access to a *function approximation* technique that allows for a compact, or compressed representation of various objects of interest, such as policies or value functions (Sutton and Barto, 1998). Recently, there is a growing body of work in understanding the interplay between reinforcement learning and function approximation. When a generative model is available, Yang and Wang (2019) proposed a computationally efficient, nearly minimax optimal RL algorithm that works with such linear function approximation for a special case when the learner has access to a polynomially sized set of "anchor state-action pairs". Lattimore et al. (2020) proposed an optimal-design based RL algorithm without the anchor state-action pairs assumption. However, for online RL where no generative model is accessible, as of today a gap between the upper bounds (Yang and Wang, 2020; Jin et al., 2020; Wang et al., 2020c; Modi et al., 2020; Zanette et al., 2020a,b; Jia et al., 2020; Ayoub et al., 2020) and the lower bounds (Du et al., 2019; Zhou et al., 2021) still exist, with or without the anchor state-action assumption. Therefore, a natural question arises:

Does there exist a computationally efficient, nearly minimax optimal RL algorithm with linear function approximation?

In this paper, we answer this question affirmatively for the special class of linear mixture MDPs, where the transition probability kernel is a linear mixture of a number of basis kernels (Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021). Following ideas developed for the tabular case (e.g., Azar et al. 2013) we replace the conservative Hoeffding-type confidence bounds used in UCRL-VTR of Ayoub et al. (2020) with a Bernstein-type confidence bound that is based on a new, Bernstein-type variant of the standard self-normalized concentration inequality of Abbasi-Yadkori et al. (2011). In detail, our contributions are listed as follows.

- We propose a Bernstein-type self-normalized concentration inequality for vector-valued martingales, which improves the dominating term of the analog inequality of Abbasi-Yadkori et al. (2011) from $R\sqrt{d}$ to $\sigma\sqrt{d} + R$, where R and σ^2 are the magnitude and the variance of the noise respectively, and d is the dimension of the vectors involved. Our concentration inequality is a non-trivial extension of the Bernstein inequality from the scalar case to the vector case.

- With the Bernstein-type tail inequality, we consider a linear bandit problem as a "warm-up" example, whose noise at round t is R-bounded and of σ_t^2-variance. Note that bandits can be seen as a special instance of episodic RL where the length of the episode equals one. We propose a new algorithm called Weighted OFUL, which adapts a new linear regression scheme called *weighted ridge regression*. We prove that Weighted OFUL enjoys an $\widetilde{O}(R\sqrt{dT} + d\sqrt{\sum_{t=1}^T \sigma_t^2})$ regret, which strictly improves the regret $\widetilde{O}(Rd\sqrt{T})$ obtained for the OFUL algorithm by Abbasi-Yadkori et al. (2011).

- We further apply the new tail inequality to the design and analysis of online RL algorithms for the aforementioned linear mixture MDPs (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021). In the episodic setting, we propose a new algorithm, UCRL-VTR$^+$, which can be seen as an extension of UCRL-VTR studied by Jia et al. (2020); Ayoub et al. (2020). The key idea of UCRL-VTR$^+$ is to utilize weighted ridge regression and a new estimator for the variance of the value function. We show that UCRL-VTR$^+$ attains an $\widetilde{O}(dH\sqrt{T}+\sqrt{dH^3}\sqrt{T}+d^2H^3+d^3H^2)$ regret, where T is the number of interactions with the MDP and H is the episode length. We also prove a nearly matching lower bound $\Omega(dH\sqrt{T})$ on the regret. When $d \geq H$ and $T \geq d^4H^2 + d^3H^3$, our UCRL-VTR$^+$ algorithm achieves an $\widetilde{O}(dH\sqrt{T})$ regret, which matches our proved lower bound. Thus, our results imply that our algorithm is minimax optimal up to logarithmic factors in the high-dimensional large-sample regime.

To the best of our knowledge, ignoring logarithmic factors, our proposed UCRL-VTR$^+$ is the first minimax optimal online RL algorithm with linear function approximation using the common case of a constant-dimension feature mapping. UCRL-VTR$^+$ is also computationally efficient with an access to a sampling or an integration oracle. The closest to our result is that of Zanette et al. (2020b) who proved their ELEANOR algorithm enjoys a regret of at most $\widetilde{O}(\sum_{h=1}^H d_h\sqrt{K})$, where d_h is the dimension of the feature mapping at the h-th stage and K is the number of episodes. ELEANOR can be shown to be nearly optimal for the special case when $d_1 = \sum_{h=2}^H d_h$, but is not optimal when $d_1 = \cdots = d_H = d$. Furthermore, as noted by the authors, ELEANOR is not computationally efficient.

Notation We use lower case letters to denote scalars, and use lower and upper case bold face letters to denote vectors and matrices respectively. We denote by $[n]$ the set $\{1,\ldots,n\}$. For a vector $\mathbf{x} \in \mathbb{R}^d$ and matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, a positive semi-definite matrix, we denote by $\|\mathbf{x}\|_2$ the vector's Euclidean norm and define $\|\mathbf{x}\|_{\boldsymbol{\Sigma}} = \sqrt{\mathbf{x}^\top\boldsymbol{\Sigma}\mathbf{x}}$. For $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$, let $\mathbf{x}\odot\mathbf{y}$ be the Hadamard (componentwise) product of $\mathbf{x}$ and $\mathbf{y}$. For two positive sequences $\{a_n\}$ and $\{b_n\}$ with $n = 1,2,\ldots$, we write $a_n = O(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \leq Cb_n$ holds for all $n \geq 1$ and write $a_n = \Omega(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \geq Cb_n$ holds for all $n \geq 1$. We use $\widetilde{O}(\cdot)$ to further hide the polylogarithmic factors. We use $\mathbb{1}\{\cdot\}$ to denote the indicator function. For $a,b \in \mathbb{R}$ satisfying $a \leq b$, we use $[x]_{[a,b]}$ to denote the function $x\cdot\mathbb{1}\{a \leq x \leq b\} + a\cdot\mathbb{1}\{x < a\} + b\cdot\mathbb{1}\{x > b\}$, which truncates its argument to the $[a,b]$ interval.

2. Related Work

The purpose of this section is to review prior works that are most relevant to our contributions.

Linear Bandits Linear bandits can be seen as the simplest version of RL with linear function approximation, where the episode length (i.e., planning horizon) $H = 1$. There is a huge body of literature on linear bandit problems (Auer, 2002; Chu et al., 2011; Li et al., 2010, 2019; Dani et al., 2008; Abbasi-Yadkori et al., 2011). Most of the linear bandit algorithms can be divided into two categories: algorithms for k-armed linear bandits, and algorithms for infinite-armed linear bandits. For the k-armed case, Auer (2002) proposed a SupLinRel algorithm, which makes use of the eigenvalue decomposition and enjoys an $O(\log^{3/2}(kT)\sqrt{dT})$ regret[2] . Li et al. (2010); Chu et al. (2011) proposed a SupLinUCB algorithm using the regularized least-squares estimator, which enjoys the same regret guarantees. Li et al. (2019) proposed a VCL-SupLinUCB algorithm with a refined confidence set design which enjoys an improved $O(\sqrt{\log(T)\log(k)dT})$ regret, which matches the lower bound up to a logarithmic factor. For the infinite-armed case, Dani et al. (2008) proposed an algorithm with a confidence ball, which enjoys $O(d\sqrt{T\log^3 T})$ regret. Abbasi-Yadkori et al. (2011) improved the regret to $O(d\sqrt{T\log^2 T})$ with a new self-normalized concentration inequality

[2] We omit the poly$(\log\log(kT))$ factors for the simplicity of comparison.

for vector-valued martingales. Li et al. (2021) further improved the regret to $O(d\sqrt{T}\log T)$, which matches the lower bound up to a logarithmic factor. However, previous works only focus on the case where the reward noise is sub-Gaussian. In this paper, we show that if the reward noise is restricted to a smaller class of distributions with bounded magnitude and variance, a better regret bound can be obtained. The main motivation to consider this problem is that linear bandits with bounded reward and variance can be seen as a special RL with linear function approximation when the episode length $H = 1$. Thus, this result immediately sheds light on the challenges involved in achieving minimax optimal regret for general RL with linear function approximation.

Reinforcement Learning with Linear Function Approximation Recent years have witnessed a flurry of activity on RL with linear function approximation (e.g., Jiang et al., 2017; Yang and Wang, 2019, 2020; Jin et al., 2020; Wang et al., 2020c; Modi et al., 2020; Dann et al., 2018; Du et al., 2019; Sun et al., 2019; Zanette et al., 2020a,b; Cai et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Weisz et al., 2021; Zhou et al., 2021; He et al., 2021). These results can be generally grouped into four categories based on their assumptions on the underlying MDP. The first category of work uses the low Bellman-rank assumption (Jiang et al., 2017) which assumes that the Bellman error "matrix" where "rows" are index by a test function and columns are indexed by a distribution generating function from the set of test functions assumes a low-rank factorization. Representative work includes Jiang et al. (2017); Dann et al. (2018); Sun et al. (2019). The second category of work considers the *linear MDP* assumption (Yang and Wang, 2019; Jin et al., 2020) which assumes taht both the transition probability function and reward function are parameterized as a linear function of a given feature mapping over state-action pairs. Representative work includes Yang and Wang (2019); Jin et al. (2020); Wang et al. (2020c); Du et al. (2019); Zanette et al. (2020a); Wang et al. (2020b); He et al. (2021). The third category of work focuses on the low inherent Bellman error assumption (Zanette et al., 2020b), which assumes the Bellman backup can be parameterized as a linear function up to some misspecification error. Zanette et al. (2020b) proposed an ELEANOR algorithm with a regret $\widetilde{O}(\sum_{h=1}^{H} d_h\sqrt{K})$, where d_h is the dimension of the feature mapping at the h-th stage within the episodes and K is the number of episodes. They also proved a lower bound $\Omega(\sum_{h=1}^{H} d_h\sqrt{K})$ under the sub-Gaussian norm assumption of the rewards and transitions but only for the special case when $d_1 = \sum_{h=2}^{H} d_h$. It can be seen that in this special case, their upper bound matches their lower bound up to logarithmic factors, and thus their algorithm is statistically near optimal. However, in the general case when $d_1 = \cdots = d_H = d$, there still exists a gap of H between their upper and lower bounds. Furthermore, as noted by the authors, the ELEANOR algorithm is not computationally efficient. The last category considers linear mixture MDPs (a.k.a., linear kernel MDPs) (Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021), which assumes that the transition probability function is parameterized as a linear function of a given feature mapping over state-action-next-state triples. Representative work includes Yang and Wang (2020); Modi et al. (2020); Jia et al. (2020); Ayoub et al. (2020); Cai et al. (2020); Zhou et al. (2021); He et al. (2021) (of these, Yang and Wang (2020) considers a special case, but their results extend to the linear mixture case seamlessly). Our work also considers linear mixture MDPs.

Bernstein Bonuses for Tabular MDPs There is a series of work proposing algorithms with nearly minimax optimal sample complexity or regret for the tabular MDP under different settings, including average-reward, discounted, and episodic MDPs (Azar et al., 2013, 2017; Zanette and Brunskill, 2019; Zhang and Ji, 2019; Simchowitz and Jamieson, 2019; Zhang et al., 2020; He et al., 2020; Zhang et al., 2021a). The key idea at the heart of these works is the usage of the law of total variance to obtain tighter bounds on the expected sum of the variances for the estimated value function.

These works have designed tighter confidence sets or upper confidence bounds by replacing the Hoeffding-type exploration bonuses with Bernstein-type exploration bonuses, and obtained more accurate estimates of the optimal value function, a technique pioneered by Lattimore and Hutter (2012). Our work shows how this idea extends to algorithms with linear function approximation. To the best of our knowledge, our work is the first work using Bernstein bonus and law of total variance to achieve nearly minimax optimal regret for RL with linear function approximation.

3. Preliminaries

We consider RL with linear function approximation for episodic MDPs. In the following, we will introduce the necessary background and definitions. For further background, the reader is advised to consult, e.g., Puterman (2014).

Inhomogeneous, episodic MDP We denote an inhomogeneous, episodic MDP by a tuple $M = (\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the length of the episode, $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the deterministic reward function, and $\mathbb{P}_h$ is the transition probability function at stage h so that for $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$, $\mathbb{P}_h(s'|s, a)$ is the probability of arriving at stage $h + 1$ at state s' provided that the state at stage h is s and action a is chosen at this stage. For the sake of simplicity, we restrict ourselves to countable state and finite action spaces. A policy $\pi = \{\pi_h\}_{h=1}^H$ is a collection of H functions, where each of them maps a state s to an action a. For $(s, a) \in \mathcal{S} \times \mathcal{A}$, we define the action-values $Q_h^\pi(s, a)$ and (state) values $V_h^\pi(s)$ as follows:

$$Q_h^\pi(s, a) = \mathbb{E}_{\pi,h,s,a}\left[\sum_{h'=h}^{H} r_h(s_{h'}, a_{h'})\right],\ V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)),\ V_{H+1}^\pi(s) = 0.$$

In the definition of Q_h^π, $\mathbb{E}_{\pi,h,s,a}$ means an expectation over the probability measure over state-action pairs of length $H - h + 1$ that is induced by the interconnection of policy π and the MDP M when initializing the process to start at stage h with the pair (s, a). In particular, the probability of sequence $(s_h, a_h, s_{h+1}, a_{h+1}, \ldots, s_H, a_H)$ under this sequence is $\mathbf{1}(s_h = s)\mathbf{1}(a_h = a)\mathbb{P}_h(s_{h+1}|s_h, a_h)\mathbf{1}_{\pi_{h+1}(s_{h+1})=a_{h+1}} \ldots \mathbb{P}_{H-1}(s_H|s_{H-1}, a_{H-1})\mathbf{1}_{\pi_H(s_H)=a_H}$. The optimal value function $V_h^*(\cdot)$ and the optimal action-value function $Q_h^*(\cdot, \cdot)$ are defined by $V_h^*(s) = \sup_\pi V_h^\pi(s)$ and $Q_h^*(s, a) = \sup_\pi Q_h^\pi(s, a)$, respectively. For any function $V : \mathcal{S} \to \mathbb{R}$, we introduce the shorthands

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

where V^2 stands for the function whose value at s is $V^2(s)$. Using this notation, the Bellman equations for policy π and the Bellman optimality equation can be written as

$$Q_h^\pi(s, a) = r_h(s, a) + [\mathbb{P}_h V_{h+1}^\pi](s, a),\ Q_h^*(s, a) = r_h(s, a) + [\mathbb{P}_h V_{h+1}^*](s, a).$$

Note that both hold *simultaneously* for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$.

In the **online learning setting**, a learning agent who does not know the kernels $\{\mathbb{P}_h\}_h$ but, for the sake of simplicity, knows the rewards $\{r_h\}_h$, aims to learn to take good actions by interacting with the environment. For each $k \geq 1$, at the beginning of the k-th episode, the environment picks the initial state s_1^k and the agent chooses a policy π^k to be followed in this episode. As the agent follows the policy through the episode, it observes the sequence of states $\{s_h^k\}_h$ with

$s_{h+1}^k \sim \mathbb{P}_h(\cdot|s_h^k, \pi^k(s_h^k))$. The goal is to design a learning algorithm that constructs the sequence $\{\pi^k\}_k$ based on past information so that the K-episode regret,

$$\text{Regret}(M, K) = \sum_{k=1}^{K} \big[V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\big]$$

is kept small. In this paper, we focus on proving high probability bounds on the regret Regret(M, K), as well as lower bounds in expectation.

Linear Mixture MDPs We consider a special class of MDPs called *linear mixture MDPs* (a.k.a., linear kernel MDPs), where the transition probability kernel is a linear mixture of a number of basis kernels. This class has been considered by a number of previous authors (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021) and is defined as follows: Firstly, let $\boldsymbol{\phi}(s'|s,a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ be a feature mapping satisfying that for any bounded function $V : \mathcal{S} \to [0, 1]$ and any tuple $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\|\boldsymbol{\phi}_V(s,a)\|_2 \leq 1, \text{where } \boldsymbol{\phi}_V(s,a) = \sum_{s' \in \mathcal{S}} \boldsymbol{\phi}(s'|s,a)V(s')\,. \tag{1}$$

We define **episodic linear mixture MDPs** as follows:

Definition 1 (Jia et al. 2020; Ayoub et al. 2020) *$M = (\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$ is called an inhomogeneous, episodic B-bounded linear mixture MDP if there exist vectors $\boldsymbol{\theta}_h \in \mathbb{R}^d$ with $\|\boldsymbol{\theta}_h\|_2 \leq B$ and $\boldsymbol{\phi}(\cdot|\cdot,\cdot)$ satisfying (1), such that $\mathbb{P}_h(s'|s,a) = \langle \boldsymbol{\phi}(s'|s,a), \boldsymbol{\theta}_h \rangle$ for any state-action-next-state triplet $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and stage h.*

Note that in the learning problem, the vectors introduced in the above definition are initially unknown to the learning agent. In the rest of this paper, we assume that the learning agent is given access to $\boldsymbol{\phi}$ and the unknown episodic linear mixture MDP is parameterized by $\boldsymbol{\Theta}^* = \{\boldsymbol{\theta}_h^*\}_{h=1}^H$. We denote this MDP by $M_{\boldsymbol{\Theta}^*}$.

4. Challenges and New Technical Tools

To motivate our approach, we start this section with a recap of previous work addressing online learning in episodic linear mixture MDPs. This allows us to argue for how this work falls short of achieving minimax optimal regret and motivates us to develop new theoretical tools to achieve that.

4.1. Barriers to Minimax Optimality in RL with Linear Function Approximation

To understand the key technical challenges that underly achieving minimax optimality in RL with linear function approximation, we first look into the UCRL with "value-targeted regression" (UCRL-VTR) method of Jia et al. (2020) (for a longer exposition, with refined results see Ayoub et al. (2020)) for episodic linear mixture MDPs. The key idea of UCRL-VTR is using a model-based supervised learning framework to learn the underlying unknown parameter vector $\boldsymbol{\theta}_h^*$ of linear mixture MDP, and use the learned parameter vector $\boldsymbol{\theta}_{k,h}$ to build an optimistic estimator $Q_{k,h}(\cdot,\cdot)$ for the optimal action-value function $Q^*(\cdot,\cdot)$. In detail, for any stage h of the k-th episode, the following equation holds: For value functions $V_k = \{V_{k,h}\}_h$ constructed based on data received before episode k and the state action pair (s_h^k, a_h^k) visited in stage h of episode k,

$$[\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) = \bigg\langle \sum_{s'} \boldsymbol{\phi}(s'|s_h^k, a_h^k) V_{k,h+1}(s'), \boldsymbol{\theta}_h^* \bigg\rangle = \langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^* \rangle,$$

where the first equation holds due to the definition of linear mixture MDPs (cf. Definition 1), the second equation holds due to the definition of $\boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)$ in (1). As it turns out, taking actions that maximize the value shown above with appropriately constructed value functions V_k is sufficient for minimizing regret. Therefore, learning the underlying $\boldsymbol{\theta}_h^*$ can be regarded as solving a "linear bandit" problem (Part V, Lattimore and Szepesvári, 2020), where the context is $\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k) \in \mathbb{R}^d$, and the noise is $V_{k,h+1}(s_{h+1}^k) - [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k)$. Previous work (Jia et al., 2020; Ayoub et al., 2020) proposed an estimator $\boldsymbol{\theta}_{k,h}$ as the minimizer to the following regularized linear regression problem:

$$\boldsymbol{\theta}_{k,h} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \lambda \|\boldsymbol{\theta}\|_2^2 + \sum_{j=1}^{k-1} \big[\langle \boldsymbol{\phi}_{V_{j,h+1}}(s_h^j, a_h^j), \boldsymbol{\theta} \rangle - V_{j,h+1}(s_{h+1}^j)\big]^2. \tag{2}$$

By using the standard self-normalized concentration inequality for vector-valued martingales of Abbasi-Yadkori et al. (2011), one can show then that, with high probability, $\boldsymbol{\theta}_h^*$ lies in the ellipsoid

$$\mathcal{C}_{k,h} = \Big\{ \boldsymbol{\theta} : \Big\| \boldsymbol{\Sigma}_{k,h}^{1/2} (\boldsymbol{\theta} - \boldsymbol{\theta}_{k,h}) \Big\|_2 \leq \beta_k \Big\}$$

which is centered at $\boldsymbol{\theta}_{k,h}$, with shape parameter $\boldsymbol{\Sigma}_{k,h} = \lambda \mathbf{I} + \sum_{j=1}^{k-1} \boldsymbol{\phi}_{V_{j,h+1}}(s_h^j, a_h^j) \boldsymbol{\phi}_{V_{j,h+1}}(s_h^j, a_h^j)^\top$ and where β_k is the radius chosen to be proportional to the magnitude of the value function $V_{k,h+1}(\cdot)$, which eventually gives $\beta_k = \widetilde{O}(\sqrt{d}H)$. It follows that if we define

$$Q_{k,h}(\cdot,\cdot) = \Big[r_h(\cdot,\cdot) + \max_{\boldsymbol{\theta} \in \mathcal{C}_{k,h}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot) \rangle \Big]_{[0,H]},$$

then, with high probability, $Q_{k,1}(\cdot,\cdot)$ is an overestimate of $Q_1^*(\cdot,\cdot)$, and the summation of "suboptimality gaps" can be bounded by $\sum_{k=1}^K \sum_{h=1}^H \beta_k \|\boldsymbol{\Sigma}_{k,h}^{-1/2} \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\|_2$. This leads to the $\widetilde{O}(dH^{3/2}\sqrt{T})$ regret by further applying the elliptical potential lemma from linear bandits (Abbasi-Yadkori et al., 2011).

However, we note that the above reasoning has a number of shortcomings. First, it chooses the confidence radius β_k proportional to the *magnitude* of the value function $V_{k,h+1}(\cdot)$ rather than its *variance* $[\mathbb{V}_h V_{k,h+1}](\cdot,\cdot)$. This is known to be too conservative: Tabular RL is a special case of linear mixture MDPs and here it is known by the *law of total variance* (Lattimore and Hutter, 2012; Azar et al., 2013) that the variance of the value function is smaller than its magnitude by a factor $\sqrt{H}$. This inspires us to derive a Bernstein-type self-normalized concentration bound for vector-valued martingales which is sensitive to the variance of the martingale terms. Second, even if we were able to build such a tighter concentration bound, we still need to carefully design an algorithm because the variances of the value functions $\{\mathbb{V}_h V_{k,h+1}(s_h^k, a_h^k)\}_h$ at different stages of the episodes are non-uniform: We face a so-called *heteroscedastic* linear bandit problem. Naively choosing a uniform upper bound for all the variances $\{[\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)\}_h$ yields no improvement compared with previous results. To address this challenge, we will need to build variance estimates and use these in a weighted least-squares estimator to achieve a better aggregation of the heteroscedastic data.

4.2. A Bernstein Self-normalized Concentration Inequality for Vector-valued Martingales

One of the key results of this paper is the following Bernstein self-normalized concentration inequality:

Theorem 2 (Bernstein inequality for vector-valued martingales) *Let $\{\mathcal{G}_t\}_{t=1}^\infty$ be a filtration, $\{\mathbf{x}_t, \eta_t\}_{t\ge1}$ be 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 $t \ge 1$ let $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t\rangle + \eta_t$ and suppose that $\eta_t, \mathbf{x}_t$ also satisfy*

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

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

$$\forall t > 0,\ \Big\|\sum_{i=1}^t \mathbf{x}_i\eta_i\Big\|_{\mathbf{Z}_t^{-1}} \le \beta_t,\ \|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} \le \beta_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2, \tag{3}$$

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

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

Proof The proof adapts the proof technique of Dani et al. (2008); for details see Appendix B.1. ■

Theorem 2 can be viewed as a non-trivial extension of the Bernstein concentration inequality from scalar-valued martingales to self-normalized vector-valued martingales. It is a strengthened version of self-normalized tail inequality for vector-valued martingales when the magnitude and the variance of the noise are bounded. Abbasi-Yadkori et al. (2011) considered the setting where η_t is R-sub-Gaussian and showed that (3) holds when $\beta_t = R\sqrt{d\log((1+tL^2/\lambda)/\delta)} = \widetilde{O}(R\sqrt{d})$, while our result improves this to $\beta_t = \widetilde{O}(\sigma\sqrt{d}+R)$. A more detailed comparison between Theorem 2 and previous results is given in Appendix A.1.

4.3. Weighted Ridge Regression and Heteroscedastic Linear Bandits

In this subsection we consider the problem of linear bandits where the learner is given at the end of each round an upper bound on the (conditional) variance of the noise in the responses as input. This abstract problem is studied to work out the tools needed to handle the heteroscedasticity of the noise that arises in the linear mixture MDPs in a cleaner setting. In more details, let $\{\mathcal{D}_t\}_{t=1}^\infty$ be a fixed sequence of decision sets. The agent selects an action $\mathbf{a}_t \in \mathcal{D}_t$ and then observes the reward $r_t = \langle \boldsymbol{\mu}^*, \mathbf{a}_t\rangle + \epsilon_t$, where $\boldsymbol{\mu}^* \in \mathbb{R}^d$ is a vector unknown to the agent and ϵ_t is a random noise satisfying the following properties almost surely:

$$\forall t,\ |\epsilon_t| \le R,\ \mathbb{E}[\epsilon_t|\mathbf{a}_{1:t}, \epsilon_{1:t-1}] = 0,\ \mathbb{E}[\epsilon_t^2|\mathbf{a}_{1:t}, \epsilon_{1:t-1}] \le \sigma_t^2,\ \|\mathbf{a}_t\|_2 \le A. \tag{4}$$

As noted above, the learner gets to observe σ_t together with r_t after each choice it makes. We assume that σ_t is $(\mathbf{a}_{1:t}, \epsilon_{1:t-1})$-measurable. The goal of the agent is to minimize its *pseudo-regret*, defined as follows:

$$\text{Regret}(T) = \sum_{t=1}^T\langle \mathbf{a}_t^*, \boldsymbol{\mu}^*\rangle - \sum_{t=1}^T\langle \mathbf{a}_t, \boldsymbol{\mu}^*\rangle,\ \text{where } \mathbf{a}_t^* = \mathop{\text{argmax}}_{\mathbf{a}\in\mathcal{D}_t}\langle \mathbf{a}, \boldsymbol{\mu}^*\rangle.$$

Our problem setup is similar to the setting studied by Kirschner and Krause (2018), where it is not the variance, but the sub-Gaussianity parameter that the learner observes at the end of the rounds. The learner's goal is then to make use of this information to achieve a smaller regret as a function of the sum of squared variances (a "second-order bound"). This is also related to the Gaussian side-observation setting and partial monitoring with feedback graphs considered in Wu et al. (2015).

To make use of the variance information, we propose *Weighted OFUL*, which is an extension of the "Optimism in the Face of Uncertainty for Linear bandits" algorithm (OFUL) of Abbasi-Yadkori et al. (2011). The algorithm's pseudocode is shown in Algorithm 1.

Algorithm 1 Weighted OFUL

```
Require: Regularization parameter $\lambda > 0$, and $B$, an upper bound on the $\ell_2$-norm of $\boldsymbol{\mu}^*$
1: $\mathbf{A}_0 \leftarrow \lambda\mathbf{I}$, $\mathbf{c}_0 \leftarrow \mathbf{0}$, $\widehat{\boldsymbol{\mu}}_0 \leftarrow \mathbf{A}_0^{-1}\mathbf{c}_0$, $\widehat{\beta}_0 = 0$, $\mathcal{C}_0 \leftarrow \{\boldsymbol{\mu} : \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_0\|_{\mathbf{A}_0} \le \widehat{\beta}_0 + \sqrt{\lambda}B\}$
2: for $t = 1, \ldots, T$ do
3:     Observe $\mathcal{D}_t$
4:     Let $(\mathbf{a}_t, \widetilde{\boldsymbol{\mu}}_t) \leftarrow \mathrm{argmax}_{\mathbf{a} \in \mathcal{D}_t, \boldsymbol{\mu} \in \mathcal{C}_{t-1}} \langle \mathbf{a}, \boldsymbol{\mu} \rangle$
5:     Select $\mathbf{a}_t$ and observe $(r_t, \sigma_t)$, set $\bar{\sigma}_t$ based on $\sigma_t$, set radius $\widehat{\beta}_t$ as defined in (6)
6:     $\mathbf{A}_t \leftarrow \mathbf{A}_{t-1} + \mathbf{a}_t\mathbf{a}_t^\top/\bar{\sigma}_t^2$, $\mathbf{c}_t \leftarrow \mathbf{c}_{t-1} + r_t\mathbf{a}_t/\bar{\sigma}_t^2$, $\widehat{\boldsymbol{\mu}}_t \leftarrow \mathbf{A}_t^{-1}\mathbf{c}_t$, $\mathcal{C}_t \leftarrow \{\boldsymbol{\mu} : \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_t\|_{\mathbf{A}_t} \le \widehat{\beta}_t + \sqrt{\lambda}B\}$
7: end for
```

In round t, Weighted OFUL selects the estimate $\widehat{\boldsymbol{\mu}}_t$ of the unknown $\boldsymbol{\mu}^*$ as the minimizer to the following *weighted ridge regression* problem:

$$\widehat{\boldsymbol{\mu}}_t \leftarrow \underset{\boldsymbol{\mu} \in \mathbb{R}^d}{\mathrm{argmin}}\, \lambda\|\boldsymbol{\mu}\|_2^2 + \sum_{i=1}^{t} [\langle \boldsymbol{\mu}, \mathbf{a}_i \rangle - r_i]^2/\bar{\sigma}_i^2, \tag{5}$$

where $\bar{\sigma}_i$ is a selected upper bound of σ_i. The closed-form solution to (5) is in Line 6 of Algorithm 1. The term "weighted" refers to the normalization constant $\bar{\sigma}_i$ used in (5). The estimator in (5) is closely related to the best linear unbiased estimator (BLUE) (Henderson, 1975). In particular, in the language of linear regression, with $\lambda = 0$ and when $\bar{\sigma}_t^2$ is the variance of r_t, with a fixed design, $\widehat{\boldsymbol{\mu}}_t$ is known to be the lowest variance estimator of $\boldsymbol{\mu}^*$ in the class of linear unbiased estimators. Note that both Lattimore et al. (2015) and Kirschner and Krause (2018) used a similar weighted ridge-regression estimator for their respective problem settings.

By adapting the new Bernstein-type self-normalized concentration inequality in Theorem 2, we obtain the following bound on the regret of Weighted OFUL:

Theorem 3 *Suppose that for all $t \ge 1$ and all $\mathbf{a} \in \mathcal{D}_t$, $\langle \mathbf{a}, \boldsymbol{\mu}^* \rangle \in [-1, 1]$, $\|\boldsymbol{\mu}^*\|_2 \le B$. Set $\bar{\sigma}_t = \max\{R/\sqrt{d}, \sigma_t\}$, $\lambda = 1/B^2$ and*

$$\widehat{\beta}_0 = 0,\ \widehat{\beta}_t = 8\sqrt{d\log(1 + tA^2/([\bar{\sigma}_{min}^t]^2 d\lambda))\log(4t^2/\delta)} + 4R/\bar{\sigma}_{min}^t \cdot \log(4t^2/\delta),\ t \ge 1\,. \tag{6}$$

where $\bar{\sigma}_{min}^t = \min_{1 \le i \le t} \bar{\sigma}_i$. Then, with probability at least $1 - \delta$, the regret of Weighted OFUL for the first T rounds is bounded as follows:

$$\mathit{Regret}(T) = \widetilde{O}\bigg(R\sqrt{dT} + d\sqrt{\sum_{t=1}^{T} \sigma_t^2}\bigg). \tag{7}$$

Proof See Appendix B.2. ■

Remark 4 *Comparing* (7) *of Theorem* 3 *with the regret bound Regret*$(T) = \widetilde{O}(Rd\sqrt{T})$ *achieved by OFUL in* Abbasi-Yadkori et al. (2011)*, it can be seen that the regret of Weighted OFUL is strictly better than that of OFUL since* $\sigma_t \le R$.

5. Optimal Exploration for Episodic Linear Mixture MDPs

In this section, equipped with the new technical tools discussed in Section 4, we propose a new algorithm UCRL-VTR$^+$ for episodic linear mixture MDPs (see Definition 1). We also prove its near minimax optimality by providing matching upper and lower bounds.

5.1. The Proposed Algorithm

Algorithm 2 UCRL-VTR$^+$ for Episodic Linear Mixture MDPs

```
Require: Regularization parameter $\lambda$, an upper bound $B$ of the $\ell_2$-norm of $\boldsymbol{\theta}_h^*$
For $h \in [H]$, set $\widehat{\boldsymbol{\Sigma}}_{1,h}, \widetilde{\boldsymbol{\Sigma}}_{1,h} \leftarrow \lambda\mathbf{I}$, $\widehat{\mathbf{b}}_{1,h}, \widetilde{\mathbf{b}}_{1,h} \leftarrow \mathbf{0}$, $\widehat{\boldsymbol{\theta}}_{1,h}, \widetilde{\boldsymbol{\theta}}_{1,h} \leftarrow \mathbf{0}$, $V_{1,H+1}(\cdot) \leftarrow 0$
for $k = 1, \ldots, K$ do
    for $h = H, \ldots, 1$ do
        $Q_{k,h}(\cdot,\cdot) \leftarrow \Big[r_h(\cdot,\cdot) + \langle \widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\rangle + \widehat{\beta}_k \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\Big\|_2\Big]_{[0,H]}$, where $\widehat{\beta}_k$
        is defined in (14)
        $\pi_h^k(\cdot) \leftarrow \mathrm{argmax}_{a\in\mathcal{A}} Q_{k,h}(\cdot,a)$
        $V_{k,h}(\cdot) \leftarrow \max_{a\in\mathcal{A}} Q_{k,h}(\cdot,a)$
    end for
    Receive $s_1^k$
    for $h = 1, \ldots, H$ do
        Take action $a_h^k \leftarrow \pi_h^k(s_h^k)$, receive $s_{h+1}^k \sim \mathbb{P}_h(\cdot|s_h^k, a_h^k)$
        Set $[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)$ as in (13) and $E_{k,h}$ as in (15)
        $\bar{\sigma}_{k,h} \leftarrow \sqrt{\max\big\{H^2/d, [\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k) + E_{k,h}\big\}}$ {Variance upper bound}
        $\widehat{\boldsymbol{\Sigma}}_{k+1,h} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k,h} + \bar{\sigma}_{k,h}^{-2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)^\top$ {"Covariance", 1st moment}
        $\widehat{\mathbf{b}}_{k+1,h} \leftarrow \widehat{\mathbf{b}}_{k,h} + \bar{\sigma}_{k,h}^{-2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)V_{k,h+1}(s_{h+1}^k)$ {Response, 1st moment}
        $\widetilde{\boldsymbol{\Sigma}}_{k+1,h} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{k,h} + \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k)\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k)$ {"Covariance", 2nd moment}
        $\widetilde{\mathbf{b}}_{k+1,h} \leftarrow \widetilde{\mathbf{b}}_{k,h} + \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k)V_{k,h+1}^2(s_{h+1}^k)$ {Response, 2nd moment}
        $\widehat{\boldsymbol{\theta}}_{k+1,h} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k+1,h}^{-1}\widehat{\mathbf{b}}_{k+1,h}$, $\widetilde{\boldsymbol{\theta}}_{k+1,h} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{k+1,h}^{-1}\widetilde{\mathbf{b}}_{k+1,h}$ {1st and 2nd moment parameters}
    end for
end for
```

At a high level, UCRL-VTR$^+$ is an improved version of the UCRL-VTR algorithm by Jia et al. (2020) and refined and generalized by Ayoub et al. (2020). UCRL-VTR$^+$, shares the basic structure of UCRL-VTR, which constructs the optimistic estimate of the optimal action-value function at k-th episode and h-th stage as follows, following the optimism in the face of uncertainty principle:

$$Q_{k,h}(\cdot,\cdot) = \Big[r_h(\cdot,\cdot) + \max_{\boldsymbol{\theta}\in\widehat{\mathcal{C}}_{k,h}} \langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\rangle\Big]_{[0,H]}. \tag{8}$$

where the confidence set $\widehat{\mathcal{C}}_{k,h}$ constructed is an ellipsoid in the parameter space, centered at the parameter vector $\widehat{\boldsymbol{\theta}}_{k,h}$ and shape given by the "covariance" matrix $\widehat{\boldsymbol{\Sigma}}_{k,h}$ and having a radius of $\widehat{\beta}_k$:

$$\widehat{\mathcal{C}}_{k,h} = \Big\{\boldsymbol{\theta} : \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{k,h})\Big\|_2 \leq \widehat{\beta}_k\Big\}, \tag{9}$$

Given the choice of $\widehat{\mathcal{C}}_{k,h}$, it is not hard to see that the update in Line 4 is equivalent to (8). Given $\{Q_{k,h}\}_h$, in each episode k, at h-th stage, UCRL-VTR$^+$ executes actions that are greedy with respect to $Q_{k,h}$ (Line 5).

Weighted Ridge Regression and Optimistic Estimates of Value Functions The key novelty of UCRL-VTR$^+$ is the use of the covariance matrix $\widehat{\boldsymbol{\Sigma}}_{k,h}$ (Line 13) and the parameter vector $\widehat{\boldsymbol{\theta}}_{k,h}$ (Line 17) based on weighted ridge regression (cf. Section 4) to learn the underlying $\boldsymbol{\theta}_h^*$. To understand the mechanism behind UCRL-VTR$^+$, recall the discussion in Section 4.1: $V_{k,h+1}(s_{h+1}^k)$ and $\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)$ can be seen as the stochastic reward and context of a linear bandits problem. Then, letting $\sigma_{k,h}^2 = [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)$ be the variance of the value function, the analysis in Section 4 suggests that one should use a weighted ridge regression estimator, such as

$$\widehat{\boldsymbol{\theta}}_{k,h} = \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}\in\mathbb{R}^d} \lambda\|\boldsymbol{\theta}\|_2^2 + \sum_{j=1}^{k-1}\big[\langle\boldsymbol{\phi}_{V_{j,h+1}}(s_h^j, a_h^j), \boldsymbol{\theta}\rangle - V_{j,h+1}(s_{h+1}^j)\big]^2/\bar{\sigma}_{j,h}^2, \tag{10}$$

where $\bar{\sigma}_{j,h}$ is an appropriate upper bound on $\sigma_{j,h}$. We propose to set

$$\bar{\sigma}_{k,h} = \sqrt{\max\big\{H^2/d, [\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k) + E_{k,h}\big\}},$$

where $[\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k)$ is a scalar-valued empirical estimate for the variance of the value function $V_{k,h+1}$ under the transition probability $\mathbb{P}_h(\cdot|s_k, a_k)$, and $E_{k,h}$ is an offset term that is used to guarantee that $[\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k) + E_{k,h}$ upper bounds $\sigma_{k,h}^2$ with high probability. The detailed specifications of these are deferred later. Moreover, by construction, we have $\bar{\sigma}_{k,h} \geq H/\sqrt{d}$. Our construction of $\bar{\sigma}_{k,h}$ shares a similar spirit as the variance estimator used in *empirical Bernstein inequalities* (Audibert et al., 2009; Maurer and Pontil, 2009), which proved to be pivotal to achieve nearly minimax optimal sample complexity/regret in tabular MDPs (Azar et al., 2013, 2017; Zanette and Brunskill, 2019; He et al., 2020).

Several nontrivial questions remain to be resolved. First, we need to specify how to calculate the empirical variance $[\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k)$. Second, in order to ensure $Q_{k,h}(\cdot,\cdot)$ is an overestimate of $Q_h^*(\cdot,\cdot)$, we need to choose an appropriate $\widehat{\beta}_k$ such that $\widehat{\mathcal{C}}_{k,h}$ contains $\boldsymbol{\theta}_h^*$ with high probability. Third, we need to select $E_{k,h}$ to guarantee that $[\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k) + E_{k,h}$ upper bounds $\sigma_{k,h}^2$ with high probability.

Variance Estimator To address the first question, we recall that by definition, we have

$$\begin{aligned}[\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k) &= [\mathbb{P}_h V_{k,h+1}^2](s_h^k, a_h^k) - \big([\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k)\big)^2 \\ &= \big\langle \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\big\rangle - \big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\big\rangle\big]^2, \qquad (11)\end{aligned}$$

where the second equality holds due to the definition of linear mixture MDPs. By (11) we conclude that the expectation of $V_{k,h+1}^2(s_{h+1}^k)$ over the next state, s_{h+1}^k, is a linear function of $\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)$. Therefore, we use $\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h}\rangle$ to estimate this term, where $\widetilde{\boldsymbol{\theta}}_{k,h}$ is the solution to the following ridge regression problem:

$$\widetilde{\boldsymbol{\theta}}_{k,h} = \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}\in\mathbb{R}^d} \lambda\|\boldsymbol{\theta}\|_2^2 + \sum_{j=1}^{k-1} \big[\big\langle \boldsymbol{\phi}_{V_{j,h+1}^2}(s_h^j, a_h^j), \boldsymbol{\theta}\big\rangle - V_{j,h+1}^2(s_{h+1}^j)\big]^2. \qquad (12)$$

The closed-form solution to (12) is in Line 17. In addition, we use $\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\rangle$ to estimate the second term in (11). Meanwhile, since $[\mathbb{P}_h V_{k,h+1}^2](s_h^k, a_h^k) \in [0, H^2]$ and $[\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) \in [0, H]$ hold, we add clipping to control the range of our variance estimator, which gives the final expression of $[\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k)$:

$$[\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k) = \Big[\Big\langle \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h}\Big\rangle\Big]_{[0,H^2]} - \Big[\Big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\Big\rangle\Big]_{[0,H]}^2. \qquad (13)$$

More discussions about the algorithm design are in Appendix C.1. Meanwhile, UCRL-VTR$^+$ is computationally efficient for some specific family of $\boldsymbol{\phi}(\cdot|\cdot,\cdot)$ given access to an integration oracle. Detailed discussions are in Appendix C.2.

Confidence Set To address the choice of $\widehat{\beta}_k$ and $E_{k,h}$, we need the following key technical lemma:

Lemma 5 *Let $\widehat{\mathcal{C}}_{k,h}$ be defined in (9) and set $\widehat{\beta}_k$ as*

$$\widehat{\beta}_k = 8\sqrt{d\log(1+k/\lambda)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B\,. \qquad (14)$$

Then, with probability at least $1-3\delta$, we have that simultaneously for all $k \in [K]$ and $h \in [H]$,

$$\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h},\ |[\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)| \le E_{k,h},$$

where $E_{k,h}$ is defined as follows:

$$E_{k,h} = \min\Big\{H^2, 2H\check{\beta}_k\Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\|_2\Big\} + \min\Big\{H^2, \widetilde{\beta}_k\Big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\Big\|_2\Big\}, \qquad (15)$$

with

$$\begin{aligned}\check{\beta}_k &= 8d\sqrt{\log(1+k/\lambda)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B, \\ \widetilde{\beta}_k &= 8\sqrt{dH^4\log(1+kH^4/(d\lambda))\log(4k^2H/\delta)} + 4H^2\log(4k^2H/\delta) + \sqrt{\lambda}B.\end{aligned}$$

Proof See Appendix D.1. ■

Lemma 5 shows that with high probability, for all stages h and episodes k, $\boldsymbol{\theta}_h^*$ lies in the confidence set centered at its estimate $\widehat{\boldsymbol{\theta}}_{k,h}$, and the error between the estimated variance and the true variance is bounded by the offset term $E_{k,h}$. Equipped with Lemma 5, we can verify the following facts: First, since $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$, it can be easily verified that $\big\langle \widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\big\rangle + \widehat{\beta}_k \big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\big\|_2 \geq \big\langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\big\rangle = [\mathbb{P}_h V_{k,h+1}](\cdot,\cdot)$, which shows that our constructed $Q_{k,h}(\cdot,\cdot)$ in Line 4 is indeed an overestimate of $Q_h^*(\cdot,\cdot)$. Second, recalling the definition of $\bar{\sigma}_{k,h}$ defined in Line 12, since $\big|[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k,a_h^k)\big| \leq E_{k,h}$, we have $\bar{\sigma}_{k,h}^2 \geq [\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k) + E_{k,h} \geq [\mathbb{V}_h V_{k,h+1}](s_h^k,a_h^k)$, which shows that $\bar{\sigma}_{k,h}$ is indeed an overestimate of the true variance $[\mathbb{V}_h V_{k,h+1}](s_h^k,a_h^k)$.

5.2. Regret Upper Bound

Now we present the regret upper bound of UCRL-VTR$^+$.

Theorem 6 *Set $\lambda = 1/B^2$. Then, with probability at least $1-5\delta$, the regret of UCRL-VTR$^+$ on MDP $M_{\boldsymbol{\Theta}^*}$ is upper bounded as follows:*

$$\textit{Regret}\big(M_{\boldsymbol{\Theta}^*}, K\big) = \widetilde{O}\Big(\sqrt{d^2H^2 + dH^3}\sqrt{T} + d^2H^3 + d^3H^2\Big),\ T = KH. \tag{16}$$

Proof [Sketch] The detailed proof is given in Appendix D.2. By Lemma 5, it suffices to prove the result on the event $\mathcal{E}$ when the conclusions of this lemma hold. Hence, in what follows assume that this event holds. By using the standard regret decomposition and using the definition of the confidence sets $\{\widehat{\mathcal{C}}_{k,h}\}_{k,h}$, we can show that the total regret is bounded by the summation of the bonus terms, $\sum_{k=1}^K\sum_{h=1}^H \widehat{\beta}_k\big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\big\|_2$, which, by the Cauchy-Schwarz inequality, can be further bounded by $\widehat{\beta}_K\sqrt{dH\sum_{k=1}^K\sum_{h=1}^H\bar{\sigma}_{k,h}^2}$. Finally, by the definition of $\bar{\sigma}_{k,h}^2$ we have $\bar{\sigma}_{k,h}^2 \leq H^2/d + E_{k,h} + [\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k) \leq H^2/d + 2E_{k,h} + [\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)$. Therefore the summation of $\bar{\sigma}_{k,h}^2$ can be bounded as

$$\begin{aligned}\sum_{k=1}^K\sum_{h=1}^H\bar{\sigma}_{k,h}^2 &\leq H^3K/d + 2\sum_{k=1}^K\sum_{h=1}^H E_{k,h} + \sum_{k=1}^K\sum_{h=1}^H[\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)\\ &= \widetilde{O}(HT + H^2T/d + dH^3\sqrt{T}),\end{aligned} \tag{17}$$

where the equality holds since by the *law of total variance* (Lattimore and Hutter, 2012; Azar et al., 2013), $\sum_{k=1}^K\sum_{h=1}^H[\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k) = \widetilde{O}(HT)$, and $\sum_{k=1}^K\sum_{h=1}^H E_{k,h} = \widetilde{O}(dH^3\sqrt{T} + d^{1.5}H^{2.5}\sqrt{T})$ by the elliptical potential lemma. ■

Remark 7 *When $d \geq H$ and $T \geq d^4H^2 + d^3H^3$, the regret in (16) can be simplified to $\widetilde{O}(dH\sqrt{T})$. Compared with the regret $\widetilde{O}(dH^{3/2}\sqrt{T})$ of UCRL-VTR in Jia et al. (2020); Ayoub et al. (2020)[3], the regret of UCRL-VTR$^+$ is improved by a factor of $\sqrt{H}$.*

[3] Jia et al. (2020); Ayoub et al. (2020) report a regret of order $\widetilde{O}(dH\sqrt{T})$. However, these works considered the time-homogeneous case where $\mathbb{P}_1 = \cdots = \mathbb{P}_H$. In particular, in the time-homogeneous setting parameters are shared between the stages of an episode, and this reduces the regret. When UCRL-VTR is modified for the inhomogenous case, the regret picks up an additional $\sqrt{H}$ factor. Similar observation has also been made by Jin et al. (2018).

5.3. Lower Bound

In this subsection, we present a lower bound for episodic linear mixture MDPs, which shows the optimality of UCRL-VTR$^+$.

Theorem 8 *Let $B > 1$ and suppose $K \geq \max\{(d-1)^2H/2, (d-1)/(32H(B-1))\}$, $d \geq 4$, $H \geq 3$. Then for any algorithm there exists an episodic, B-bounded linear mixture MDP parameterized by $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_H)$ such that the expected regret is lower bounded as follows:*

$$\mathbb{E}_{\boldsymbol{\Theta}} \textit{Regret}\big(M_{\boldsymbol{\Theta}}, K\big) \geq \Omega\big(dH\sqrt{T}\big),$$

where $T = KH$ and $\mathbb{E}_{\boldsymbol{\Theta}}$ denotes the expectation over the probability distribution generated by the interconnection of the algorithm and the MDP.

Proof [Sketch] We construct a hard-to-learn MDP instance M. The detailed construction and proof are given in Appendix E.1 and E.2. We show that learning the optimal policy of such an MDP is no *harder* than minimizing the regret on H linear bandit problems, where the payoff for the first $H/2$ bandits is $\Omega(H)Z$. Here Z is a Bernoulli random variable with mean equal to $\Theta(1/H)$. Utilizing existing lower bound results for linear bandits (Lattimore and Szepesvári, 2020) yields our result. ■

Remark 9 *Theorem 8 shows that for any algorithm running on episodic linear mixture MDPs, its regret is lower bounded by $\Omega(dH\sqrt{T})$. The lower bound together with the upper bound of UCRL-VTR$^+$ in Theorem 6 shows that UCRL-VTR$^+$ is minimax optimal up to logarithmic factors.*

6. Conclusion and Future Work

In this paper, we proposed a new Bernstein-type concentration inequality for self-normalized vector-valued martingales, which was shown to tighten existing confidence sets for linear bandits when the reward noise has low variance σ_t^2 and is almost surely uniformly bounded by a constant $R > 0$. This also allowed us to derive a bandit algorithm for the stochastic linear bandit problem with changing actions sets. The proposed algorithm uses weighted least-squares estimates and achieves a second-order regret bound of order $\widetilde{O}(R\sqrt{dT} + d\sqrt{\sum_{t=1}^T \sigma_t^2})$, which is a significant improvement on the dimension dependence in the low-noise regime. Based on the new tail inequality, we propose a new, computationally efficient algorithm, UCRL-VTR$^+$ for episodic MDPs with an $\widetilde{O}(dH\sqrt{T} + \sqrt{dH^3}\sqrt{T} + d^2H^3 + d^3H^2)$ regret.

We would like to point out that our current regret bound is nearly minimax optimal only for the "large dimension" and "large sample" cases. In particular, UCRL-VTR$^+$ is nearly minimax optimal only when $d \geq H$ and $T \geq d^4H^2 + d^3H^3$. It remains to be seen whether the range-restrictions on the dimension and the sample size can be loosened or altogether eliminated.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. Dongruo Zhou and Quanquan Gu are partially supported by the National Science Foundation CAREER Award 1906169 and IIS-1904183. Csaba Szepesvári gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, Amii and NSERC. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83, 2020.

J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Alex Ayoub, Zeyu Jia, Csaba Szepesvári, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474, 2020.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3): 325–349, 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294, 2020.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, pages 2137–2143, 2008.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. In *Advances in neural information processing systems*, pages 1422–1432, 2018.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2019.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060, 2020.

D.A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Minimax optimal reinforcement learning for discounted MDPs. *arXiv preprint arXiv:2010.00587*, 2020.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, 2021.

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

Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *L4DC*, 2020.

Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398, 2018.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. 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, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4868–4878, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384, 2018.

T. Lattimore, K. Crammer, and Cs. Szepesvári. Linear multi-resource allocation with semi-bandit feedback. In *Advances in Neural Information Processing Systems*, pages 964–972, September 2015.

Tor Lattimore and Marcus Hutter. PAC bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvári, and Gellért Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *International Conference on Machine Learning*, pages 5662–5670, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, pages 2173–2174, 2019.

Yingkai Li, Yining Wang, Xi Chen, and Yuan Zhou. Tight regret bounds for infinite-armed linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 370–378. PMLR, 2021.

Odalric-Ambrym Maillard, Timothy A Mann, and Shie Mannor. How hard is my mdp?" the distribution-norm to the rescue". *Advances in Neural Information Processing Systems*, 27:1835–1843, 2014.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *COLT*, 2009.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020, 2020.

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

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems 31*, pages 5192–5202, 2018.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, pages 1153–1162, 2019.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pages 2898–2933, 2019.

Richard S. Sutton and Andrew G. Barto. *Introduction to reinforcement learning*, volume 135. MIT Press, Cambridge, 1998.

Aristide Tossou, Debabrota Basu, and Christos Dimitrakakis. Near-optimal optimistic reinforcement learning using empirical Bernstein inequalities. *arXiv preprint arXiv:1905.12425*, 2019.

Ruosong Wang, Simon S Du, Lin Yang, and Sham Kakade. Is long horizon rl more difficult than short horizon rl? *Advances in Neural Information Processing Systems*, 33, 2020a.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33, 2020b.

Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2020c.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

Y. Wu, A. György, and Cs. Szepesvári. Online learning with Gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, pages 1360–1368, September 2015.

Lin Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004, 2019.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964, 2020a.

Andrea Zanette, Alessandro Lazaric, Mykel J. Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent Bellman error. In *International Conference on Machine Learning*, pages 10978–10989, 2020b.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, pages 2827–2836, 2019.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Advances in Neural Information Processing Systems 33*, 2020.

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? A near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, 2021a.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*, 2021b.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted MDPs with feature mapping. In *International Conference on Machine Learning*, 2021.

Appendix A. Additional Discussions in Section 4

A.1. Comparison between Theorem 2 and previous results

It is worth to compare Theorem 2 with a few Hoeffding-Azuma-type results proved in prior work (Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011). In particular, Dani et al. (2008) considered the setting where η_t is R-bounded and showed that for large enough t, the following holds with probability at least $1-\delta$:

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} \leq R\max\{\sqrt{128d\log(tL^2)\log(t^2/\delta)}, 8/3\cdot\log(t^2/\delta)\}.$$

Rusmevichientong and Tsitsiklis (2010) considered a more general setting than Dani et al. (2008) where η_t is R-sub-Gaussian and showed that (3) holds when $\beta_t = 2\kappa^2 R\sqrt{\log t}\sqrt{d\log t + \log(t^2/\delta)}$, where $\kappa = \sqrt{3+2\log(L^2/\lambda + d)}$. Abbasi-Yadkori et al. (2011) considered the same setting as Rusmevichientong and Tsitsiklis (2010) where η_t is R-sub-Gaussian and showed that (3) holds when $\beta_t = R\sqrt{d\log((1+tL^2/\lambda)/\delta)}$, which improves the bound of Rusmevichientong and Tsitsiklis (2010) in terms of logarithmic factors. By selecting proper λ, all these results yield an $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} = \widetilde{O}(R\sqrt{d})$ bound. As a comparison, with the choice $\lambda = \sigma^2 d/\|\boldsymbol{\mu}^*\|_2^2$, our result gives

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} = \widetilde{O}(\sigma\sqrt{d} + R). \tag{18}$$

Note that for any random variable, its standard deviation is always upper bounded by its magnitude or sub-Gaussian norm, therefore our result strictly improves the mentioned previous results. This improvement is due to the fact that here we consider a subclass of sub-Gaussian noise variables which allows us to derive a tighter upper bound. Indeed, Exercise 20.1 in the book of Lattimore and Szepesvári (2020) shows that the previous inequalities are tight in the worst-case for R-sub-Gaussian noise.

Even more closely related are results by Lattimore et al. (2015); Kirschner and Krause (2018) and Faury et al. (2020). In all these papers the strategy is to use a weighted ridge regression estimator, which we will also make use of in the next section. In particular, Lattimore et al. (2015) study the special case of Bernoulli payoffs. For this special case, with our notation, they show a result implying that with high probability $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} = \widetilde{O}(\sigma\sqrt{d})$. The lack of the scale term R is due to that Bernoulli's are single-parameter: The variance and the mean control each other, which the proof exploits. As such, this result does not lead in a straightforward way to ours, where the scale and variance are independently controlled. A similar comment applies to the result of Kirschner and Krause (2018) who considered the case when the noise in the responses are sub-Gaussian.

For the case of $R=1$, $L=1$ and $\mathbb{E}[\eta_t^2|\mathcal{G}_t] \leq \sigma_t^2$, the recent work of Faury et al. (2020) also proposed a Bernstein-type concentration inequality (cf. Theorem 1 in their paper) and showed that this gives rise to better results in the context of logistic bandits. Their result can be extended to arbitrary R and L (see Appendix A.2), which gives that with high probability,

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} = \widetilde{O}\big(\sigma\sqrt{d} + \sqrt{d\|\boldsymbol{\mu}^*\|_2 RL}\big), \tag{19}$$

where the second term in (19) has a polynomial dependence on $d, \|\boldsymbol{\mu}^*\|_2, R, L$, whereas in (18) the second term is only a function of R. This is a significant difference. In particular, in the linear mixture MDP setting, we have $\sigma = \widetilde{O}(\sqrt{H})$, $\|\boldsymbol{\mu}^*\|_2 = O(B)$, $R = O(H)$ and $L = O(H)$. Plugging these into both bounds, we see that our new result gives $\widetilde{O}(\sqrt{dH} + H)$, while (19) gives the worse bound $\widetilde{O}(\sqrt{dH} + \sqrt{dB}H)$. As it will be clear from the further details of our derivations given in Section 5, as a result of the above difference, their bound would not result in a minimax optimal bound on the regret in our setting.

A.2. Derivation of the Bound in (19)

In this subsection, we derive the bound in (19) by the concentration inequality proved in Faury et al. (2020). The following proposition is a restatement of Theorem 1 in Faury et al. (2020).

Proposition 10 (Theorem 1, Faury et al. 2020) *Let $\{\mathcal{G}_t\}_{t=1}^\infty$ be a filtration, where $\mathbf{x}_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{G}_{t+1}$-measurable. Suppose $\eta_t, \mathbf{x}_t$ satisfy that*

$$|\eta_t| \le 1,\ \mathbb{E}[\eta_t|\mathcal{G}_t] = 0,\ \mathbb{E}[\eta_t^2|\mathcal{G}_t] \le \sigma_t^2,\ \|\mathbf{x}_t\|_2 \le 1,$$

Let $\mathbf{H}_t = \lambda\mathbf{I} + \sum_{i=1}^t \sigma_t^2\mathbf{x}_i\mathbf{x}_i^\top$. Then for any $0 < \delta < 1, \lambda > 0$, with probability at least $1-\delta$ we have

$$\forall t > 0,\ \left\|\sum_{i=1}^t \mathbf{x}_i\eta_i\right\|_{\mathbf{H}_t^{-1}} \le \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}}\log\left(\frac{\det(\mathbf{H}_t)^{1/2}\lambda^{-d/2}}{\delta}\right) + \frac{2d\log 2}{\sqrt{\lambda}}. \tag{20}$$

In the following, we first extend the above bound to the general case, where $|\eta_t| \le R, \mathbb{E}[\eta_t^2|\mathcal{G}_t] \le \sigma^2, \|\mathbf{x}_t\|_2 \le L$. In specific, we have

$$|\eta_t/R| \le 1,\ \mathbb{E}[\eta_t/R|\mathcal{G}_t] = 0,\ \mathbb{E}[\eta_t^2/R^2|\mathcal{G}_t] \le \sigma^2/R^2,\ \|\mathbf{x}_t/L\|_2 \le 1,$$

Therefore, by Proposition 10, let

$$\bar{\mathbf{H}}_t = \lambda\mathbf{I} + \sum_{i=1}^t \sigma^2\mathbf{x}_i\mathbf{x}_i^\top/(R^2L^2),$$

the following holds with probability at least $1-\delta$,

$$\begin{aligned}\forall t > 0,\ \left\|\bar{\mathbf{H}}_t^{-1/2}\sum_{i=1}^t \mathbf{x}_i\eta_i/(RL)\right\|_2 &\le \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}}\log\left(\det(\bar{\mathbf{H}}_t)^{1/2}\lambda^{-d/2}\right) + \frac{2d\log 2 + 2\log(1/\delta)}{\sqrt{\lambda}} \\ &\le \frac{\sqrt{\lambda}}{2} + \frac{d}{\sqrt{\lambda}}\log(1+t\sigma^2/\lambda) + \frac{2d\log 2 + 2\log(1/\delta)}{\sqrt{\lambda}},\end{aligned} \tag{21}$$

where the second inequality holds since $\det(\bar{\mathbf{H}}_t) \le \|\bar{\mathbf{H}}_t\|_2^d \le (\lambda + t\sigma^2)^d$. Set $\lambda \leftarrow \lambda\sigma^2/(R^2L^2)$, then (21) becomes

$$\forall t > 0,\ \left\|\sum_{i=1}^t \mathbf{x}_i\eta_i\right\|_{\mathbf{Z}_t^{-1}} \le \sigma\left(\frac{\sigma\sqrt{\lambda}}{2RL} + \frac{dRL}{\sigma\sqrt{\lambda}}\log(1+tR^2L^2/\lambda) + \frac{2d\log 2 + 2\log(1/\delta)}{\sigma\sqrt{\lambda}}RL\right), \tag{22}$$

Now we are going to bound $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t}$ by (22). By the definition of $\boldsymbol{\mu}_t$, we have

$$\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1}\mathbf{b}_t = \mathbf{Z}_t^{-1}\sum_{i=1}^t \mathbf{x}_i(\mathbf{x}_i^\top\boldsymbol{\mu}^* + \eta_i) = \boldsymbol{\mu}^* - \lambda\mathbf{Z}_t^{-1}\boldsymbol{\mu}^* + \mathbf{Z}_t^{-1}\sum_{i=1}^t \mathbf{x}_i\eta_i,$$

then $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t}$ can be bounded as

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} = \bigg\|\mathbf{Z}_t^{-1/2}\sum_{i=1}^t \mathbf{x}_i\eta_i + \lambda \mathbf{Z}_t^{-1/2}\boldsymbol{\mu}^*\bigg\|_2 \le \bigg\|\sum_{i=1}^t \mathbf{x}_i\eta_i\bigg\|_{\mathbf{Z}_t^{-1}} + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2, \tag{23}$$

where the first equality holds due to triangle inequality and $\mathbf{Z}_t \succeq \lambda\mathbf{I}$. Next, substituting (22) into (23) yields

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} \le \frac{\sigma^2\sqrt{\lambda}}{2RL} + \frac{dRL}{\sqrt{\lambda}}\log(1 + tR^2L^2/\lambda) + \frac{2d\log 2 + 2\log(1/\delta)}{\sqrt{\lambda}}RL + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2.$$

Finally, set $\lambda = \widetilde{\Theta}(dR^2L^2/(\sigma^2 + RL\|\boldsymbol{\mu}^*\|_2))$ to minimize the above upper bound, we have $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} \le \widetilde{O}(\sigma\sqrt{d} + \sqrt{dRL\|\boldsymbol{\mu}^*\|_2})$.

Appendix B. Proofs of Theorems in Section 4

B.1. Proof of Theorem 2

We follow the proof in Dani et al. (2008) with a refined analysis. Let us start with recalling two well known results that we will need:

Lemma 11 (Freedman 1975) *Let $M, v > 0$ be fixed constants. Let $\{x_i\}_{i=1}^n$ be a stochastic process,$\{\mathcal{G}_i\}_i$ be a filtration so that so that for all $i \in [n]$ x_i is $\mathcal{G}_i$-measurable, while almost surely $\mathbb{E}[x_i|\mathcal{G}_{i-1}] = 0$, $|x_i| \le M$ and*

$$\sum_{i=1}^n \mathbb{E}(x_i^2|\mathcal{G}_i) \le v\,.$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\sum_{i=1}^n x_i \le \sqrt{2v\log(1/\delta)} + 2/3 \cdot M\log(1/\delta).$$

Lemma 12 (Lemma 11, Abbasi-Yadkori et al. 2011) *For any $\lambda > 0$ and sequence $\{\mathbf{x}_t\}_{t=1}^T \subset \mathbb{R}^d$ for $t \in \{0, 1, \dots, T\}$, define $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^t \mathbf{x}_i\mathbf{x}_i^\top$. Then, provided that $\|\mathbf{x}_t\|_2 \le L$ holds for all $t \in [T]$, we have*

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

Recall that for $t \ge 0$, $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^t \mathbf{x}_i\mathbf{x}_i^\top$. Since $\mathbf{Z}_t = \mathbf{Z}_{t-1} + \mathbf{x}_t\mathbf{x}_t^\top$, by the matrix inversion lemma

$$\mathbf{Z}_t^{-1} = \mathbf{Z}_{t-1}^{-1} - \frac{\mathbf{Z}_{t-1}^{-1}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{Z}_{t-1}^{-1}}{1 + w_t^2}. \tag{24}$$

We need the following definitions:

$$\mathbf{d}_0 = 0,\ Z_0 = 0,\ \mathbf{d}_t = \sum_{i=1}^{t} \mathbf{x}_i \eta_i,\ Z_t = \|\mathbf{d}_t\|_{\mathbf{Z}_t^{-1}},\ w_t = \|\mathbf{x}_t\|_{\mathbf{Z}_{t-1}^{-1}},\ \mathcal{E}_t = \mathbb{1}\{0 \le s \le t, Z_s \le \beta_s\}\,, \tag{25}$$

where $t \ge 1$ and we define $\beta_0 = 0$. Recalling that x_t is $\mathcal{G}_t$-measurable and η_t is $\mathcal{G}_{t+1}$-measurable, we find that d_t, Z_t and $\mathcal{E}_t$ are $\mathcal{G}_{t+1}$-measurable while w_t is $\mathcal{G}_t$ measurable. We now prove the following result:

Lemma 13 *Let $\mathbf{d}_i, w_i, \mathcal{E}_i$ be as defined in (25). Then, with probability at least $1 - \delta/2$, simultaneously for all $t \ge 1$ it holds that*

$$\sum_{i=1}^{t} \frac{2\eta_i \mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1} \mathbf{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1} \le 3\beta_t^2/4.$$

Proof We have

$$\left| \frac{2\mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1} \mathbf{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1} \right| \le \frac{2\|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}} [\|\mathbf{d}_{i-1}\|_{\mathbf{Z}_{i-1}^{-1}} \mathcal{E}_{i-1}]}{1 + w_i^2} \le \frac{2 w_i \beta_{i-1}}{1 + w_i^2} \le \min\{1, 2w_i\}\beta_{i-1}, \tag{26}$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to the definition of $\mathcal{E}_{i-1}$, the last inequality holds by algebra. For simplicity, let ℓ_i denote

$$\ell_i = \frac{2\eta_i \mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1} \mathbf{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1}. \tag{27}$$

We are preparing to apply Freedman's inequality from Lemma 11 to $\{\ell_i\}_i$ and $\{\mathcal{G}_i\}_i$. First note that $\mathbb{E}[\ell_i | \mathcal{G}_i] = 0$. Meanwhile, by (26), the inequalities

$$|\ell_i| \le R\beta_{i-1} \min\{1, 2w_i\} \le R\beta_{i-1} \le R\beta_t \tag{28}$$

almost surely hold (the last inequality follows since $\{\beta_i\}_i$ is increasing). We also have

$$\begin{aligned}
\sum_{i=1}^{t} \mathbb{E}[\ell_i^2 | \mathcal{G}_i] &\le \sigma^2 \sum_{i=1}^{t} \left(\frac{2\mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1} \mathbf{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1} \right)^2 \\
&\le \sigma^2 \sum_{i=1}^{t} [\min\{1, 2w_i\}\beta_{i-1}]^2 \\
&\le 4\sigma^2 \beta_t^2 \sum_{i=1}^{t} \min\{1, w_i^2\} \\
&\le 8\sigma^2 \beta_t^2 d \log(1 + tL^2/(d\lambda)),
\end{aligned} \tag{29}$$

where the first inequality holds since $\mathbb{E}[\eta_i^2 | \mathcal{G}_i] \le \sigma^2$, the second inequality holds due to (26), the third inequality holds again since $\{\beta_i\}_i$ is increasing, the last inequality holds due to Lemma 12.

Therefore, by (28) and (29), using Lemma 11, we know that for any t, with probability at least $1 - \delta/(4t^2)$, we have

$$\begin{aligned}\sum_{i=1}^{t} \ell_i &\leq \sqrt{16\sigma^2\beta_t^2 d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)} + 2/3 \cdot R\beta_t \log(4t^2/\delta)\\ &\leq \frac{\beta_t^2}{4} + 16\sigma^2 d\log(1+tL^2/(d\lambda))\log(4t^2/\delta) + \frac{\beta_t^2}{4} + 4R^2\log^2(4t^2/\delta)\\ &\leq \beta_t^2/2 + \frac{1}{4}\big(8\sigma\sqrt{d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)} + 4R\log(4t^2/\delta)\big)^2\\ &= 3\beta_t^2/4, \end{aligned} \tag{30}$$

where the first inequality holds due to Lemma 11, the second inequality holds due to $2\sqrt{|ab|} \leq |a| + |b|$, the last equality holds due to the definition of β_t. Taking union bound for (30) from $t = 1$ to ∞ and using the fact that $\sum_{t=1}^{\infty} t^{-2} < 2$ finishes the proof. ■

We also need the following lemma.

Lemma 14 *Let w_i be as defined in (25). Then, with probability at least $1 - \delta/2$, simultaneously for all t ≥ 1 it holds that*

$$\sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1+w_i^2} \leq \beta_t^2/4.$$

Proof We are preparing to apply Freedman's inequality (Lemma 11) to $\{\ell_i\}_i$ and $\{\mathcal{G}_i\}_i$ where now

$$\ell_i = \frac{\eta_i^2 w_i^2}{1+w_i^2} - \mathbb{E}\bigg[\frac{\eta_i^2 w_i^2}{1+w_i^2}\bigg|\mathcal{G}_i\bigg]. \tag{31}$$

Clearly, for any i, we have $\mathbb{E}[\ell_i|\mathcal{G}_i] = 0$ almost surely (a.s.). We further have that a.s.

$$\begin{aligned}\sum_{i=1}^{t} \mathbb{E}[\ell_i^2|\mathcal{G}_i] &\leq \sum_{i=1}^{t} \mathbb{E}\bigg[\frac{\eta_i^4 w_i^4}{(1+w_i^2)^2}\bigg|\mathcal{G}_i\bigg]\\ &\leq R^2\sum_{i=1}^{t} \mathbb{E}\bigg[\frac{\eta_i^2 w_i^2}{1+w_i^2}\bigg|\mathcal{G}_i\bigg]\\ &\leq R^2\sigma^2\sum_{i=1}^{t}\frac{w_i^2}{1+w_i^2}\\ &\leq 2R^2\sigma^2 d\log(1+tL^2/(d\lambda)), \end{aligned} \tag{32}$$

where the first inequality holds due to the fact $\mathbb{E}(X - \mathbb{E}X)^2 \leq \mathbb{E}X^2$, the second inequality holds since $|\eta_t| \leq R$ a.s., the third inequality holds since $\mathbb{E}[\eta_i^2|\mathcal{G}_i] \leq \sigma^2$ a.s. and w_i is $\mathcal{G}_i$-measurable, the fourth inequality holds due to the fact $w_i^2/(1+w_i^2) \leq \min\{1, w_i^2\}$ and Lemma 12. Furthermore, by the fact that $|\eta_i| \leq R$ a.s., we have

$$|\ell_i| \leq \bigg|\frac{\eta_i^2 w_i^2}{1+w_i^2}\bigg| + \bigg|\mathbb{E}\bigg[\frac{\eta_i^2 w_i^2}{1+w_i^2}\bigg|\mathcal{G}_i\bigg]\bigg| \leq 2R^2 \text{ a.s.} \tag{33}$$

Therefore, by (32) and (33), using Lemma 11, we know that for any t, with probability at least $1-\delta/(4t^2)$, we have that a.s.,

$$
\begin{aligned}
\sum_{i=1}^{t}\frac{\eta_i^2 w_i^2}{1+w_i^2} &\leq \sum_{i=1}^{t}\mathbb{E}\bigg[\frac{\eta_i^2 w_i^2}{1+w_i^2}\bigg|\mathcal{G}_i\bigg] + \sqrt{4R^2\sigma^2 d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)} + 4/3\cdot R^2\log(4t^2/\delta)\\
&\leq \sigma^2\sum_{i=1}^{t}\frac{w_i^2}{1+w_i^2} + 2R\sigma\sqrt{d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)} + 2R^2\log(4t^2/\delta)\\
&\leq 2\sigma^2 d\log(1+tL^2/(d\lambda)) + 2R\sigma\sqrt{d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)} + 2R^2\log(4t^2/\delta)\\
&\leq 1/4\cdot\big(8\sigma\sqrt{d}\sqrt{\log(1+tL^2/(d\lambda))\log(4t^2/\delta)} + 4R\log(4t^2/\delta)\big)^2\\
&= \beta_t^2/4,
\end{aligned}
\tag{34}
$$

where the first inequality holds due to Lemma 11, the second inequality holds due to $\mathbb{E}[\eta_i^2|\mathcal{G}_i]\leq\sigma^2$, the third inequality holds due to the fact $w_i^2/(1+w_i^2)\leq\min\{1,w_i^2\}$ and Lemma 12, the last inequality holds due to the definition of β_t. Taking union bound for (34) from $t=1$ to ∞ and using the fact that $\sum_{t=1}^{\infty}t^{-2}<2$ finishes the proof. ■

With this, we are ready to prove Theorem 2.

Proof [Proof of Theorem 2] We first give a crude upper bound on Z_t. We have

$$
\begin{aligned}
Z_t^2 &= (\mathbf{d}_{t-1}+\mathbf{x}_t\eta_t)^\top\mathbf{Z}_t^{-1}(\mathbf{d}_{t-1}+\mathbf{x}_t\eta_t)\\
&= \mathbf{d}_{t-1}^\top\mathbf{Z}_t^{-1}\mathbf{d}_{t-1} + 2\eta_t\mathbf{x}_t^\top\mathbf{Z}_t^{-1}\mathbf{d}_{t-1} + \eta_t^2\mathbf{x}_t^\top\mathbf{Z}_t^{-1}\mathbf{x}_t\\
&\leq Z_{t-1}^2 + \underbrace{2\eta_t\mathbf{x}_t^\top\mathbf{Z}_t^{-1}\mathbf{d}_{t-1}}_{I_1} + \underbrace{\eta_t^2\mathbf{x}_t^\top\mathbf{Z}_t^{-1}\mathbf{x}_t}_{I_2},
\end{aligned}
$$

where the inequality holds since $\mathbf{Z}_t\succeq\mathbf{Z}_{t-1}$. For term I_1, from the matrix inversion lemma (cf. (24)), we have

$$
\begin{aligned}
I_1 &= 2\eta_t\bigg(\mathbf{x}_t^\top\mathbf{Z}_{t-1}^{-1}\mathbf{d}_{t-1} - \frac{\mathbf{x}_t^\top\mathbf{Z}_{t-1}^{-1}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{Z}_{t-1}^{-1}\mathbf{d}_{t-1}}{1+w_t^2}\bigg)\\
&= 2\eta_t\bigg(\mathbf{x}_t^\top\mathbf{Z}_{t-1}^{-1}\mathbf{d}_{t-1} - \frac{w_t^2\mathbf{x}_t^\top\mathbf{Z}_{t-1}^{-1}\mathbf{d}_{t-1}}{1+w_t^2}\bigg)\\
&= \frac{2\eta_t\mathbf{x}_t^\top\mathbf{Z}_{t-1}^{-1}\mathbf{d}_{t-1}}{1+w_t^2}.
\end{aligned}
$$

For term I_2, again from the matrix inversion lemma (cf. (24)), we have

$$
I_2 = \eta_t^2\bigg(\mathbf{x}_t^\top\mathbf{Z}_{t-1}^{-1}\mathbf{x}_t^\top - \frac{\mathbf{x}_t^\top\mathbf{Z}_{t-1}^{-1}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{Z}_{t-1}^{-1}\mathbf{x}_t}{1+w_t^2}\bigg) = \eta_t^2\bigg(w_t^2 - \frac{w_t^4}{1+w_t^2}\bigg) = \frac{\eta_t^2 w_t^2}{1+w_t^2}.
$$

Therefore, we have

$$
Z_t^2 \leq \sum_{i=1}^{t}\frac{2\eta_i\mathbf{x}_i^\top\mathbf{Z}_{i-1}^{-1}\mathbf{d}_{i-1}}{1+w_i^2} + \sum_{i=1}^{t}\frac{\eta_i^2 w_i^2}{1+w_i^2}. \tag{35}
$$

Consider now the event $\mathcal{E}$ where the conclusions of Lemma 13 and Lemma 14 hold. We claim that on this event for any $i \ge 0$, $Z_i \le \beta_i$. We prove this by induction on i. Let the said event hold. The base case of $i = 0$ holds since $\beta_0 = 0 = Z_0$, by definition. Now fix some $t \ge 1$ and assume that for all $0 \le i < t$, we have $Z_i \le \beta_i$. This implies that $\mathcal{E}_1 = \mathcal{E}_2 = \cdots = \mathcal{E}_{t-1} = 1$. Then by (35), we have

$$Z_t^2 \le \sum_{i=1}^{t} \frac{2\eta_i \mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1}\mathbf{d}_{i-1}}{1+w_i^2} + \sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1+w_i^2} = \sum_{i=1}^{t} \frac{2\eta_i \mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1}\mathbf{d}_{i-1}}{1+w_i^2}\mathcal{E}_{i-1} + \sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1+w_i^2}. \tag{36}$$

Since on the event $\mathcal{E}$ the conclusions of Lemma 13 and Lemma 14 hold, we have

$$\sum_{i=1}^{t} \frac{2\eta_i \mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1}\mathbf{d}_{i-1}}{1+w_i^2}\mathcal{E}_{i-1} \le 3\beta_t^2/4, \ \sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1+w_i^2} \le \beta_t^2/4. \tag{37}$$

Therefore, substituting (37) into (36), we have $Z_t \le \beta_t$, which ends the induction. Taking the union bound, the events in Lemma 13 and Lemma 14 hold with probability at least $1-\delta$, which implies that with probability at least $1-\delta$, for any t, $Z_t \le \beta_t$.

Finally, we bound $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t}$ as follows. First,

$$\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1}\mathbf{b}_t = \mathbf{Z}_t^{-1}\sum_{i=1}^{t} \mathbf{x}_i(\mathbf{x}_i^\top \boldsymbol{\mu}^* + \eta_i) = \boldsymbol{\mu}^* - \lambda \mathbf{Z}_t^{-1}\boldsymbol{\mu}^* + \mathbf{Z}_t^{-1}\mathbf{d}_t\,.$$

Then, on $\mathcal{E}$ we have

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} = \big\|\mathbf{d}_t - \lambda\boldsymbol{\mu}^*\big\|_{\mathbf{Z}_t^{-1}} \le Z_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2 \le \beta_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2, \tag{38}$$

where the first inequality holds due to triangle inequality and $\mathbf{Z}_t \succeq \lambda\mathbf{I}$, while the last one holds since we have shown that on $\mathcal{E}$, $Z_t \le \beta_t$ for all $t \ge 0$, thus finishing the proof.

■

B.2. Proof of Theorem 3

Proof [Proof of Theorem 3] By the assumption on ϵ_t, we know that

$$|\epsilon_t/\bar{\sigma}_t| \le R/\bar{\sigma}^t_{\text{min}},\ \mathbb{E}[\epsilon_t|\mathbf{a}_{1:t},\epsilon_{1:t-1}] = 0,\ \mathbb{E}[(\epsilon_t/\bar{\sigma}_t)^2|\mathbf{a}_{1:t},\epsilon_{1:t-1}] \le 1,\ \|\mathbf{a}_t/\bar{\sigma}_t\|_2 \le A/\bar{\sigma}^t_{\text{min}},$$

Then, taking $\mathcal{G}_t = \sigma(\mathbf{a}_{1:t},\epsilon_{1:t-1})$, using that σ_t is $\mathcal{G}_t$-measurable, we can apply Theorem 2 to $(\boldsymbol{x}_t,\eta_t) = (\boldsymbol{a}_t/\sigma_t, \epsilon_t/\sigma_t)$ to get that with probability at least $1-\delta$,

$$\forall t \ge 1,\ \big\|\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*\big\|_{\mathbf{A}_t} \le \widehat{\beta}_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2 \le \widehat{\beta}_t + \sqrt{\lambda}B, \tag{39}$$

where $\widehat{\beta}_t = 8\sqrt{d\log(1+tA^2/([\bar{\sigma}^t_{\text{min}}]^2 d\lambda))\log(4t^2/\delta)} + 4R/\bar{\sigma}^t_{\text{min}}\cdot\log(4t^2/\delta)$. Thus, in the remainder of the proof, we will assume that the event $\mathcal{E}$ when (39) is true holds and proceed to bound the regret on this event.

Note that on $\mathcal{E}$, $\boldsymbol{\mu}^* \in \mathcal{C}_t$. Recall that $\widetilde{\boldsymbol{\mu}}_t$ is the optimistic parameter choice of the algorithm (cf. Line 4 in Algorithm 1). Then, using the standard argument for linear bandits, the pseudo-regret for round t is bounded by

$$\langle \mathbf{a}_t^*, \boldsymbol{\mu}^*\rangle - \langle \mathbf{a}_t, \boldsymbol{\mu}^*\rangle \le \langle \mathbf{a}_t, \widetilde{\boldsymbol{\mu}}_t\rangle - \langle \mathbf{a}_t, \boldsymbol{\mu}^*\rangle = \langle \mathbf{a}_t, \widetilde{\boldsymbol{\mu}}_t - \widehat{\boldsymbol{\mu}}_{t-1}\rangle + \langle \mathbf{a}_t, \widehat{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\mu}^*\rangle, \tag{40}$$

where the inequality holds due to the choice $\widetilde{\boldsymbol{\mu}}_t$. To further bound (40), we have

$$\begin{aligned}
&\langle \mathbf{a}_t, \widetilde{\boldsymbol{\mu}}_t - \widehat{\boldsymbol{\mu}}_{t-1}\rangle + \langle \mathbf{a}_t, \widehat{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\mu}^*\rangle \\
&\quad \leq \|\mathbf{a}_t\|_{\mathbf{A}_{t-1}^{-1}}(\|\widetilde{\boldsymbol{\mu}}_t - \widehat{\boldsymbol{\mu}}_{t-1}\|_{\mathbf{A}_{t-1}} + \|\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_{t-1}\|_{\mathbf{A}_{t-1}}) \\
&\quad \leq 2(\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\|\mathbf{a}_t\|_{\mathbf{A}_{t-1}^{-1}}, \qquad (41)
\end{aligned}$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second one holds since $\widetilde{\boldsymbol{\mu}}_t, \boldsymbol{\mu}^* \in \mathcal{C}_{t-1}$. Meanwhile, we have $0 \leq \langle \mathbf{a}_t^*, \boldsymbol{\mu}^*\rangle - \langle \mathbf{a}_t, \boldsymbol{\mu}^*\rangle \leq 2$. Thus, substituting (41) into (40) and summing up (40) for $t = 1, \ldots, T$, we have

$$\text{Regret}(T) = \sum_{t=1}^T \big[\langle \mathbf{a}_t^*, \boldsymbol{\mu}^*\rangle - \langle \mathbf{a}_t, \boldsymbol{\mu}^*\rangle\big] \leq 2\sum_{t=1}^T \min\Big\{1, \bar{\sigma}_t(\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}\Big\}. \qquad (42)$$

To further bound the right-hand side above, we decompose the set $[T]$ into a union of two disjoint subsets $[T] = \mathcal{I}_1 \cup \mathcal{I}_2$, where

$$\mathcal{I}_1 = \Big\{t \in [T] : \|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}} \geq 1\Big\},\ \mathcal{I}_2 = [T] \setminus \mathcal{I}_1. \qquad (43)$$

Then the following upper bound of $|\mathcal{I}_1|$ holds:

$$|\mathcal{I}_1| \leq \sum_{t \in \mathcal{I}_1} \min\Big\{1, \|\mathbf{a}_t/\bar{\sigma}_t\|^2_{\mathbf{A}_{t-1}^{-1}}\Big\} \leq \sum_{t=1}^T \min\Big\{1, \|\mathbf{a}_t/\bar{\sigma}_t\|^2_{\mathbf{A}_{t-1}^{-1}}\Big\} \leq 2d\log(1 + TA^2/(d\lambda[\bar{\sigma}_{\text{min}}^T]^2)), \qquad (44)$$

where the first inequality holds since $\|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}} \geq 1$ for $t \in \mathcal{I}_1$, the third inequality holds due to Lemma 12 together with the fact $\|\mathbf{a}_t/\bar{\sigma}_t\|_2 \leq A/\bar{\sigma}_{\text{min}}^T$. Therefore, by (42),

$$\begin{aligned}
&\text{Regret}(T)/2 = \\
&\quad \sum_{t \in \mathcal{I}_1} \min\Big\{1, \bar{\sigma}_t(\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}\Big\} + \sum_{t \in \mathcal{I}_2} \min\Big\{1, \bar{\sigma}_t(\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}\Big\} \\
&\quad \leq \Bigg[\sum_{t \in \mathcal{I}_1} 1\Bigg] + \sum_{t \in \mathcal{I}_2} (\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\bar{\sigma}_t\|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}} \\
&\quad = |\mathcal{I}_1| + \sum_{t \in \mathcal{I}_2} (\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\bar{\sigma}_t \min\Big\{1, \|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}\Big\} \\
&\quad \leq 2d\log(1 + TA^2/(d\lambda[\bar{\sigma}_{\text{min}}^T]^2)) + \sum_{t=1}^T (\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\bar{\sigma}_t \min\Big\{1, \|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}\Big\}, \qquad (45)
\end{aligned}$$

where the first inequality holds since for any x real, $\min\{1, x\} \leq 1$ and also $\min\{1, x\} \leq x$, the second inequality holds since $\|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}} \leq 1$ for $t \in \mathcal{I}_2$ and the last one holds due to (44). Finally, to further bound (45), notice that

$$\sum_{t=1}^T (\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\bar{\sigma}_t \min\Big\{1, \|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}\Big\}$$

$$
\begin{aligned}
&\leq \sqrt{\sum_{t=1}^{T}(\widehat{\beta}_{t-1}+\sqrt{\lambda}B)^2\bar{\sigma}_t^2}\sqrt{\sum_{t=1}^{T}\min\Big\{1, \|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}^2\Big\}}\\
&\leq \sqrt{\sum_{t=1}^{T}(\widehat{\beta}_{t-1}+\sqrt{\lambda}B)^2\bar{\sigma}_t^2}\sqrt{2d\log(1+TA^2/(d\lambda[\bar{\sigma}_{\text{min}}^T]^2))}, \qquad (46)
\end{aligned}
$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second one holds due to Lemma 12 and the the fact that $\|\mathbf{a}_t/\sigma_t\|_2 \leq A/\bar{\sigma}_{\text{min}}^T$. Substituting (46) into (45), we have

$$
\begin{aligned}
\text{Regret}(T) \leq\, & 2\sqrt{2d\log(1+TA^2/(d\lambda[\bar{\sigma}_{\text{min}}^T]^2))}\sqrt{\sum_{t=1}^{T}(\widehat{\beta}_{t-1}+\sqrt{\lambda}B)^2\bar{\sigma}_t^2}\\
&+4d\log\big(1+TA^2/(d\lambda[\bar{\sigma}_{\text{min}}^T]^2)\big), \qquad (47)
\end{aligned}
$$

Next, since $\bar{\sigma}_t = \max\{R/\sqrt{d}, \sigma_t\}$, then we have $\bar{\sigma}_{\text{min}}^t \geq R/\sqrt{d}$. Therefore, with $\lambda = 1/B^2$, we have

$$
\log(1+TA^2/(d\lambda[\bar{\sigma}_{\text{min}}^T]^2)) \leq \log(1+TB^2A^2/R^2) = \widetilde{O}(1), \qquad (48)
$$

and

$$
\begin{aligned}
\widehat{\beta}_t+\sqrt{\lambda}B &= 8\sqrt{d\log(1+tA^2/([\bar{\sigma}_{\text{min}}^t]^2d\lambda))\log(4t^2/\delta)}+4R/\bar{\sigma}_{\text{min}}^t\cdot\log(4t^2/\delta)+\sqrt{\lambda}B\\
&\leq 8\sqrt{d\log(1+TB^2A^2/R^2)\log(4T^2/\delta)}+4\sqrt{d}\log(4T^2/\delta)+1\\
&= \widetilde{O}(\sqrt{d}). \qquad (49)
\end{aligned}
$$

Substituting (48) and (49) into (47), we have our second result.

$$
\text{Regret}(T) = \widetilde{O}\bigg(d\sqrt{\sum_{t=1}^{T}\bar{\sigma}_t^2}\bigg) = \widetilde{O}\bigg(d\sqrt{\sum_{t=1}^{T}(R^2/d+\sigma_t^2)}\bigg) = \widetilde{O}\bigg(R\sqrt{dT}+d\sqrt{\sum_{t=1}^{T}\sigma_t^2}\bigg),
$$

where the second equality holds since $\bar{\sigma}_t^2 = \max\{R^2/d, \sigma_t^2\} \leq R^2/d+\sigma_t^2$, the third equality holds since $\sqrt{|x|+|y|} \leq \sqrt{|x|}+\sqrt{|y|}$. ■

Appendix C. Additional Discussions on UCRL-VTR$^+$

C.1. UCRL-VTR$^+$ with single estimation sequence

Currently UCRL-VTR$^+$ uses two estimate sequences $\check{\boldsymbol{\theta}}_{k,h}$ and $\widetilde{\boldsymbol{\theta}}_{k,h}$ to estimate the first-order moment $\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\rangle$ and second-order moment $\langle\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\rangle$ separately. We would like to point out that it is possible to use only one sequence to estimate both. Such an estimator can be constructed as a weighted ridge regression estimator based on both $\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)$'s and $\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)$, and the corresponding responses $V_{k,h+1}(s_{h+1}^k)$ and $V_{k,h+1}^2(s_{h+1}^k)$. However, since second-order moments generally have larger variance than the first-order moments, we need to use different weights for the square loss evaluated at $\{\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), V_{k,h+1}(s_{h+1}^k)\}_{k,h}$ and $\{\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), V_{k,h+1}^2(s_{h+1}^k)\}_{k,h}$. Also, by merging the data, even with using perfect weighting, we would expect to win at best a (small) constant factor on the regret since the effect of not merging the data can be seen as not worse than throwing away "half of the data". As a result, for the sake of simplicity, we chose to use two estimate sequences instead of one in our algorithm.

C.2. Computational Efficiency of UCRL-VTR$^+$

Similar to UCRL-VTR (Ayoub et al., 2020), the computational complexity of UCRL-VTR$^+$ depends on the specific family of feature mapping $\boldsymbol{\phi}(\cdot|\cdot,\cdot)$. As an example, let us consider a special class of linear mixture MDPs studied by Yang and Wang (2020); Zhou et al. (2021). In this setting, $\boldsymbol{\phi}(s'|s,a) = \boldsymbol{\psi}(s') \odot \boldsymbol{\mu}(s,a)$, $\boldsymbol{\psi}(\cdot) : \mathcal{S} \to \mathbb{R}^d$ and $\boldsymbol{\mu}(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ are two features maps and $\odot$ denotes componentwise product. Recall that, by assumption, the action space $\mathcal{A}$ is finite.

We now argue that UCRL-VTR$^+$ is computationally efficient for this class of MDPs as long as we have access to an integration oracle $\mathcal{O}$ underlying the basis kernels. In particular, the assumption is that $\sum_{s'} \psi(s')V(s')$ can be evaluated at the cost of evaluating V at $p(d)$ states with some polynomial p. Now, for $1 \le h \le H$, $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ let

$$Q_{h,\boldsymbol{\theta},\boldsymbol{\Sigma}}(\cdot,\cdot) = \big[r_h(\cdot,\cdot) + \langle \boldsymbol{\theta}, \boldsymbol{\mu}(\cdot,\cdot)\rangle + \|\boldsymbol{\Sigma}\boldsymbol{\mu}(\cdot,\cdot)\|_2\big]_{[0,H]}.$$

It is easy to verify that for any k, h, $Q_{k,h} = Q_{h,\boldsymbol{\theta}_{k,h},\boldsymbol{\Sigma}_{k,h}}$ where $\boldsymbol{\theta}_{k,h} = \widehat{\boldsymbol{\theta}}_{k,h} \odot [\sum_{s'} \boldsymbol{\psi}(s')V_{k,h+1}(s')]$ and the (i,j)-th entry of $\boldsymbol{\Sigma}_{k,h}$ is $\widehat{\beta}_k(\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2})_{i,j}[\sum_{s'} \boldsymbol{\psi}_j(s')V_{k,h+1}(s')]$. Now notice that $\boldsymbol{\theta}_{k,H} = \mathbf{0}$, $\boldsymbol{\Sigma}_{k,H} = \mathbf{0}$. Thus, for $1 \le h \le H-1$, assuming that $\boldsymbol{\theta}_{k,h+1}$ and $\boldsymbol{\Sigma}_{k,h+1}$ have been calculated, evaluating $V_{k,h+1}$ at any state $s \in \mathcal{S}$ costs $O(d^2|\mathcal{A}|)$ arithmetic operations. Now, calculating $\boldsymbol{\theta}_{k,h}$ and $\boldsymbol{\Sigma}_{k,h}$ costs $O(d^2)$ arithmetic operations given access $\widehat{\boldsymbol{\theta}}_{k,h}$ and $\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}$, in addition to $p(d)$ evaluations of $V_{k,h+1}$. Since each evaluation of $V_{k,h+1}$ takes $O(d^2|\mathcal{A}|)$ operations, as established, calculating $\boldsymbol{\theta}_{k,h}$ and $\boldsymbol{\Sigma}_{k,h}$ cost a total of $O(p(d)d^2|\mathcal{A}|)$ operations. From this, it is clear that calculating the H actions to be taken in episode k takes a total of $O(p(d)d^2|\mathcal{A}|H)$ operations (Line 10). It also follows that calculating either $\boldsymbol{\phi}_{V_{k,h+1}}$ or $\boldsymbol{\phi}_{V_{k,h+1}^2}$ at any state-action pair costs $O(p(d)d^2|\mathcal{A}|)$ operations.

To calculate the quantities appearing in Lines 11–17, first $\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)$ and $\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)$ ($h \in [H]$) are evaluated at the cost of $O(p(d)d^2|\mathcal{A}|H)$. It is then clear that the rest of the calculation costs at most $O(d^3H)$: the most expensive step is to obtain $\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}$ (the cost could be reduced to $O(d^2H)$ by using the matrix inversion lemma and organizing the calculation of $Q_{k,h}$ slightly differently). It follows that the total computational complexity of UCRL-VTR$^+$ is $O(\text{poly}(d)|\mathcal{A}|HK) = O(\text{poly}(d)|\mathcal{A}|T)$. For many other MDP models, UCRL-VTR$^+$ can still be computationally efficient. Please refer to Ayoub et al. (2020) for a detailed discussion.

Appendix D. Proof of Upper Bound Results in Section 5

Let $\mathbb{P}$ be the distribution over $(\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$ induced by the interconnection of UCRL-VTR$^+$ (treated as a nonstationary, history dependent policy) and the episodic MDP M. Further, let $\mathbb{E}$ be the corresponding expectation operator. Note that the only source of randomness are the stochastic transitions in the MDP, hence, all random variables can be defined over the sample space $\Omega = (\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$. Thus, we work with the probability space given by the triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is the product σ-algebra generated by the discrete σ-algebras underlying $\mathcal{S}$ and $\mathcal{A}$, respectively.

For $1 \le k \le K$, $1 \le h \le H$, let $\mathcal{F}_{k,h}$ be the σ-algebra generated by the random variables representing the state-action pairs up to and including those that appear stage h of episode k. That is, $\mathcal{F}_{k,h}$ is generated by

$$
\begin{aligned}
&s_1^1, a_1^1, \dots, s_h^1, a_h^1, \dots, s_H^1, a_H^1\,,\\
&s_1^2, a_1^2, \dots, s_h^2, a_h^2, \dots, s_H^2, a_H^2\,,\\
&\qquad\qquad\vdots\\
&s_1^k, a_1^k, \dots, s_h^k, a_h^k\,.
\end{aligned}
$$

Note that, by construction,

$$
\bar{\mathbb{V}}_{k,h} V_{k,h+1}(s_h^k, a_h^k), E_{k,h}, \bar{\sigma}_{k,h}, \widehat{\boldsymbol{\Sigma}}_{k+1,h}, \widetilde{\boldsymbol{\Sigma}}_{k+1,h},
$$

are $\mathcal{F}_{k,h}$-measurable, $\widehat{\boldsymbol{b}}_{k+1,h}, \widetilde{\boldsymbol{b}}_{k+1,h}, \widehat{\boldsymbol{\theta}}_{k+1,h}, \widetilde{\boldsymbol{\theta}}_{k+1,h}$ are $\mathcal{F}_{k,h+1}$-measurable, and $Q_{k,h}, V_{k,h}, \pi_h^k, \phi_{V_{k,h+1}}$ are $\mathcal{F}_{k-1,H}$ measurable. Note also that $Q_{k,h}, V_{k,h}, \pi_h^k, \phi_{V_{k,h+1}}$ are *not* $\mathcal{F}_{k-1,h}$ measurable: The get their values only after episode $k-1$ is *over*, due to their "backwards" construction.

D.1. Proof of Lemma 5

The main idea of the proof is to use a (crude) two-step, "peeling" device. Let $\check{\mathcal{C}}_{k,h}, \widetilde{\mathcal{C}}_{k,h}$ denote the following confidence sets:

$$
\begin{aligned}
\check{\mathcal{C}}_{k,h} &= \Big\{\boldsymbol{\theta} : \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{k,h})\Big\|_2 \le \check{\beta}_k\Big\},\\
\widetilde{\mathcal{C}}_{k,h} &= \Big\{\boldsymbol{\theta} : \Big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{1/2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_{k,h})\Big\|_2 \le \widetilde{\beta}_k\Big\}.
\end{aligned}
$$

Note that $\widehat{\mathcal{C}}_{k,h} \subset \check{\mathcal{C}}_{k,h}$: The "leading term" in the definition of $\check{\beta}_k$ is larger than that in $\widehat{\beta}_k$ by a factor of $\sqrt{d}$. The idea of our proof is to show that $\boldsymbol{\theta}_h^*$ is included in $\check{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h}$ with high probability (for this, a standard self-normalized tail inequality suffices) and then use that when this holds, the weights used in constructing $\widehat{\boldsymbol{\theta}}_{k,h}$ are sufficiently precise to "balance" the noise term, which allows to reduce $\check{\beta}_k$ by the extra $\sqrt{d}$ factor without significantly increasing the probability of the bad event when $\boldsymbol{\theta}_h^* \notin \widehat{\mathcal{C}}_{k,h}$.

We start with the following lemma.

Lemma 15 *Let $V_{k,h+1}, \widehat{\boldsymbol{\theta}}_{k,h}, \widehat{\boldsymbol{\Sigma}}_{k,h}, \widetilde{\boldsymbol{\theta}}_{k,h}, \widetilde{\boldsymbol{\Sigma}}_{k,h}$ be defined in Algorithm 2, then we have*

$$
\begin{aligned}
&\big|\mathbb{V}_h V_{k,h+1}(s_h^k, a_h^k) - \bar{\mathbb{V}}_{k,h} V_{k,h+1}(s_h^k, a_h^k)\big|\\
&\quad\le \min\Big\{H^2, \Big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\Big\|_2\Big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{1/2}\big(\widetilde{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^*\big)\Big\|_2\Big\}\\
&\qquad + \min\Big\{H^2, 2H\Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\|_2\Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}\big(\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^*\big)\Big\|_2\Big\}.
\end{aligned}
$$

Proof We have

$$
\big|[\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)\big|
$$

$$
\begin{aligned}
&= \Big| \big[\big\langle \boldsymbol{\phi}_{V^2_{k,h+1}}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h}\big\rangle\big]_{[0,H^2]} - \big\langle \boldsymbol{\phi}_{V^2_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\big\rangle \\
&\qquad + \big(\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\big\rangle\big)^2 - \big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\big\rangle\big]^2_{[0,H]}\Big| \\
&\le \underbrace{\Big| \big[\big\langle \boldsymbol{\phi}_{V^2_{k,h+1}}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h}\big\rangle\big]_{[0,H^2]} - \big\langle \boldsymbol{\phi}_{V^2_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\big\rangle\Big|}_{I_1} \\
&\qquad + \underbrace{\Big| \big(\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\big\rangle\big)^2 - \big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\big\rangle\big]^2_{[0,H]}\Big|}_{I_2},
\end{aligned}
$$

where the inequality holds due to the triangle inequality. We bound I_1 first. We have $I_1 \le H^2$ since both terms in I_1 belong to the interval $[0, H^2]$. Furthermore,

$$
\begin{aligned}
I_1 &\le \Big|\big\langle \boldsymbol{\phi}_{V^2_{k,h+1}}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h}\big\rangle - \big\langle \boldsymbol{\phi}_{V^2_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\big\rangle\Big| \\
&= \Big|\big\langle \boldsymbol{\phi}_{V^2_{k,h+1}}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^*\big\rangle\Big| \\
&\le \Big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V^2_{k,h+1}}(s_h^k, a_h^k)\Big\|_2 \Big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{1/2}\big(\widetilde{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^*\big)\Big\|_2,
\end{aligned}
$$

where the first inequality holds since $\langle \boldsymbol{\phi}_{V^2_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\rangle \in [0, H^2]$ and the second inequality holds due to the Cauchy-Schwarz inequality. Thus, we have

$$
I_1 \le \min\Big\{H^2, \Big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V^2_{k,h+1}}(s_h^k, a_h^k)\Big\|_2 \Big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{1/2}\big(\widetilde{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^*\big)\Big\|_2\Big\}. \tag{50}
$$

For the term I_2, since both terms in I_2 belong to the interval $[0, H^2]$, we have $I_2 \le H^2$. Meanwhile,

$$
\begin{aligned}
I_2 &= \Big|\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\big\rangle + \big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\big\rangle\big]_{[0,H]}\Big| \\
&\qquad \cdot \Big|\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\big\rangle - \big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\big\rangle\big]_{[0,H]}\Big| \\
&\le 2H\Big|\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\big\rangle - \big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\big\rangle\Big| \\
&= 2H\Big|\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^* - \widehat{\boldsymbol{\theta}}_{k,h}\big\rangle\Big| \\
&\le 2H\Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\|_2 \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}\big(\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^*\big)\Big\|_2,
\end{aligned} \tag{51}
$$

where the first inequality holds since both terms in this line are less than H and the fact $\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\big\rangle \in [0, H]$, the second inequality holds due to the Cauchy-Schwarz inequality. Thus, we have

$$
I_2 \le \min\Big\{H^2, 2H\Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\|_2 \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}\big(\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^*\big)\Big\|_2\Big\}. \tag{52}
$$

Combining (50) and (52) gives the desired result. ■

Proof [Proof of Lemma 5] Fix $h \in [H]$. We first show that with probability at least $1 - \delta/H$, for all k, $\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{k,h}$. To show this, we apply Theorem 2. Let $\mathbf{x}_i = \bar{\sigma}_{i,h}^{-1}\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)$ and $\eta_i = \bar{\sigma}_{i,h}^{-1}V_{i,h+1}(s_{h+1}^i) - \bar{\sigma}_{i,h}^{-1}\langle\boldsymbol{\phi}_{V_{i,h+1}}(s_{i,h}, a_{i,h}), \boldsymbol{\theta}_h^*\rangle$, $\mathcal{G}_i = \mathcal{F}_{i,h}$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}_h^*$, $y_i = \langle\boldsymbol{\mu}^*, \mathbf{x}_i\rangle + \eta_i$,

$\mathbf{Z}_i = \lambda \mathbf{I} + \sum_{i'=1}^{i} \mathbf{x}_{i'} \mathbf{x}_{i'}^\top$, $\mathbf{b}_i = \sum_{i'=1}^{i} \mathbf{x}_{i'} y_{i'}$ and $\boldsymbol{\mu}_i = \mathbf{Z}_i^{-1}\mathbf{b}_i$. Then it can be verified that $y_i = \bar{\sigma}_{i,h}^{-1} V_{i,h+1}(s_{h+1}^i)$ and $\boldsymbol{\mu}_i = \widehat{\boldsymbol{\theta}}_{i+1,h}$. Moreover, almost surely,

$$\|\mathbf{x}_i\|_2 \le \bar{\sigma}_{i,h}^{-1} H \le \sqrt{d}, \;\; |\eta_i| \le \bar{\sigma}_{i,h}^{-1} H \le \sqrt{d}, \;\; \mathbb{E}[\eta_i | \mathcal{G}_i] = 0, \;\; \mathbb{E}[\eta_i^2 | \mathcal{G}_i] \le d\,,$$

where we used that $V_{i,h+1}$ takes values in $[0, H]$ and that $\|\boldsymbol{\phi}_{V_{i,h+1}}(s,a)\|_2 \le H$ by (1). Since we also have that $\boldsymbol{x}_i$ is $\mathcal{G}_i$ measurable and η_i is $\mathcal{G}_{i+1}$ measurable, by Theorem 2, we obtain that with probability at least $1 - \delta/H$, for all $k \le K$,

$$\big\|\boldsymbol{\theta}_h^* - \widehat{\boldsymbol{\theta}}_{k,h}\big\|_{\widehat{\boldsymbol{\Sigma}}_{k,h}} \le 8d\sqrt{\log(1 + k/\lambda)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B = \check{\beta}_k, \quad (53)$$

implying that with probability $1 - \delta/H$, for any $k \le K$, $\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{k,h}$.

An argument, which is analogous to the one just used (except that now the range of the "noise" matches the range of "squared values" and is thus bounded by H^2, rather than being bounded by $\sqrt{d}$) gives that with probability at least $1 - \delta/H$, for any $k \le K$ we have

$$\big\|\boldsymbol{\theta}_h^* - \widetilde{\boldsymbol{\theta}}_{k,h}\big\|_{\widetilde{\boldsymbol{\Sigma}}_{k,h}} \le 8\sqrt{dH^4\log(1 + kH^4/(d\lambda))\log(4k^2H/\delta)} + 4H^2\log(4k^2H/\delta) + \sqrt{\lambda}B = \widetilde{\beta}_k, \quad (54)$$

which implies that with the said probability, $\boldsymbol{\theta}_h^* \in \widetilde{\mathcal{C}}_{k,h}$.

We now show that $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$ with high probability. We again apply Theorem 2. Let $\mathbf{x}_i = \bar{\sigma}_{i,h}^{-1}\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)$ and

$$\eta_i = \bar{\sigma}_{i,h}^{-1}\,\mathbb{1}\{\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h}\}\big[V_{i,h+1}(s_{h+1}^i) - \langle \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i), \boldsymbol{\theta}_h^*\rangle\big],$$

$\mathcal{G}_i = \mathcal{F}_{i,h}$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}_h^*$. Clearly $\mathbb{E}[\eta_i|\mathcal{G}_i] = 0$, $|\eta_i| \le \bar{\sigma}_{i,h}^{-1}H \le \sqrt{d}$ since $|V_{i,h+1}(\cdot)| \le H$ and $\bar{\sigma}_{i,h} \ge H/\sqrt{d}$, $\|\mathbf{x}_i\|_2 \le \bar{\sigma}_{i,h}^{-1}H \le \sqrt{d}$. Furthermore, owning to that $\mathbb{1}\{\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h}\}$ is $\mathcal{G}_i$-measurable, it holds that

$$\begin{aligned}
\mathbb{E}[\eta_i^2|\mathcal{G}_i] &= \bar{\sigma}_{i,h}^{-2}\,\mathbb{1}\{\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h}\}[\mathbb{V}_h V_{i,h+1}](s_h^i, a_h^i)\\
&\le \bar{\sigma}_{i,h}^{-2}\,\mathbb{1}\{\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h}\}\bigg[[\bar{\mathbb{V}}_{i,h}V_{i,h+1}](s_h^i, a_h^i)\\
&\qquad + \min\Big\{H^2, \Big\|\widetilde{\boldsymbol{\Sigma}}_{i,h}^{-1/2}\boldsymbol{\phi}_{V_{i,h+1}^2}(s_h^i, a_h^i)\Big\|_2\Big\|\widetilde{\boldsymbol{\Sigma}}_{i,h}^{1/2}\big(\widetilde{\boldsymbol{\theta}}_{i,h} - \boldsymbol{\theta}_h^*\big)\Big\|_2\Big\}\\
&\qquad + \min\Big\{H^2, 2H\Big\|\widehat{\boldsymbol{\Sigma}}_{i,h}^{-1/2}\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\Big\|_2\Big\|\widehat{\boldsymbol{\Sigma}}_{i,h}^{1/2}\big(\widehat{\boldsymbol{\theta}}_{i,h} - \boldsymbol{\theta}_h^*\big)\Big\|_2\Big\}\bigg]\\
&\le \bar{\sigma}_{i,h}^{-2}\bigg[[\bar{\mathbb{V}}_{i,h}V_{i,h+1}](s_h^i, a_h^i) + \min\Big\{H^2, \widetilde{\beta}_i\Big\|\widetilde{\boldsymbol{\Sigma}}_{i,h}^{-1/2}\boldsymbol{\phi}_{V_{i,h+1}^2}(s_h^i, a_h^i)\Big\|_2\Big\}\\
&\qquad + \min\Big\{H^2, 2H\check{\beta}_i\Big\|\widehat{\boldsymbol{\Sigma}}_{i,h}^{-1/2}\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\Big\|_2\Big\}\bigg]\\
&= 1,
\end{aligned}$$

where the first inequality holds due to Lemma 15, the second inequality holds due to the indicator function, the last equality holds due to the definition of $\bar{\sigma}_{i,h}$. Now, let $y_i = \langle \boldsymbol{\mu}^*, \mathbf{x}_i\rangle + \eta_i$, $\mathbf{Z}_i =$

$\lambda\mathbf{I} + \sum_{i'=1}^{i} \mathbf{x}_{i'}\mathbf{x}_{i'}^\top$, $\mathbf{b}_i = \sum_{i'=1}^{i} \mathbf{x}_{i'}y_{i'}$ and $\boldsymbol{\mu}_i = \mathbf{Z}_i^{-1}\mathbf{b}_i$. Then, by Theorem 2, with probability at least $1-\delta/H, \forall k \le K$,

$$\|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|_{\mathbf{Z}_i} \le 8\sqrt{d\log(1+k/\lambda)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B = \widehat{\beta}_k, \tag{55}$$

where the equality uses the definition of $\widehat{\beta}_k$. Let $\mathcal{E}'$ be the event when $\boldsymbol{\theta}_h^* \in \cap_{k\le K}\check{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h}$ and (55) hold. By the union bound, $\mathbb{P}(\mathcal{E}') \ge 1-3\delta/H$.

We now show that $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$ holds on $\mathcal{E}'$. For this note that on $\mathcal{E}'$, for all $k \le K$, $\boldsymbol{\mu}_k = \widehat{\boldsymbol{\theta}}_{k+1,h}$ for any $k \le K$. Indeed, on this event, for any $i \le K$,

$$\begin{aligned} y_i &= \bar{\sigma}_{i,h}^{-1}\big(\langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\rangle + \mathbb{1}\{\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h}\}\big[V_{i,h+1}(s_{h+1}^i) - \langle \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i), \boldsymbol{\theta}^*\rangle\big]\big) \\ &= \bar{\sigma}_{i,h}^{-1}V_{i,h+1}(s_{h+1}^i), \end{aligned}$$

which does imply the claim. Therefore, by the definition of $\widehat{\mathcal{C}}_{k,h}$ and since on $\mathcal{E}'$ (55) holds, we get that on $\mathcal{E}'$, the relation $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$ also holds. Finally, taking union bound over h and substituting (53) and (54) into Lemma 15 shows that with probability at least $1-3\delta$,

$$\boldsymbol{\theta}_h^* \in \cap_{k,h}\widehat{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h} \tag{56}$$

To finish our proof, it is thus sufficient to show that on the event when (56) holds, it also holds that

$$\big|[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_hV_{k,h+1}](s_h^k, a_h^k)\big| \le E_{k,h}.$$

However, this is immediate from Lemma 15 and the definition of $E_{k,h}$. ■

D.2. Proof of Theorem 6

In this subsection we prove Theorem 6. The proof is broken down into a number of lemmas. However, first we need the Azuma-Hoeffding inequality:

Lemma 16 (Azuma-Hoeffding inequality, Azuma 1967) *Let $M > 0$ be a constant. Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}_i$ ($\mathbb{E}[x_i|\mathcal{G}_i] = 0$ a.s. and x_i is $\mathcal{G}_{i+1}$-measurable) such that for all $i \in [n]$, $|x_i| \le M$ holds almost surely. Then, for any $0 < \delta < 1$, with probability at least $1-\delta$, we have*

$$\sum_{i=1}^n x_i \le M\sqrt{2n\log(1/\delta)}.$$

For the remainder of this subsection, let $\mathcal{E}$ denote the event when the conclusion of Lemma 5 holds. Then Lemma 5 suggests $\mathbb{P}(\mathcal{E}) \ge 1-3\delta$. We introduce another two events $\mathcal{E}_1$ and $\mathcal{E}_2$:

$$\begin{aligned} \mathcal{E}_1 = \bigg\{&\forall h' \in [H], \sum_{k=1}^K\sum_{h=h'}^H \Big[[\mathbb{P}_h(V_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)\Big] \\ &\le 4H\sqrt{2T\log(H/\delta)}\bigg\}, \end{aligned}$$

$$\mathcal{E}_2 = \bigg\{ \sum_{k=1}^K \sum_{h=1}^H [\mathbb{V}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k) \le 3(HT + H^3 \log(1/\delta)) \bigg\}.$$

Then we have $\mathbb{P}(\mathcal{E}_1) \ge 1-\delta$ and $\mathbb{P}(\mathcal{E}_2) \ge 1-\delta$. The first one holds since $[\mathbb{P}_h(V_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)$ forms a martingale difference sequence and $|[\mathbb{P}_h(V_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)| \le 4H$. Applying the Azuma-Hoeffding inequality (Lemma 16), we find that with probability at least $1-\delta$, simultaneously for all $h' \in [H]$, we have

$$\sum_{k=1}^K \sum_{h=h'}^H \Big[[\mathbb{P}_h(V_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)\Big] \le 4H\sqrt{2T\log(H/\delta)}, \quad (57)$$

which implies $\mathbb{P}(\mathcal{E}_1) \ge 1-\delta$. That $\mathbb{P}(\mathcal{E}_2) \ge 1-\delta$ holds is due to the following lemma:

Lemma 17 (Total variance lemma, Lemma C.5, Jin et al. 2018) *With probability at least $1-\delta$, we have*

$$\sum_{k=1}^K \sum_{h=1}^H [\mathbb{V}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k) \le 3(HT + H^3 \log(1/\delta)).$$

Remark 18 *Maillard et al. (2014); Zanette and Brunskill (2019) considered a setting where the variance of the optimal value function V^* is bounded by some quantity $\mathbb{Q}^*$. Under this setting, to bound the summation of variances of value functions, we can also obtain a tighter bound based on $\mathbb{Q}^*$ instead of H and T, as shown in Lemma 17.*

We now prove the following three lemmas based on $\mathcal{E}, \mathcal{E}_1, \mathcal{E}_2$.

Lemma 19 *Let $Q_{k,h}, V_{k,h}$ be defined in Algorithm 2. Then, on the event $\mathcal{E}$, for any s, a, k, h we have that $Q_h^*(s,a) \le Q_{k,h}(s,a)$, $V_h^*(s) \le V_{k,h}(s)$.*

Proof Since $\mathcal{E}$ holds, we have for any $k \in [K]$ and $h \in [H]$, $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$. We prove the statement by induction. The statement holds for $h = H+1$ since $Q_{k,H+1}(\cdot,\cdot) = 0 = Q_{H+1}^*(\cdot,\cdot)$. Assume the statement holds for $h+1$. That is, $Q_{k,h+1}(\cdot,\cdot) \ge Q_{h+1}^*(\cdot,\cdot)$, $V_{k,h+1}(\cdot) \ge V_{h+1}^*(\cdot)$. Given s, a, if $Q_{k,h}(s,a) \ge H$, then $Q_{k,h}(s,a) \ge H \ge Q_h^*(s,a)$. Otherwise, we have

$$\begin{aligned}
&Q_{k,h}(s,a) - Q_h^*(s,a)\\
&= \langle \boldsymbol{\phi}_{V_{k,h+1}}(s,a), \widehat{\boldsymbol{\theta}}_{k,h}\rangle + \widehat{\beta}_k \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s,a)\Big\|_2 - \langle \boldsymbol{\phi}_{V_{k,h+1}}(s,a), \boldsymbol{\theta}_h^*\rangle\\
&\qquad + \mathbb{P}_h V_{k,h+1}(s,a) - \mathbb{P}_h V_{h+1}^*(s,a)\\
&\ge \widehat{\beta}_k \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s,a)\Big\|_2 - \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}(\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^*)\Big\|_2 \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s,a)\Big\|_2\\
&\qquad + \mathbb{P}_h V_{k,h+1}(s,a) - \mathbb{P}_h V_{h+1}^*(s,a)\\
&\ge \mathbb{P}_h V_{k,h+1}(s,a) - \mathbb{P}_h V_{h+1}^*(s,a)\\
&\ge 0,
\end{aligned}$$

where the first inequality holds due to Cauchy-Schwarz, the second inequality holds by the assumption that $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$, the third inequality holds by the induction assumption and because $\mathbb{P}_h$ is a monotone operator with respect to the partial ordering of functions. Therefore, for all s, a, we have $Q_{k,h}(s,a) \geq Q_h^*(s,a)$, which implies $V_{k,h}(s) \geq V_h^*(s)$, finishing the inductive step and thus the proof. ■

Lemma 20 *Let $V_{k,h}, \bar{\sigma}_{k,h}$ be defined in Algorithm 2. Then, on the event $\mathcal{E} \cap \mathcal{E}_1$, we have*

$$\sum_{k=1}^K \Big[V_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k)\Big] \leq 2\widehat{\beta}_K \sqrt{\sum_{k=1}^K \sum_{h=1}^H \bar{\sigma}_{k,h}^2} \sqrt{2Hd\log(1+K/\lambda)} + 4H\sqrt{2T\log(H/\delta)},$$

$$\sum_{k=1}^K \sum_{h=1}^H \mathbb{P}_h[V_{k,h+1} - V_{h+1}^{\pi^k}](s_h^k, a_h^k) \leq 2\widehat{\beta}_K \sqrt{\sum_{k=1}^K \sum_{h=1}^H \bar{\sigma}_{k,h}^2} \sqrt{2dH^3\log(1+K/\lambda)} + 4H^2\sqrt{2T\log(H/\delta)}.$$

Proof Assume that $\mathcal{E} \cap \mathcal{E}_1$ holds. We have

$$\begin{aligned} V_{k,h}(s_h^k) - V_h^{\pi^k}(s_h^k) &\leq \langle \widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\rangle - [\mathbb{P}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k) + \widehat{\beta}_k \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2} \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\|_2 \\ &\leq \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}(\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^*)\Big\|_2 \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2} \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\|_2 \\ &\quad + [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{P}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k) + \widehat{\beta}_k \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2} \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\|_2 \\ &\leq [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{P}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k) + 2\widehat{\beta}_k \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2} \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\|_2, \end{aligned} \tag{58}$$

where the first inequality holds due to the definition of $V_{k,h}$ and the Bellman equation for $V_h^{\pi^k}$, the second inequality holds due to Cauchy-Schwarz inequality and because we are in a linear MDP, the third inequality holds by the fact that on $\mathcal{E}$, $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$. Meanwhile, since $V_{k,h}(s_h^k) - V_h^{\pi^k}(s_h^k) \leq H$, we also have

$$\begin{aligned} &V_{k,h}(s_h^k) - V_h^{\pi^k}(s_h^k) \\ &\leq \min\Big\{H, 2\widehat{\beta}_k \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2} \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\|_2 + [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{P}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k)\Big\} \\ &\leq \min\Big\{H, 2\widehat{\beta}_k \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2} \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\|_2\Big\} + [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{P}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k) \\ &\leq 2\widehat{\beta}_k \bar{\sigma}_{k,h} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2} \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2\Big\} + [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{P}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k), \end{aligned} \tag{59}$$

where the second inequality holds since the optimal value function dominates the value function of any policy, and thus on $\mathcal{E}$, by Lemma 19, $V_{k,h+1}(\cdot) \geq V_{h+1}^{\pi^k}(\cdot)$, the third inequality holds since $2\widehat{\beta}_k \bar{\sigma}_{k,h} \geq \sqrt{d} \cdot H/\sqrt{d} \geq H$. By (59) we have

$$V_{k,h}(s_h^k) - V_h^{\pi^k}(s_h^k) - [V_{k,h+1}(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)] \tag{60}$$

$$\begin{aligned} &\leq 2\widehat{\beta}_k \bar{\sigma}_{k,h} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2} \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2\Big\} \\ &\quad + \mathbb{P}_h[V_{k,h+1} - V_{h+1}^{\pi^k}](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k). \end{aligned} \tag{61}$$

Summing up these inequalities for $k \in [K]$ and $h = h', \dots, H$,

$$
\begin{aligned}
&\sum_{k=1}^{K} \Big[V_{k,h'}(s_{k,h'}) - V_{h'}^{\pi^k}(s_{k,h'})\Big] \\
&\leq 2\sum_{k=1}^{K}\sum_{h=h'}^{H} \widehat{\beta}_k \bar{\sigma}_{k,h} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2\Big\} \\
&\qquad + \sum_{k=1}^{K}\sum_{h=h'}^{H} \Big[[\mathbb{P}_h(V_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)\Big] \\
&\leq \underbrace{2\sum_{k=1}^{K}\sum_{h=1}^{H} \widehat{\beta}_k \bar{\sigma}_{k,h} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2\Big\}}_{I_1} + 4H\sqrt{2T\log(H/\delta)}, \qquad (62)
\end{aligned}
$$

where the first inequality holds by a telescoping argument and since $V_{k,H+1}(\cdot) = V_{h+1}^{\pi^k}(\cdot) = 0$, the second inequality holds due to $\mathcal{E}_1$. To further bound I_1, we have

$$
\begin{aligned}
I_1 &\leq \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\beta}_k^2 \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2^2\Big\}} \\
&\leq \widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2^2\Big\}} \\
&\leq \widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)}, \qquad (63)
\end{aligned}
$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds since $\widehat{\beta}_k \leq \widehat{\beta}_K$, the third inequality holds due to Lemma 12 with the fact that $\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\|_2 \leq \|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\|_2 \cdot \sqrt{d}/H \leq \sqrt{d}$. Substituting (63) into (62) gives

$$
\sum_{k=1}^{K}\Big[V_{k,h'}(s_{k,h'}) - V_{h'}^{\pi^k}(s_{k,h'})\Big] \leq 2\widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)} + 4H\sqrt{2T\log(H/\delta)}\,. \qquad (64)
$$

Choosing $h' = 1$ here we get the first inequality that was to be proven. To get the second inequality, note that

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{P}_h[V_{k,h+1} - V_{h+1}^{\pi^k}](s_h^k, a_h^k) \\
&= \sum_{k=1}^{K}\sum_{h=2}^{H}[V_{k,h} - V_h^{\pi^k}](s_h^k)
\end{aligned}
$$

$$+\sum_{k=1}^{K}\sum_{h=1}^{H}\Big[[\mathbb{P}_h(V_{k,h+1}-V_{h+1}^{\pi^k})](s_h^k,a_h^k)-[V_{k,h+1}-V_{h+1}^{\pi^k}](s_{h+1}^k)\Big]$$
$$\le 2\widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2dH^3\log(1+KH/(d\lambda))}+4H^2\sqrt{2T\log(H/\delta)},$$

where to get the last inequality we sum up (64) for $h'=2,\dots,H$, and use the inequality that defines $\mathcal{E}_1$, which is followed by loosening the resulting bound. ■

The next lemma is concerned with bounding $\sum_{k=1}^K\sum_{h=1}^H\bar{\sigma}_{k,h}^2$ on $\mathcal{E}\cap\mathcal{E}_2$:

Lemma 21 *Let $V_{k,h},\bar{\sigma}_{k,h}$ be defined in Algorithm 2. Then, on the event $\mathcal{E}\cap\mathcal{E}_2$, we have*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2\le H^2T/d+3(HT+H^3\log(1/\delta))+2H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{P}_h[V_{k,h+1}-V_{h+1}^{\pi^k}](s_h^k,a_h^k)$$
$$+2\widetilde{\beta}_K\sqrt{T}\sqrt{2dH\log(1+KH^4/(d\lambda))}+7\check{\beta}_KH^2\sqrt{T}\sqrt{2dH\log(1+K/\lambda)}.$$

Proof Assume that $\mathcal{E}\cap\mathcal{E}_2$ holds. Since we are on $\mathcal{E}$, by Lemma 19, for all k,h, $V_{k,h}(\cdot)\ge V_h^*(\cdot)\ge V_h^{\pi^k}(\cdot)$. Now, we calculate

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2\\
&\quad\le\sum_{k=1}^{K}\sum_{h=1}^{H}\Big[H^2/d+[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)+E_{k,h}\Big]\\
&\quad=H^2T/d+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\Big[[\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{h+1}^{\pi^k}](s_h^k,a_h^k)\Big]}_{I_1}+2\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}E_{k,h}}_{I_2}\\
&\qquad+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_hV_{h+1}^{\pi^k}](s_h^k,a_h^k)}_{I_3}+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\Big[[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)-E_{k,h}\Big]}_{I_4},\end{aligned}\tag{65}$$

where the first inequality holds due to the definition of $\bar{\sigma}_{k,h}$. To bound I_1, we have

$$\begin{aligned}I_1&\le\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}_hV_{k,h+1}^2](s_h^k,a_h^k)-[\mathbb{P}_h(V_{h+1}^{\pi^k})^2](s_h^k,a_h^k)\\
&\le 2H\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}_h(V_{k,h+1}-V_{h+1}^{\pi^k})](s_h^k,a_h^k),\end{aligned}$$

where the first inequality holds since $V_{h+1}^{\pi^k}(\cdot) \leq V_{h+1}^*(\cdot) \leq V_{k,h+1}(\cdot)$, the second inequality holds since $V_{h+1}^{\pi^k}(\cdot), V_{k,h+1}(\cdot) \leq H$. To bound I_2, we have

$$
\begin{aligned}
I_2 &\leq 2\sum_{k=1}^{K}\sum_{h=1}^{H}\widetilde{\beta}_k \min\Big\{1, \Big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\Big\|_2\Big\} \\
&\quad + 4H\sum_{k=1}^{K}\sum_{h=1}^{H}\check{\beta}_k\bar{\sigma}_{k,h}\min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2\Big\} \\
&\leq 2\widetilde{\beta}_K\sqrt{T}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\Big\{1, \Big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\Big\|_2^2\Big\}} \\
&\quad + 7\check{\beta}_K H^2\sqrt{T}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2^2\Big\}} \\
&\leq 2\widetilde{\beta}_K\sqrt{T}\sqrt{2dH\log(1+KH^4/(d\lambda))} + 7\check{\beta}_K H^2\sqrt{T}\sqrt{2dH\log(1+K/\lambda)},
\end{aligned}
$$

where the first inequality holds since $\widetilde{\beta}_k \geq H^2$ and $\check{\beta}_k\bar{\sigma}_{k,h} \geq \sqrt{d}\cdot H/\sqrt{d} = H$, the second inequality holds due to Cauchy-Schwarz inequality, $\widetilde{\beta}_k \leq \widetilde{\beta}_K$, $\check{\beta}_k \leq \check{\beta}_K$, and the following bound on $\bar{\sigma}_{k,h}$ due to the definitions of $\bar{\sigma}_{k,h}$, $[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k)$ and $E_{k,h}$:

$$
\bar{\sigma}_{k,h}^2 = \max\big\{H^2/d, [\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) + E_{k,h}\big\} \leq \max\big\{H^2/d, H^2 + 2H^2\big\} = 3H^2\,.
$$

Finally, the third inequality holds due to Lemma 12 together with the facts that $\big\|\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\big\|_2 \leq H^2$ and $\big\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\big\|_2 \leq \big\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\big\|_2\cdot\sqrt{d}/H \leq \sqrt{d}$. To bound I_3, since $\mathcal{E}_2$ holds, we have

$$
I_3 \leq 3(HT + H^3\log(1/\delta)).
$$

Finally, due to Lemma 5, we have $I_4 \leq 0$. Substituting I_1, I_2, I_3, I_4 into (65) ends our proof. ■

With all above lemmas, we are ready to prove Theorem 6.

Proof [Proof of Theorem 6] By construction, taking a union bound, we have with probability $1-5\delta$ that $\mathcal{E}\cap\mathcal{E}_1\cap\mathcal{E}_2$ holds. In the remainder of the proof, assume that we are on this event. Thus, we can also use the conclusions of Lemmas 19, 20 and 21. We bound the regret as

$$
\begin{aligned}
\mathrm{Regret}(M_{\boldsymbol{\theta}^*}, K) &\leq \sum_{k=1}^{K}\Big[V_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k)\Big] \\
&\leq 2\widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+KH/(d\lambda))} + 4H\sqrt{2T\log(H/\delta)} \\
&= \widetilde{O}\bigg(\sqrt{dH}\sqrt{d}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2} + H\sqrt{T}\bigg), \qquad (66)
\end{aligned}
$$

where the first inequality holds due to Lemma 19, the second inequality holds due to Lemma 20, the equality holds since when $\lambda = 1/B^2$,

$$\widehat{\beta}_K = 8\sqrt{d\log(1+K/\lambda)\log(4K^2H/\delta)} + 4\sqrt{d}\log(4K^2H/\delta) + \sqrt{\lambda}B = \widetilde{\Theta}(\sqrt{d}).$$

It remains to bound $\sum_{k=1}^K\sum_{h=1}^H \bar{\sigma}_{k,h}^2$. For this we have

$$\begin{aligned}\sum_{k=1}^K\sum_{h=1}^H \bar{\sigma}_{k,h}^2 &\leq H^2T/d + 3(HT + H^3\log(1/\delta)) + 2H\sum_{k=1}^K\sum_{h=1}^H \mathbb{P}_h[V_{k,h+1} - V_{h+1}^{\pi^k}](s_h^k, a_h^k)\\ &\quad + 2\widetilde{\beta}_K\sqrt{T}\sqrt{2dH\log(1+KH^4/(d\lambda))} + 7\check{\beta}_K H^2\sqrt{T}\sqrt{2dH\log(1+K/\lambda)}\\ &\leq H^2T/d + 3(HT + H^3\log(1/\delta)) + 2H\\ &\quad\cdot\Bigg(2\widehat{\beta}_K\sqrt{\sum_{k=1}^K\sum_{h=1}^H \bar{\sigma}_{k,h}^2}\sqrt{2dH^3\log(1+K/\lambda)} + 4H^2\sqrt{2T\log(H/\delta)}\Bigg)\\ &\quad + 2\widetilde{\beta}_K\sqrt{T}\sqrt{2dH\log(1+KH^4/(d\lambda))} + 7\check{\beta}_K H^2\sqrt{T}\sqrt{2dH\log(1+K/\lambda)}\\ &= \widetilde{O}\Bigg(\sqrt{\sum_{k=1}^K\sum_{h=1}^H \bar{\sigma}_{k,h}^2}\sqrt{d^2H^5} + H^2T/d + TH + \sqrt{T}d^{1.5}H^{2.5} + \sqrt{T}H^3\Bigg).\end{aligned} \tag{67}$$

where the first inequality holds due to Lemma 21, the second inequality holds due to Lemma 20, the last equality holds due to the fact that $\widehat{\beta}_K = \widetilde{O}(\sqrt{d})$, $\lambda = 1/B^2$,

$$\begin{aligned}\check{\beta}_K &= 8d\sqrt{\log(1+K/\lambda)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B = \widetilde{\Theta}(d),\\ \widetilde{\beta}_K &= 8\sqrt{dH^4\log(1+KH^4/(d\lambda))\log(4k^2H/\delta)} + 4H^2\log(4k^2H/\delta) + \sqrt{\lambda}B = \widetilde{\Theta}(\sqrt{d}H^2).\end{aligned}$$

Therefore, by the fact that $x \leq a\sqrt{x} + b$ implies $x \leq c(a^2+b)$ with some $c > 0$, (67) yields that

$$\begin{aligned}\sum_{k=1}^K\sum_{h=1}^H \bar{\sigma}_{k,h}^2 &\leq \widetilde{O}\big(d^2H^5 + H^2T/d + TH + \sqrt{T}d^{1.5}H^{2.5} + \sqrt{T}H^3\big)\\ &= \widetilde{O}\big(d^2H^5 + d^4H^3 + TH + H^2T/d\big),\end{aligned} \tag{68}$$

where the equality holds since $\sqrt{T}d^{1.5}H^{2.5} \leq (TH^2/d + d^4H^3)/2$ and $\sqrt{T}H^3 \leq (H^2T/d + H^4d)/2$. Substituting (68) into (66), we have

$$\text{Regret}\big(M_{\mathbf{\Theta}^*}, K\big) = \widetilde{O}\Big(\sqrt{d^2H^2 + dH^3}\sqrt{T} + d^2H^3 + d^3H^2\Big),$$

finishing the proof. ■

Remark 22 *To derive our upper bound of regret, we actually only need a weaker assumption on reward functions r_h such that for any policy π, we have $0 \leq \sum_{h=1}^H r_h(s_h, a_h) \leq H$, where $a_h = \pi_h(s_h)$, $s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h)$. Therefore, under the assumption $0 \leq \sum_{h=1}^H r_h(s_h, a_h) \leq 1$ studied in Dann and Brunskill (2015); Jiang and Agarwal (2018); Wang et al. (2020a); Zhang et al. (2021a), by simply rescaling all parameters in Algorithm 2 by a factor of $1/H$, UCRL-VTR$^+$ achieves the regret $\widetilde{O}(\sqrt{d^2 + dH}\sqrt{T} + d^2H^2 + d^3H)$. Zhang et al. (2021a) has shown that in the tabular,* homogeneous *case with this normalization the regret is $\widetilde{O}(\sqrt{|\mathcal{S}||\mathcal{A}|K} + |\mathcal{S}|^2|\mathcal{A}|)$, regardless of the value of H. It remains to be an interesting open question whether this can be also achieved in homogeneous linear mixture MDPs.*

Appendix E. Proof of Lower Bound Results in Section 5

E.1. Overview of the Lower Bound Construction

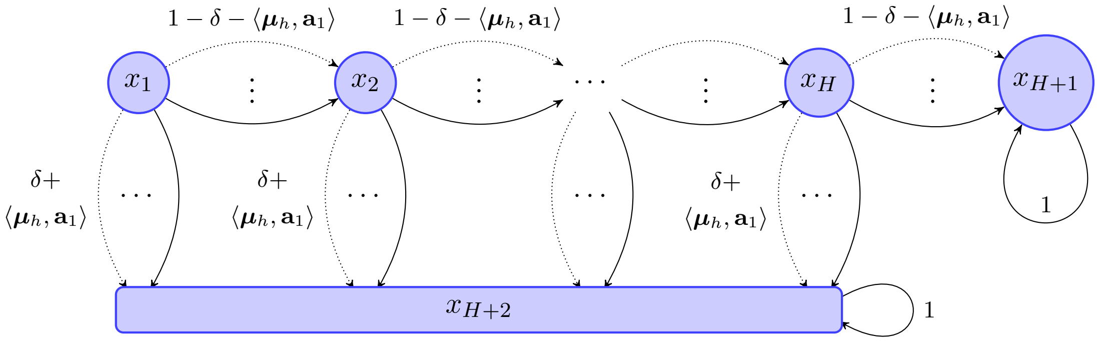
$1-\delta-\langle\boldsymbol{\mu}_h,\mathbf{a}_1\rangle$
$1-\delta-\langle\boldsymbol{\mu}_h,\mathbf{a}_1\rangle$
$1-\delta-\langle\boldsymbol{\mu}_h,\mathbf{a}_1\rangle$
x_1
x_2
$\cdots$
x_H
x_{H+1}
$\delta+$ $\langle\boldsymbol{\mu}_h,\mathbf{a}_1\rangle$
$\delta+$ $\langle\boldsymbol{\mu}_h,\mathbf{a}_1\rangle$
$\delta+$ $\langle\boldsymbol{\mu}_h,\mathbf{a}_1\rangle$
1
x_{H+2}
1

Figure 1: The transition kernel $\mathbb{P}_h$ of the class of hard-to-learn linear mixture MDPs. The kernel $\mathbb{P}_h$ is parameterized by $\boldsymbol{\mu}_h \in \{-\Delta,\Delta\}^{d-1}$ for some small Δ, $\delta = 1/H$ and the actions are from $\mathbf{a} \in \{+1,-1\}^{d-1}$. The learner knows this structure, but does not know $\boldsymbol{\mu} = (\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_H)$.

To prove the lower bound, we construct a hard instance $M(\mathcal{S},\mathcal{A},H,\{r_h\},\{\mathbb{P}_h\})$ based on the hard-to-learn MDPs introduced in Zhou et al. (2021). The transitions for stage h of the MDP are shown in Figure 1. The state space $\mathcal{S}$ consists of states $x_1,\ldots x_{H+2}$, where x_{H+1} and x_{H+2} are absorbing states. There are 2^{d-1} actions and $\mathcal{A} = \{-1,1\}^{d-1}$. Regardless of the stage $h \in [H]$, no transition incurs a reward except transitions originating at x_{H+2}, which, as a result, can be regarded as the goal state. Under $\mathbb{P}_h$, the transition structure is as follows: As noted before, x_{H+1} and x_{H+2} are absorbing regardless of the action taken. If the state is x_i with $i \le H$, under action $\boldsymbol{a} \in \{-1,1\}^{d-1}$, the next state is either x_{H+2} or x_{i+1}, with respective probabilities $\delta + \langle\boldsymbol{\mu}_h,\boldsymbol{a}\rangle$ and $1-(\delta+\langle\boldsymbol{\mu}_h,\boldsymbol{a}\rangle)$, where $\delta = 1/H$ and $\boldsymbol{\mu}_h \in \{-\Delta,\Delta\}^{d-1}$ with $\Delta = \sqrt{\delta/K}/(4\sqrt{2})$ so that the probabilities are well-defined.

This is an inhomogeneous, linear mixture MDP. In particular, $\mathbb{P}_h(s'|s,\mathbf{a}) = \langle\boldsymbol{\phi}(s'|s,\mathbf{a}),\boldsymbol{\theta}_h\rangle$, with

$$\boldsymbol{\phi}(s'|s,\mathbf{a}) = \begin{cases} (\alpha(1-\delta), -\beta\mathbf{a}^\top)^\top, & s = x_h, s' = x_{h+1}, h\in[H]\,; \\ (\alpha\delta, \beta\mathbf{a}^\top)^\top, & s = x_h, s' = x_{H+2}, h\in[H]\,; \\ (\alpha, \mathbf{0}^\top)^\top, & s \in \{x_{H+1}, x_{H+2}\}, s' = s\,; \\ \mathbf{0}, & \text{otherwise}\,. \end{cases},$$

$$\boldsymbol{\theta}_h = (1/\alpha, \boldsymbol{\mu}_h^\top/\beta)^\top,\ h\in[H],$$

where $\alpha = \sqrt{1/(1+\Delta(d-1))}$, $\beta = \sqrt{\Delta/(1+\Delta(d-1))}$. It can be verified that $\boldsymbol{\phi}(\cdot|\cdot,\cdot)$ and $\{\boldsymbol{\theta}_h\}$ satisfy the requirements of a B-bounded linear mixture MDPs. In particular, (1) holds. Indeed, if we let $V : \mathcal{S} \to [0,1]$ be any bounded function then for $s = x_{H+1}$ or $s = x_{H+2}$, $\boldsymbol{\phi}_V(s,\boldsymbol{a}) =$

$\sum_{s'} \boldsymbol{\phi}(s'|s,\boldsymbol{a})V(s') = (\alpha V(s), \mathbf{0}^\top)^\top$ and hence $||\boldsymbol{\phi}_V(s,\boldsymbol{a})||_2 \leq 1$, while for $s = x_h$ with $h \in [H]$, we have

$$||\boldsymbol{\phi}_V(s,\boldsymbol{a})||_2^2 = \alpha^2(V(x_{H+2})\delta + V(x_{h+1})(1-\delta))^2 + \beta^2(V(x_{H+2}) - V(x_{h+1}))^2||\boldsymbol{a}||_2^2$$
$$\leq \alpha^2 + (d-1)\beta^2$$
$$= 1. \tag{69}$$

Meanwhile, since $K \geq (d-1)/(32H(B-1))$, we have

$$\|\boldsymbol{\theta}_h\|_2^2 = \frac{1}{\alpha^2} + \frac{\|\boldsymbol{\mu}_h\|_2^2}{\beta^2} = (1+\Delta(d-1))^2 = (1+\sqrt{\delta/K}/4\sqrt{2}\cdot(d-1))^2 \leq B^2.$$

The initial state in each episode k is $s_{k,1} = x_1$. Note that if the agent transitions to x_{H+2} it remains there until the end of the episode. Due to the special structure of the MDP, at any stage $h \in [H]$, either the state is x_{H+2} or it is x_h. Further, state x_h can only be reached one way, through states $x_1, x_2, \ldots, x_{h-1}$. As such, knowing the current state is equivalent to knowing the history from the beginning of the episode and hence policies that simply decide at the beginning of the episode what actions to take upon reaching a state are as powerful as those that can use the "within episode" history.

Now, clearly, since the only rewarding transitions are those from x_{H+2}, the optimal strategy in stage h when in state x_h is to take action $\mathrm{argmax}_{\boldsymbol{a}\in\mathcal{A}}\langle\boldsymbol{\mu}_h, \boldsymbol{a}\rangle$. Intuitively, the learning problem is not *harder* than minimizing the regret on H linear bandit problems with a shared action set $\mathcal{A} = \{-1,+1\}^{d-1}$ and where the payoff on bandit $h \leq H/2$ of taking action $\boldsymbol{a} \in \mathcal{A}$ is $\Omega(H)Z$, where Z is drawn from a Bernoulli with parameter $\delta + \langle\boldsymbol{\mu}_h, \boldsymbol{a}\rangle$. Some calculation shows that the reverse is also true: Thanks to the choice of δ, $(1-\delta)^{H/2} \approx$ const, hence there is sufficiently high probability of reaching all stages including stage $H/2$, even under the optimal policy. Hence, the MDP learning problem is not easier than solving the first $\Omega(H/2)$ bandit problems. Choosing $\Delta = \Theta(\sqrt{\delta/K})$, for K large enough, $(d-1)\Delta \leq \delta$ so the probabilities are well defined. Furthermore, on each of the bandit, the regret is at least $\Omega(dH\sqrt{K\delta})$. Since there are $\Omega(H/2)$ bandit problems, plugging in the choice of δ, we find that the total regret is $\Omega(dH\sqrt{KH})$ and the result follows by noting that $T = KH$.

Remark 23 *Our lower bound analysis can be adapted to prove a lower bound for linear MDPs proposed in (Yang and Wang, 2019; Jin et al., 2020). In specific, based on our constructed linear mixture MDP M in the proof sketch of Theorem 8, we can construct a linear MDP $\bar{M}(\mathcal{S}, \mathcal{A}, H, \{\bar{r}_h\}, \{\bar{\mathbb{P}}_h\})$ as follows. For each stage $h \in [H]$, the transition probability kernel $\bar{\mathbb{P}}_h$ and the reward function $\bar{r}_h$ are defined as $\bar{\mathbb{P}}_h(s'|s,\mathbf{a}) = \langle\boldsymbol{\phi}(s,\mathbf{a}), \boldsymbol{\mu}_h(s')\rangle$ and $\bar{r}_h(s,\mathbf{a}) = \langle\boldsymbol{\phi}(s,\mathbf{a}), \boldsymbol{\xi}_h\rangle$, where $\boldsymbol{\phi}(s,a), \boldsymbol{\mu}(s') \in \mathbb{R}^{d+1}$ are two feature mappings, and $\boldsymbol{\xi}_h \in \mathbb{R}^{d+1}$ is a parameter vector. Here, we choose $\boldsymbol{\phi}(s,\mathbf{a}), \boldsymbol{\mu}_h(s'), \boldsymbol{\xi}_h \in \mathbb{R}^{d+1}$ as follows:*

$$\boldsymbol{\phi}(s,\mathbf{a}) = \begin{cases} (\alpha, \beta\mathbf{a}^\top, 0)^\top, & s = x_h,\ h \in [H+1]; \\ (0, \mathbf{0}^\top, 1)^\top, & s = x_{H+2}. \end{cases},$$

$$\boldsymbol{\mu}_h(s') = \begin{cases} ((1-\delta)/\alpha, -\boldsymbol{\mu}_h^\top/\beta, 0)^\top, & s' = x_{h+1}; \\ (\delta/\alpha, \boldsymbol{\mu}_h^\top/\beta, 1)^\top, & s' = x_{H+2}; \\ \mathbf{0}, & \textit{otherwise}, \end{cases}$$

and $\boldsymbol{\xi}_h = (\mathbf{0}^\top, 1)^\top$. It can be verified that $\max\{\|\boldsymbol{\xi}_h\|_2, \|\boldsymbol{\mu}_h(\mathcal{S})\|_2\} \leq \sqrt{d+1}$, and $\|\boldsymbol{\phi}(s, \mathbf{a})\|_2 \leq 1$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. In addition, for any $h \in [H]$, we have $\mathbb{P}_h(s'|s, a) = \bar{\mathbb{P}}_h(s'|s, a)$ and $r_h(s, a) = \bar{r}_h(s, a)$ when $s = x_h$ or x_{H+2}. Since at stage h, s can be either x_h or x_{H+2}, we can show that the constructed linear MDP $\bar{M}$ has the same transition probability as the the linear mixture MDP M, which suggests the same lower bound $\Omega(dH\sqrt{T})$ in Theorem 8 also holds for linear MDP.

E.2. Proof of Theorem 8

We select $\delta = 1/H$ as suggested in Appendix E.1. For brevity, with a slight abuse of notation, we will use $M_{\boldsymbol{\mu}}$ to denote the MDP described in Appendix E.1 corresponding to the parameters $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_H)$. We will use $\mathbb{E}_{\boldsymbol{\mu}}$ denote the expectation underlying the distribution generated from the interconnection of a policy and MDP $M_{\boldsymbol{\mu}}$; since the policy is not denoted, we tacitly assume that the identity of the policy will always be clear from the context. We will similarly use $\mathbb{P}_{\boldsymbol{\mu}}$ to denote the corresponding probability measure.

We start with a lemma that will be the basis of our argument that shows that the regret in our MDP can be lower bounded by the regret of $H/2$ bandit instances:

Lemma 24 *Suppose $H \geq 3$ and $3(d-1)\Delta \leq \delta$. Fix $\boldsymbol{\mu} \in (\{-\Delta, \Delta\}^{d-1})^H$. Fix a possibly history dependent policy π and define $\bar{\boldsymbol{a}}_h^\pi = \mathbb{E}_{\boldsymbol{\mu}}[\boldsymbol{a}_h \,|\, s_h = x_h, s_1 = x_1]$: the expected action taken by the policy when it visits state x_h in stage h provided that the initial state is x_1. Then, letting V^* (V^π) be the optimal value function (the value function of policy π, respectively), we have*

$$V_1^*(x_1) - V_1^\pi(x_1) \geq \frac{H}{10} \sum_{h=1}^{H/2} \Big(\max_{\mathbf{a} \in \mathcal{A}} \langle \boldsymbol{\mu}_h, \mathbf{a} \rangle - \langle \boldsymbol{\mu}_h, \bar{\boldsymbol{a}}_h^\pi \rangle \rangle \Big).$$

Proof Fix $\boldsymbol{\mu}$. Since $\boldsymbol{\mu}$ is fixed, we drop the subindex from $\mathbb{P}$ and $\mathbb{E}$. Since $\mathcal{A} = \{+1, -1\}^{d-1}$ and $\boldsymbol{\mu}_h \in \{-\Delta, \Delta\}^{d-1}$, we have $(d-1)\Delta = \max_{\mathbf{a} \in \mathcal{A}} \langle \boldsymbol{\mu}_h, \mathbf{a} \rangle$. Recall the definition of the value of policy π in state x_1:

$$V_1^\pi(x_1) = \mathbb{E}\bigg[\sum_{h=1}^H r_h(s_h, a_h) \bigg| s_1 = x_1, a_h \sim \pi_h(\cdot | s_1, a_1, \ldots, s_{h-1}, a_{h-1}, s_h) \bigg]. \tag{70}$$

Note that by the definition of our MDPs, only x_{H+2} satisfies that $r_h(x_{H+2}, \mathbf{a}) = 1$, all other rewards are zero. Also, once entered, the process does not leave x_{H+2}. Therefore,

$$V_1^\pi(x_1) = \sum_{h=1}^{H-1} (H-h) \mathbb{P}(N_h | s_1 = x_1). \tag{71}$$

where N_h is the event of visiting state x_h in stage h and then entering x_{H+2}:

$$N_h = \{ s_{h+1} = x_{H+2}, s_h = x_h \}. \tag{72}$$

By the law of total probability, the Markov property and the definition of $M_{\boldsymbol{\mu}}$,

$$\mathbb{P}(s_{h+1} = x_{H+2} | s_h = x_h, s_1 = x_1)$$

$$
\begin{aligned}
&= \sum_{\boldsymbol{a}\in\mathcal{A}} \mathbb{P}(s_{h+1} = x_{H+2} | s_h = x_h, a_h = \boldsymbol{a})\mathbb{P}(a_h = \boldsymbol{a} | s_h = x_h, s_1 = x_1) \\
&= \sum_{\boldsymbol{a}\in\mathcal{A}} (\delta + \langle \boldsymbol{\mu}_h, \boldsymbol{a}\rangle)\mathbb{P}(a_h = \boldsymbol{a} | s_h = x_h, s_1 = x_1) \\
&= \delta + \langle \boldsymbol{\mu}_h, \bar{\boldsymbol{a}}_h^\pi \rangle \,,
\end{aligned}
$$

where the last equality used that by definition, $\bar{\boldsymbol{a}}_h^\pi = \sum_{\boldsymbol{a}\in\mathcal{A}} \mathbb{P}(a_h = \boldsymbol{a} | s_h = x_h, s_1 = x_1)\boldsymbol{a}$. It also follows that $\mathbb{P}(s_{h+1} = x_{h+1} | s_h = x_h, s_1 = x_1) = 1 - (\delta + \langle \boldsymbol{\mu}_h, \bar{\boldsymbol{a}}_h^\pi \rangle)$. Hence,

$$
\mathbb{P}(N_h) = (\delta + \langle \boldsymbol{\mu}_h, \bar{\boldsymbol{a}}_h^\pi \rangle) \prod_{j=1}^{h-1} (1 - \delta - \langle \boldsymbol{\mu}_j, \bar{\boldsymbol{a}}_j^\pi \rangle) \,. \tag{73}
$$

Defining $a_h = \langle \boldsymbol{\mu}_h, \bar{\boldsymbol{a}}_h^\pi \rangle$, we get that

$$
V_1^\pi(x_1) = \sum_{h=1}^{H} (H-h)(a_h + \delta) \prod_{j=1}^{h-1} (1 - a_j - \delta) \,.
$$

Working backwards, it is not hard to see that the optimal policy must take at stage the action that maximizes $\langle \boldsymbol{\mu}_h, \boldsymbol{a} \rangle$. Since $\max_{a\in\mathcal{A}} \langle \boldsymbol{\mu}_h, \boldsymbol{a} \rangle = (d-1)\Delta$, we get

$$
V_1^*(x_1) = \sum_{h=1}^{H} (H-h)(1 - (d-1)\Delta - \delta)^{h-1}((d-1)\Delta + \delta).
$$

For $i \in [H]$, introduce

$$
S_i = \sum_{h=i}^{H} (H-h) \prod_{j=i}^{h-1} (1 - a_j - \delta)(a_h + \delta), \; T_i = \sum_{h=i}^{H} (H-h)(1 - (d-1)\Delta - \delta)^{h-i}((d-1)\Delta + \delta).
$$

Then $V_1^*(x_1) - V_1^\pi(x_1) = T_1 - S_1$. To lower bound $T_1 - S_1$, first note that

$$
S_i = (H-i)(a_i + \delta) + S_{i+1}(1 - a_i - \delta), \; T_i = (H-i)((d-1)\Delta + \delta) + T_{i+1}(1 - (d-1)\Delta - \delta),
$$

which gives that

$$
T_i - S_i = (H - i - T_{i+1})((d-1)\Delta - a_i) + (1 - a_i - \delta)(T_{i+1} - S_{i+1}). \tag{74}
$$

Therefore by induction, we get that

$$
T_1 - S_1 = \sum_{h=1}^{H-1} ((d-1)\Delta - a_h)(H - h - T_{h+1}) \prod_{j=1}^{h-1} (1 - a_j - \delta). \tag{75}
$$

To further bound (75), first we note that T_h can be written as the following closed-form expression:

$$
T_h = \frac{(1 - (d-1)\Delta - \delta)^{H-h} - 1}{(d-1)\Delta + \delta} + H - h + 1 - (1 - (d-1)\Delta - \delta)^{H-h},
$$

Hence, for any $h \leq H/2$,

$$H - h - T_{h+1} = \frac{1 - (1 - (d-1)\Delta - \delta)^{H-h}}{(d-1)\Delta + \delta} + (1 - (d-1)\Delta - \delta)^{H-h}$$
$$\geq \frac{1 - (1 - (d-1)\Delta - \delta)^{H/2}}{(d-1)\Delta + \delta} \geq H/3, \tag{76}$$

where the last inequality holds since $3(d-1)\Delta \leq \delta = 1/H$ and $H \geq 3$. Furthermore we have

$$\prod_{j=1}^{h-1}(1 - a_j - \delta) \geq (1 - 4\delta/3)^H \geq 1/3, \tag{77}$$

where the first inequality holds since $a_j \leq (d-1)\Delta, 3(d-1)\Delta \leq \delta$, the second one holds since $\delta = 1/H$ and $H \geq 3$. Therefore, substituting (76) and (77) into (75), we have

$$V_1^*(x_1) - V_1^\pi(x_1) = T_1 - S_1 \geq \frac{H}{10} \cdot \sum_{h=1}^{H/2}((d-1)\Delta - a_h),$$

which finishes the proof. ■

We also need a lower bound on the regret on linear bandits with the hypercube action set $\mathcal{A} = \{-1, 1\}^{d-1}$, Bernoulli bandits with linear mean payoff. While the proof technique used is standard (cf. Lattimore and Szepesvári 2020), we give the full proof as the "scaling" of the reward parameters is nonstandard:

Lemma 25 *Fix a positive real $0 < \delta \leq 1/3$, and positive integers K, d and assume that $K \geq d^2/(2\delta)$. Let $\Delta = \sqrt{\delta/K}/(4\sqrt{2})$ and consider the linear bandit problems $\mathcal{L}_{\boldsymbol{\mu}}$ parameterized with a parameter vector $\boldsymbol{\mu} \in \{-\Delta, \Delta\}^d$ and action set $\mathcal{A} = \{-1, 1\}^d$ so that the reward distribution for taking action $\mathbf{a} \in \mathcal{A}$ is a Bernoulli distribution $B(\delta + \langle \boldsymbol{\mu}^*, \mathbf{a} \rangle)$. Then for any bandit algorithm $\mathcal{B}$, there exists a $\boldsymbol{\mu}^* \in \{-\Delta, \Delta\}^d$ such that the expected pseudo-regret of $\mathcal{B}$ over first K steps on bandit $\mathcal{L}_{\boldsymbol{\mu}^*}$ is lower bounded as follows:*

$$\mathbb{E}_{\boldsymbol{\mu}^*} \textit{Regret}(K) \geq \frac{d\sqrt{K\delta}}{8\sqrt{2}}.$$

Note that the expectation is with respect to a distribution that depends both on $\mathcal{B}$ and $\boldsymbol{\mu}^*$, but since $\mathcal{B}$ is fixed, this dependence is hidden.

Proof Let $\mathbf{a}_k \in \mathcal{A} = \{-1, 1\}^d$ denote the action chosen in round k. Then for any $\boldsymbol{\mu} \in \{-\Delta, \Delta\}^d$, the expected pseudo regret $\mathbb{E}_{\boldsymbol{\mu}}\text{Regret}(K)$ corresponding to $\boldsymbol{\mu}$ satisfies

$$\mathbb{E}_{\boldsymbol{\mu}}\text{Regret}(K) = \sum_{k=1}^K \mathbb{E}_{\boldsymbol{\mu}}(\max_{\mathbf{a} \in \mathcal{A}} \langle \boldsymbol{\mu}, \mathbf{a} \rangle - \langle \boldsymbol{\mu}, \mathbf{a}_k \rangle)$$
$$= \Delta \sum_{k=1}^K \sum_{j=1}^d \mathbb{E}_{\boldsymbol{\mu}} \mathbb{1}\{\text{sgn}([\boldsymbol{\mu}]_j) \neq \text{sgn}([\mathbf{a}_k]_j)\}$$

$$= \Delta \sum_{j=1}^{d} \underbrace{\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\mu}} \mathbb{1}\{\operatorname{sgn}([\boldsymbol{\mu}]_j) \neq \operatorname{sgn}([\mathbf{a}_k]_j)\}}_{N_j(\boldsymbol{\mu})}, \tag{78}$$

where for a vector $\boldsymbol{x}$, we use $[\boldsymbol{x}]_j$ to denote its jth entry. Let $\boldsymbol{\mu}^j \in \{-\Delta, \Delta\}^d$ denote the vector which differs from $\boldsymbol{\mu}$ at its jth coordinate only. Then, we have

$$\begin{aligned}
2\sum_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{\mu}}\text{Regret}(K) &= \Delta \sum_{\boldsymbol{\mu}} \sum_{j=1}^{d} (\mathbb{E}_{\boldsymbol{\mu}} N_j(\boldsymbol{\mu}) + \mathbb{E}_{\boldsymbol{\mu}^j} N_j(\boldsymbol{\mu}^j)) \\
&= \Delta \sum_{\boldsymbol{\mu}} \sum_{j=1}^{d} (K + \mathbb{E}_{\boldsymbol{\mu}} N_j(\boldsymbol{\mu}) - \mathbb{E}_{\boldsymbol{\mu}^j} N_j(\boldsymbol{\mu})) \\
&\geq \Delta \sum_{\boldsymbol{\mu}} \sum_{j=1}^{d} (K - \sqrt{1/2}K\sqrt{\text{KL}(\mathcal{P}_{\boldsymbol{\mu}}, \mathcal{P}_{\boldsymbol{\mu}^j})}), \qquad (79)
\end{aligned}$$

where the inequality holds due to $N_j(\boldsymbol{\mu}) \in [0, K]$ and Pinsker's inequality (Exercise 14.4 and Eq. 14.12, Lattimore and Szepesvári 2020), $\mathcal{P}_{\boldsymbol{\mu}}$ denotes the joint distribution over the all possible reward sequences $(r_1, \ldots, r_K) \in \{0, 1\}^K$ of length K, induced by the interconnection of the algorithm and the bandit parameterized by $\boldsymbol{\mu}$. By the chain rule of relative entropy, $\text{KL}(\mathcal{P}_{\boldsymbol{\mu}}, \mathcal{P}_{\boldsymbol{\mu}^j})$ can be further decomposed as (cf. Exercise 14.11 of Lattimore and Szepesvári 2020),

$$\begin{aligned}
\text{KL}(\mathcal{P}_{\boldsymbol{\mu}}, \mathcal{P}_{\boldsymbol{\mu}^j}) &= \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[\text{KL}(\mathcal{P}_{\boldsymbol{\mu}}(r_k|\mathbf{r}_{1:k-1}), \mathcal{P}_{\boldsymbol{\mu}^j}(r_k|\mathbf{r}_{1:k-1}))] \\
&= \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[\text{KL}(B(\delta + \langle \mathbf{a}_k, \boldsymbol{\mu}\rangle), (B(\delta + \langle \mathbf{a}_k, \boldsymbol{\mu}^j\rangle))] \\
&\leq \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{2\langle \boldsymbol{\mu} - \boldsymbol{\mu}^j, \mathbf{a}_k\rangle^2}{\langle \boldsymbol{\mu}, \mathbf{a}_k\rangle + \delta}\right] \\
&\leq \frac{16K\Delta^2}{\delta}, \qquad (80)
\end{aligned}$$

where the second equality holds since the round k reward's distribution is the Bernoulli distribution $B(\delta + \langle \mathbf{a}_k, \boldsymbol{\mu}\rangle)$ in the environment parameterized by $\boldsymbol{\mu}$, the first inequality holds since for any two Bernoulli distribution $B(a)$ and $B(b)$, we have $\text{KL}(B(a), B(b)) \leq 2(a-b)^2/a$ when $a \leq 1/2, a+b \leq 1$, the second inequality holds since $\boldsymbol{\mu}$ only differs from $\boldsymbol{\mu}^j$ at j-th coordinate, $\langle \boldsymbol{\mu}, \mathbf{a}_k\rangle \geq -d\Delta \geq -\delta/2$. It can be verified that these requirements hold when $\delta \leq 1/3$, $d\Delta \leq \delta/2$. Therefore, substituting (80) into (79), we have

$$2\sum_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{\mu}}\text{Regret}(K) \geq \sum_{\boldsymbol{\mu}} \Delta d(K - \sqrt{2}K^{3/2}\Delta/\sqrt{\delta}) = \sum_{\boldsymbol{\mu}} \frac{d\sqrt{K\delta}}{4\sqrt{2}},$$

where the equality holds since $\Delta = \sqrt{\delta/K}/(4\sqrt{2})$. Selecting $\boldsymbol{\mu}^*$ which maximizes $\mathbb{E}_{\boldsymbol{\mu}}\text{Regret}(K)$ finishes the proof. ■

With this, we are ready to prove Theorem 8.

Proof [Proof of Theorem 8] We can verify that the selection of K, d, H, δ satisfy the requirement of Lemma 24 and Lemma 25. Let π^k denote the possibly nonstationary policy that is executed in episode k given the history up to the beginning of the episode. Then, by Lemma 24, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}}\text{Regret}\Big(M_{\boldsymbol{\mu}}, K\Big) &= \mathbb{E}_{\boldsymbol{\mu}}\bigg[\sum_{k=1}^{K}[V_1^*(x_1) - V_1^{\pi^k}(x_1)]\bigg] \\ &\geq \frac{H}{10}\sum_{h=1}^{H/2}\underbrace{\mathbb{E}_{\boldsymbol{\mu}}\bigg[\sum_{k=1}^{K}\Big(\max_{\mathbf{a}\in\mathcal{A}}\langle\boldsymbol{\mu}_h, \mathbf{a}\rangle - \langle\boldsymbol{\mu}_h, \bar{\boldsymbol{a}}_h^{\pi_k}\rangle\Big)\bigg]}_{I_h(\boldsymbol{\mu},\pi)}.\end{aligned} \tag{81}$$

Let $\boldsymbol{\mu}^{-h} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{h-1}, \boldsymbol{\mu}_{h+1}, \ldots, \boldsymbol{\mu}_H)$. Now, every MDP policy π gives rise to a bandit algorithm $\mathcal{B}_{\pi,h,\boldsymbol{\mu}^{-h}}$ for the linear bandit $\mathcal{L}_{\boldsymbol{\mu}_h}$ of Lemma 25. This bandit algorithm is such that the distribution of action it plays in round k matches the distribution of action played by π in stage h of episode k conditioned on the event that $s_h^k = x_h$, i.e., $\mathbb{P}_{\mu,\pi}(a_h^k = \cdot | s_h^k = x_h)$ with the tacit assumption that the first state in every episode is x_1.

As the notation suggests, the bandit algorithm depends on $\boldsymbol{\mu}^{-h}$. In particular, to play in round k, the bandit algorithm feeds π with data from the MDP kernels up until the beginning of episode k: For $i \neq h$, this can be done by just following $\mathbb{P}_i$ since the parameters of these kernels is known to $\mathcal{B}_{\pi,h,\boldsymbol{\mu}^{-h}}$. When $i = h$, since $\mathbb{P}_h$ is not available to the bandit algorithm, every time it is on stage h, if the state is x_h, it feeds the action obtained from π to $\mathcal{L}_\mu$ and if the reward is 1, it feeds π with the next state x_{H+2}, otherwise it feeds it with next state x_{h+1}. When $i = h$ and the state is not x_h, it can only be x_{H+2}, in which case the next state fed to π is x_{H+2} regardless of the action it takes. At the beginning of episode k, to ensure that state x_h is "reached", π is fed with the states x_1, x_2, $\ldots$, x_h. Then, π is queried for its action, which is the action that the bandit plays in round k. Clearly, by this construction, the distribution of action played in round k by $\mathcal{B}_{\pi,h,\boldsymbol{\mu}^{-h}}$ matches the target.

Denoting by BanditRegret$(\mathcal{B}_{\pi,h,\boldsymbol{\mu}^{-h}}, \boldsymbol{\mu}_h)$ the regret of this bandit algorithm on $\mathcal{L}_{\boldsymbol{\mu}}$, by our construction, $I_h(\boldsymbol{\mu}, \pi) = \text{BanditRegret}(\mathcal{B}_{\pi,h,\boldsymbol{\mu}^{-h}}, \boldsymbol{\mu}_h)$ for all $h \in [H/2]$. Hence,

$$\begin{aligned}\sup_{\boldsymbol{\mu}}\mathbb{E}_{\boldsymbol{\mu}}\text{Regret}\Big(M_{\boldsymbol{\mu}}, K\Big) &\geq \sup_{\boldsymbol{\mu}}\frac{H}{10}\sum_{h=1}^{H/2}\text{BanditRegret}(\mathcal{B}_{\pi,h,\boldsymbol{\mu}^{-h}}, \boldsymbol{\mu}_h) \\ &\geq \sup_{\boldsymbol{\mu}}\frac{H}{10}\sum_{h=1}^{H/2}\inf_{\widetilde{\boldsymbol{\mu}}^{-h}}\text{BanditRegret}(\mathcal{B}_{\pi,h,\widetilde{\boldsymbol{\mu}}^{-h}}, \boldsymbol{\mu}_h) \\ &= \frac{H}{10}\sum_{h=1}^{H/2}\sup_{\boldsymbol{\mu}^h}\inf_{\widetilde{\boldsymbol{\mu}}^{-h}}\text{BanditRegret}(\mathcal{B}_{\pi,h,\widetilde{\boldsymbol{\mu}}^{-h}}, \boldsymbol{\mu}_h) \\ &\geq \frac{H^2}{20}\frac{(d-1)\sqrt{K\delta}}{8\sqrt{2}},\end{aligned}$$

where the last inequality follows by Lemma 25. The result follows by plugging in $\delta = 1/H$ and $T = KH$. ■