Reward-Free Model-Based Reinforcement Learning with Linear Function Approximation

Weitong Zhang
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
weightzero@cs.ucla.edu

Dongruo Zhou
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
drzhou@cs.ucla.edu

Quanquan Gu
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
qgu@cs.ucla.edu

Abstract

We study the model-based reward-free reinforcement learning with linear function approximation for episodic Markov decision processes (MDPs). In this setting, the agent works in two phases. In the exploration phase, the agent interacts with the environment and collects samples without the reward. In the planning phase, the agent is given a specific reward function and uses samples collected from the exploration phase to learn a good policy. We propose a new provably efficient algorithm, called UCRL-RFE under the Linear Mixture MDP assumption, where the transition probability kernel of the MDP can be parameterized by a linear function over certain feature mappings defined on the triplet of state, action, and next state. We show that to obtain an ϵ-optimal policy for arbitrary reward function, UCRL-RFE needs to sample at most $\widetilde{\mathcal{O}}(H^5 d^2 \epsilon^{-2})$ episodes during the exploration phase. Here, H is the length of the episode, d is the dimension of the feature mapping. We also propose a variant of UCRL-RFE using Bernstein-type bonus and show that it needs to sample at most $\widetilde{\mathcal{O}}(H^4 d(H+d)\epsilon^{-2})$ to achieve an ϵ-optimal policy. By constructing a special class of linear Mixture MDPs, we also prove that for any reward-free algorithm, it needs to sample at least $\widetilde{\Omega}(H^2 d\epsilon^{-2})$ episodes to obtain an ϵ-optimal policy. Our upper bound matches the lower bound in terms of the dependence on ϵ and the dependence on d if $H \geq d$.

1 Introduction

In reinforcement learning (RL), the agent sequentially interacts with the environment and receives reward from it. In many real-world RL problems, the reward function is manually designed to encourage the desired behavior of the agent. Thus, engineers have to change the reward function time by time and train the agent to check whether it has achieved the desired behavior. In this case, RL algorithms need to be repeatedly executed with different reward functions and are therefore sample inefficient or even intractable. To tackle this challenge, Jin et al. [9] proposed a new reinforcement learning paradigm called *Reward-Free Exploration* (RFE), which explores the environment without using any reward function. In detail, the reward-free RL algorithm consists of two phases. The first phase is called *Exploration Phase*, where the algorithm explores the environment without receiving reward signals. The second phase is called *Planning Phase*, where the algorithm is given a specific

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

reward function and use the collected data in the first phase to learn the policy. They have shown that this exploration paradigm can learn a near-optimal policy in the planning phase given *any* reward function after collecting polynomial number of episodes in the exploration phase. Follow up work [12, 14, 27] proposed improved algorithms to achieve better or nearly optimal sample complexity.

All the aforementioned works are focused on the tabular Markov decision process (MDP), where the number of states and actions are finite. In practice, the number of states and actions can be large or even infinite, and therefore *function approximation* is required for the sake of computational tractability and generalization. However, the understanding of function approximation for reward-free exploration, even under the simplest linear function approximation, remains underexplored, with only two notable related works [17, 26]. Specifically, Wang et al. [17] studied *linear MDPs* [20, 10], where both the transition probability and the reward function admit linear representations, and proposed a reward-free RL algorithm with $\widetilde{\mathcal{O}}(d^3H^6\epsilon^{-2})$ sample complexity, where d is the dimension of the linear representation, H is the planning horizon, and ϵ is the required accuracy. They also proved that if the optimal state-action function is linear, then the reward-free exploration needs an exponential number of episodes in the planning horizon H to learn a ϵ-optimal policy. Zanette et al. [26] considered a slightly larger class of MDPs with *low inherent Bellman error* [25], and proposed an algorithm with $\widetilde{\mathcal{O}}(d^3H^5\epsilon^{-2})$ sample complexity. However, both works assume the reward function is a linear function over some feature mapping. Moreover, the lower bound proved in [17] is for a very large class of MDPs where the optimal state-action function is linear, thus it is too conservative and cannot tell the information-theoretic limits of reward-free exploration for linear MDPs or related models.

In this paper, we seek a better understanding of the statistical efficiency for reward-free RL with linear function approximation. We propose two reward-free model-based RL algorithms for the finite-horizon episodic *linear mixture/kernel MDP* [7, 3, 27], where the transition probability kernel is a linear mixture model. In detail, our contributions are highlighted as follows:

- We propose a new exploration-driven reward function and its corresponding pseudo value function for linear mixture MDPs, which will encourage the algorithm to explore the state-action pair with more uncertainty on the transition probability.
- We propose a UCRL-RFE algorithm which guides the agent to explore the state space using the exploration-driven reward function and pseudo value functions. We prove an $\widetilde{\mathcal{O}}(H^5d^2\epsilon^{-2})$ sample complexity for UCRL-RFE to achieve an ϵ-optimal policy for any reward function for time-homogeneous MDP.
- We further propose a UCRL-RFE+ algorithm which uses a Bernstein-type exploration bonus. UCRL-RFE+ can reduce the error caused by the exploration-driven reward function during the exploration phase. With a novel analysis based on total variance, we prove an $\widetilde{\mathcal{O}}(H^4d(H+d)\epsilon^{-2})$ sample complexity for UCRL-RFE+, which improves that of UCRL-RFE by a factor of $\min\{H, d\}$.
- By constructing a special class of linear mixture MDPs, we show that any reward-free algorithm needs to sample at least $\widetilde{\Omega}(H^2d\epsilon^{-2})$ episodes to achieve an ϵ-optimal policy for any reward function. This lower bound matches the upper bound of UCRL-RFE+ in terms of the dependence on the accuracy ϵ and feature dimension d when $H \geq d$.

Notation. Scalars and constants are denoted by lower and upper case letters, respectively. Vectors are denoted by lower case bold face letters $\mathbf{x}$, and matrices by upper case bold face letters $\mathbf{A}$. We denote by $[k]$ the set $\{1, 2, \cdots, k\}$ for positive integers k. For two non-negative sequence $\{a_n\}, \{b_n\}$, $a_n = \mathcal{O}(b_n)$ means that there exists a positive constant C such that $a_n \leq Cb_n$, and we use $\widetilde{\mathcal{O}}(\cdot)$ to hide the log factor in $\mathcal{O}(\cdot)$; $a_n = \Omega(b_n)$ means that there exists a positive constant C such that $a_n \geq Cb_n$, and we use $\widetilde{\Omega}(\cdot)$ to hide the log factor. $a_n = o(b_n)$ means that $\lim_{n\to\infty} a_n/b_n = 0$. We denote by S, A as the cardinality of the state set $\mathcal{S}$ and action set $\mathcal{A}$ separately. For a vector $\mathbf{x} \in \mathbb{R}^d$ and corresponding matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, we define $\|\mathbf{x}\|_{\mathbf{A}}^2 = \mathbf{x}^\top\mathbf{A}\mathbf{x}$. We denote $[x]_{(0,H)} := \max\{\min\{x, H\}, 0\}$. For vector $\mathbf{x} \in \mathbb{R}^d$, we denote by $[\mathbf{x}]_i$ the i-th element of $\mathbf{x}$.

2 Related Work

Reinforcement Learning with Function Approximation. Function approximation is extremely useful for RL when the state space and/or the action space are large or even infinite. To develop

provable RL algorithms with linear function approximation, linear MDPs [10] is probably the most widely assumed MDP model, where both the transition kernel and the reward function are linear functions of a given feature mapping. A line of works has developed RL algorithms with polynomial sample complexity or regret bounds under this setting, such as LSVI-UCB [18] and randomized LSVI [23]. Besides the linear MDP, linear mixture/kernel MDPs [7, 3, 29] has emerged as a new model which enables efficient RL with linear function approximation. In this setting, the transition kernel is a linear function over a feature mapping on the triplet of state, action, and next-state. Under this assumption, nearly minimax optimal regrets can be attained for both finite-horizon episodic MDPs and infinite-horizon discounted MDPs [28]. Note also that linear mixture MDPs do not require the reward function to be linear and therefore enables RL with arbitrary reward functions. Therefore, we also consider linear mixture MDPs in this paper.

Algorithm	Sample Complexity	Time Homo.	MDP Type	Model Based
Jin et al. [9]	$\widetilde{\mathcal{O}}(H^5S^2A\epsilon^{-2})$	$\times$	Tabular	$\surd$
RF-UCRL [12]	$\widetilde{\mathcal{O}}(H^4S^2A\epsilon^{-2})$	$\times$	Tabular	$\surd$
RF-Express [14]	$\widetilde{\mathcal{O}}(H^3S^2A\epsilon^{-2})$	$\times$	Tabular	$\surd$
SSTP [27]	$\widetilde{\mathcal{O}}(H^2S^2A\epsilon^{-2})$	$\surd$	Tabular	$\surd$
Lower bound [9]	$\Omega(H^2S^2A\epsilon^{-2})$	$\surd$	Tabular	$\surd$
Wang et al. [17]	$\widetilde{\mathcal{O}}(H^6d^3\epsilon^{-2})$	$\times$	Linear MDP	$\times$
FRANCIS [25]	$\widetilde{\mathcal{O}}(H^5d^3\epsilon^{-2})$	$\surd$	Linear MDP	$\times$
UCRL-RFE (Alg. 2)	$\widetilde{\mathcal{O}}(H^5d^2\epsilon^{-2})$	$\surd$	Linear Mixture	$\surd$
UCRL-RFE+ (Alg. 3)	$\widetilde{\mathcal{O}}(H^4d(H+d)\epsilon^{-2})$	$\surd$	Linear Mixture	$\surd$
Lower bound (Thm. 6.1)	$\widetilde{\Omega}(H^2d\epsilon^{-2})$	$\surd$	Linear MDP/Linear Mixture	$\surd$

Table 1: Comparison of episodic reward-free RL algorithms. *Time Homo.* stands for the MDP is a time-homogeneous, where the transition probabilities are the same at different stages of the episode. *Model Based* stands for the algorithm is a model-based algorithm ($\surd$) or a model-free algorithm ($\times$).

Reward-Free Exploration. As the first work on reward-free exploration, Jin et al. [9] assigned each state an exploration-driven reward function at each round to guide the algorithm to do exploration. Then they utilized the EULER [24] algorithm to minimize the total regret. Their algorithm achieves an $\widetilde{\mathcal{O}}(S^2AH^5\epsilon^{-2})$ sample complexity in the tabular setting to achieve an ϵ-optimal policy, where S is the number of states and A is the number of actions. They also proved a sample complexity lower bound as $\widetilde{\Omega}(S^2AH^2\epsilon^{-2})$. Kaufmann et al. [12] extended the UCRL [2] algorithm to the reward-free exploration. Their algorithm RF-UCRL achieves a sample complexity of $\widetilde{\mathcal{O}}(S^2AH^4\epsilon^{-2})$, which improves that of [9] by a factor of H. Ménard et al. [14] proposed RF-Express algorithm by modifying the UCB-bonus of UCRL to making it decay faster and achieved a sample complexity of $\widetilde{\mathcal{O}}(S^2AH^3\epsilon^{-2})$. Zhang et al. [27] proposed SSTP algorithm in the time-homogeneous setting, which achieves $\widetilde{\mathcal{O}}(S^2AH^2\epsilon^{-2})$ sample complexity, and matches the minimax lower bound provided in [9] up to logarithmic factors. Liu et al. [13] has shown the similarity between the self-play setting and reward-free setting. All of these works are for tabular MDPs.

Here we summarize and compare the related works on Reward Free Exploration in Table 1. Notice that our lower bound $\Omega(H^2d\epsilon^{-2})$ for linear mixture MDPs can imply the same lower bound for linear MDPs and MDPs with low inherent Bellman error, using a similar argument used in [28].

3 Preliminaries

We consider episodic Markov Decision Processes (MDP), which is denoted by a tuple $M(\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \mathbb{P})$. Here $\mathcal{S}$ is the countable state space (may be infinite), $\mathcal{A}$ is the action space, H is the length of the episode, $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function. Without loss of generality, we assume the reward function r_h is *deterministic*. $\mathbb{P}(s'|s, a)$ is the transition probability function which denotes the probability for state s to transit to state s' given action a at step h. A policy $\pi_h : \mathcal{S} \to \mathcal{A}$ is a function which maps a state s to an action a. We define the action-value

function (i.e., Q-function) $Q_h^\pi(s,a)$ as follows:

$$Q_h^\pi(s,a;\{r_h\}_h) = \mathbb{E}\bigg[\sum_{h'=h}^{H} r_{h'}(s_{h'},a_{h'})\bigg| s_h = s, a_h = a\bigg], V_h^\pi(s;\{r_h\}_h) = Q_h^\pi(s,\pi_h(s);\{r_h\}_h).$$

For simplicity, we denote $Q_h^\pi(s,a;r) = Q_h^\pi(s,a;\{r_h\}_h)$ and $V_h^\pi(s;r) = V_h^\pi(s;\{r_h\}_h)$. We define the optimal value function $\{V_h^*\}_{h=1}^H$ and the optimal action-value function $\{Q_h^*\}_{h=1}^H$ as $V_h^*(s;r) = \sup_\pi V_h^\pi(s;r)$ and $Q_h^*(s,a;r) = \sup_\pi Q_h^\pi(s,a;r)$ respectively. For any function $V : \mathcal{S} \to \mathbb{R}$, we denote $[\mathbb{P}V](s,a;r) = \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}V(s';r)$, and denote the variance of V as

$$[\mathbb{V}f](s,a) = [\mathbb{P}f^2](s,a) - \big([\mathbb{P}f](s,a)^2\big). \tag{3.1}$$

In particular, we have the following Bellman equation, as well as the Bellman optimality equation:

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

In this paper, we focus on *model-based* algorithms and make the following *linear mixture MDP* assumption, which assumes that the transition probability $\mathbb{P}$ is a linear mixture of d signed basis measures. Meanwhile, for any function V, we assume that we can do the summation $\sum_{s'\in\mathcal{S}} \boldsymbol{\phi}(s'|s,a)V(s)$ efficiently, e.g., using Monte Carlo method [21].

Definition 3.1 (Linear Mixture MDPs [3, 7, 28])**.** The unknown transition probability $\mathbb{P}$ is a linear combination of d signed basis measures $\phi_i(s'|s,a)$, i.e., $\mathbb{P}(s'|s,a) = \sum_{i=1}^d \phi_i(s'|s,a)\theta_i^*$. Meanwhile, for any $V : \mathcal{S} \to [0,1]$, $i \in [d], (s,a) \in \mathcal{S}\times\mathcal{A}$, the summation $\sum_{s'\in\mathcal{S}} \phi_i(s'|s,a)V(s')$ is computable. For simplicity, let $\boldsymbol{\phi} = [\phi_1,\ldots,\phi_d]^\top$, $\boldsymbol{\theta}^* = [\theta_1^*,\ldots,\theta_d^*]^\top$ and $\boldsymbol{\psi}_V(s,a) = \sum_{s'\in\mathcal{S}} \boldsymbol{\phi}(s'|s,a)V(s)$. Without loss of generality, we assume $\|\boldsymbol{\theta}^*\|_2 \le B$, $\|\boldsymbol{\psi}_V(s,a)\|_2 \le 1$ for all $V : \mathcal{S} \to [0,1]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Remark 3.2. A similar but notably different definition (i.e., linear MDPs [20, 10]) has been used in [17], which assumes that $\mathbb{P}(s'|s,a) = \langle\boldsymbol{\phi}(s,a), \boldsymbol{\mu}(s')\rangle$ and $r_h = \langle\boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h\rangle$, $\boldsymbol{\mu}_h(\cdot)$ is a measure and $\boldsymbol{\theta}_h$ is an unknown vector. Comparing with linear MDPs, linear mixture MDPs do not need the reward function r to be linear, which makes our algorithms more general.

With Definition 3.1, it is easy to verify that the expectation of any bounded function V is a linear function of $\boldsymbol{\psi}$:

$$[\mathbb{P}V](s,a) = \langle\boldsymbol{\psi}_V(s,a), \boldsymbol{\theta}^*\rangle. \tag{3.2}$$

Reward-free RL For reward-free RL, the algorithm can be divided into two phases: *exploration phase* and *planning phase*. In the exploration phase, the algorithm cannot access the reward function but collect K episodes by doing exploration. In the planning phase, the algorithm is given a series of reward functions and find the optimal policy based on these reward functions, using the K episodes collected in the exploration phase. We formally define (ϵ,δ)-learn and sample complexity of the algorithm as follows [9].

Definition 3.3 ((ϵ,δ)-learnability)**.** Given an MDP transition kernel set $\mathcal{P}$, reward function set $\mathcal{R}$ and a initial state distribution μ, we say a reward-free algorithm can (ϵ,δ)-learn the problem $(\mathcal{P},\mathcal{R})$ with sample complexity $K(\epsilon,\delta)$, if for any transition kernel $P \in \mathcal{P}$, after receiving $K(\epsilon,\delta)$ episodes in the exploration phase, for any reward function $r \in \mathcal{R}$, the algorithm returns a policy π in planning phase, such that with probability at least $1-\delta$, $\mathbb{E}_{s_1\sim\mu}[V_1^*(s_1;r) - V_1^\pi(s_1;r)] \le \epsilon$.

4 Algorithm and Main Results

In this section, we propose a reward-free algorithm. This algorithm works as follows: Firstly, during the *exploration phase*, it samples the MDP episodes, build an estimator $\boldsymbol{\theta}$ for the MDP parameter $\boldsymbol{\theta}^*$, and compute the covariance matrix $\boldsymbol{\Sigma}$ of the feature mappings, which characterizes the uncertainty of the estimator $\boldsymbol{\theta}$. Secondly, during the *planning phase*, the algorithm uses the collected $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ in the exploration phase to find the optimal policy π based on the given reward functions.

4.1 Planning phase algorithm

We first introduce the `PLAN` function (Algorithm 1), which is a common module in both planning phase and exploration phase. Given a series of reward functions $\{r_h\}_h$, the goal of `PLAN` function

Algorithm 1 UCRL-RFE Planning Module (PLAN)

Input: Estimated parameter and covariance $\boldsymbol{\theta}, \boldsymbol{\Sigma}$, reward $\{r_h\}_{h=1}^H$, parameter β.

1: For consistency, set $Q_{H+1}(\cdot,\cdot) \leftarrow V_{H+1}(\cdot) \leftarrow 0$
2: **for** $h = H, H-1, \cdots, 1$ **do**
3: Compute Q function as $Q_h(\cdot,\cdot) \leftarrow \left[r_h(\cdot,\cdot) + \langle \boldsymbol{\psi}_{V_{h+1}}(\cdot,\cdot), \boldsymbol{\theta}\rangle + \beta\|\boldsymbol{\psi}_{V_{h+1}}(\cdot,\cdot)\|_{\boldsymbol{\Sigma}^{-1}}\right]_{(0,H)}$
4: Compute value function $V_h(\cdot) \leftarrow \max_{a\in\mathcal{A}} Q_h(\cdot,a)$
5: Compute policy as $\pi_h(\cdot) \leftarrow \operatorname{argmax}_{a\in\mathcal{A}} Q_h(\cdot,a)$.
6: **end for**

Output: Policy $\pi \leftarrow \{\pi_h\}_{h=1}^H$ and $\{V_h\}_{h=1}^H$

is to output the optimal policies $\{\pi_h\}_h$ and Q-functions $\{Q_h\}_h$ corresponding to $\{r_h\}_h$. Suppose the unknown parameter $\boldsymbol{\theta}^*$ is known, we can compute $\{Q_h\}_h$ recursively by the following Bellman equation:

$$Q_h(s,a;r) = r_h(s,a) + [\mathbb{P}V_{h+1}](s,a;r) = r_h(s,a) + \langle \boldsymbol{\psi}_{V_{h+1}}(s,a), \boldsymbol{\theta}^*\rangle. \tag{4.1}$$

However, since $\boldsymbol{\theta}^*$ is unknown, we cannot compute Q_h as in (4.1). Instead, PLAN takes the estimated parameter $\boldsymbol{\theta}$ and the "covariance matrix" $\boldsymbol{\Sigma}$ as input. To calculate Q_h, PLAN replaces $\boldsymbol{\theta}^*$ with the estimated $\boldsymbol{\theta}$ and plus an additional exploration bonus term $\beta\|\boldsymbol{\psi}_{V_{h+1}}(\cdot,\cdot)\|_{\boldsymbol{\Sigma}^{-1}}$ to (4.1), as in Line 3 of Algorithm 1. Then PLAN takes the greedy policy of the calculated optimistic Q_h and proceeds to the previous step. Finally, the algorithm returns policy π in Line 5 as well as the estimated value functions $\{V_h\}_h$.

4.2 Exploration phase algorithm

Based on the introduced PLAN function, we propose the UCRL-RFE algorithm in Algorithm 2. In general, UCRL-RFE guides the agent to explore the unknown state space without the information of the reward functions. In detail, for the k-th episode, UCRL-RFE first defines the *exploration driven reward function* as follows:

$$r_h^k(s,a) = \min\left\{1, \frac{2\beta}{H}\sqrt{\max_{f\in\mathcal{S}\mapsto[0,H-h]} \|\boldsymbol{\psi}_f(s,a)\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}}\right\}, \tag{4.2}$$

where $\boldsymbol{\Sigma}_{1,k}$ is the "covariance matrix" of the feature mapping. Intuitively speaking, $r_h^k(s,a)$ represents the maximum possible uncertainty level of the state-action pair (s,a) caused by the randomness of the MDP transition function, which is *independent* of the true reward functions. Therefore, in order to obtain a good estimation of the optimal policy for any *given* reward functions, it suffices to obtain the optimal policy for $r_h^k(s,a)$. Thus, after obtaining $\{r_h^k\}_h$, UCRL-RFE finds the corresponding near-optimal policies $\{\pi_h^k\}_h$ using PLAN function, with the estimated parameter $\boldsymbol{\theta}_k$ and the "covariance matrix" $\boldsymbol{\Sigma}_{1,k}$ as input. UCRL-RFE uses $\{\pi_h^k\}_h$ as its exploration policy and observes the new episode $s_1^k, a_1^k, \ldots, s_H^k, a_H^k$ induced by $\{\pi_h^k\}_h$.

Next, UCRL-RFE needs to compute the parameters $\boldsymbol{\theta}_{k+1}$ and $\boldsymbol{\Sigma}_{1,k+1}$ for planning in the next episode. Similar to UCRL-VTR proposed by [7, 3], UCRL-RFE also uses a "value-targeted regression (VTR)" estimator, which computes $\boldsymbol{\theta}_{k+1}$ as the minimizer to a ridge regression problem with the target being the past value functions. The main difference between UCRL-RFE and UCRL-VTR is that, due to the lack of true reward functions, UCRL-RFE can not use the estimated value functions as its regression targets. Instead, UCRL-RFE defines the following *pseudo value function* u_h^k:

$$u_h^k = \operatorname*{argmax}_{f\in\mathcal{S}\mapsto[0,H-h]} \boldsymbol{\psi}_f^\top(s_h^k, a_h^k)\boldsymbol{\Sigma}_{1,k}^{-1}\boldsymbol{\psi}_f(s_h^k, a_h^k). \tag{4.3}$$

Here, u_h^k maximizes the "uncertainty" caused by the transition kernel, which will help the agent to explore the state space. Now given the pseudo value functions, Algorithm 2 computes the estimated $\boldsymbol{\theta}_{k+1}$ as the minimizer to the following ridge regression problem:

$$\boldsymbol{\theta}_{k+1} \leftarrow \operatorname*{argmin}_{\boldsymbol{\theta}} \lambda\|\boldsymbol{\theta}\|_2^2 + \sum_{k'=1}^{k}\sum_{h=1}^{H}\left(\langle \boldsymbol{\theta}, \boldsymbol{\psi}_{u_h^{k'}}(s_h^{k'}, a_h^{k'})\rangle - u_h^{k'}(s_{h+1}^{k'})\right)^2, \tag{4.4}$$

which has a closed-form solution as in Line 12. It also updates the covariance matrix $\mathbf{\Sigma}_{1,k+1}$ as in Line 12, by the observed feature mapping $\{\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\}_h$ in the current episode. In the end, after collecting HK state-action samples, UCRL-RFE calculates the policy $\{\pi_h\}$ as output based on $\boldsymbol{\theta}_{K+1}$ and $\mathbf{\Sigma}_{1,K+1}$.

Algorithm 2 UCRL-RFE (Hoeffding Bonus)

Input: Confident parameter β, regularization parameter λ

1: **Phase I: Exploration Phase**
2: Initialize $\mathbf{\Sigma}_{1,1} \leftarrow \lambda \mathbf{I}, \mathbf{b}_1 \leftarrow \boldsymbol{\theta}_1 \leftarrow \mathbf{0}$
3: **for** $k = 1, 2, \cdots, K$ **do**
4: Compute the exploration driven reward function $\{r_h^k(\cdot,\cdot)\}_{h=1}^H$ according to (4.2)
5: Compute exploration policy and value function as $(\{\pi_h^k\}_{h=1}^H, \{V_h^k\}_{h=1}^H) \leftarrow \texttt{PLAN}(\boldsymbol{\theta}_k, \mathbf{\Sigma}_{1,k}, \{r_h^k\}_{h=1}^H, \beta)$
6: Receive the initial state $s_1^k \sim \mu$
7: **for** $h = 1, 2, \cdots, H$ **do**
8: Take action $a_h^k \leftarrow \pi_h^k(s_h^k)$ and receive s_{h+1}^k
9: Calculate u_h^k for s_h^k, a_h^k according to (4.3)
10: Set $\mathbf{\Sigma}_{h+1,k} \leftarrow \mathbf{\Sigma}_{h,k} + \boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)^\top, \mathbf{b}_{h+1,k} \leftarrow \mathbf{b}_{h,k} + \boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)u_h^k(s_{h+1}^k)$
11: **end for**
12: Set $\mathbf{\Sigma}_{1,k+1} \leftarrow \mathbf{\Sigma}_{H+1,k}$, $\mathbf{b}_{1,k+1} \leftarrow \mathbf{b}_{H+1,k}, \boldsymbol{\theta}_{k+1} \leftarrow \mathbf{\Sigma}_{1,k+1}^{-1}\mathbf{b}_{1,k+1}$
13: **end for**
14: **Phase II: Planning Phase**
15: Receive target reward function $\{r_h\}_{h=1}^H$
16: Compute policy as $(\{\pi_h\}_{h=1}^H, \{V_h\}_{h=1}^H) \leftarrow \texttt{PLAN}(\boldsymbol{\theta}_{K+1}, \mathbf{\Sigma}_{1,K+1}, \{r_h\}_{h=1}^H, \beta)$

Output: Policy $\{\pi_h\}_{h=1}^H$

Remark 4.1. Here we do a comparison between our UCRL-RFE and the reward-free RL algorithm in [17]. The main difference is that Wang et al. [17] estimates $\boldsymbol{\theta}_k$ by regression with value function V_h^k being the target, while our UCRL-RFE does regression with the pseudo value function u_h^k being the target. That is mainly due to the different problem settings (linear MDP v.s. linear mixture MDP).

4.3 Implementation details

In general, solving the maximization problem (4.3) is hard. Here, we provide a simple approximate solution to the problem (4.2) and (4.3) for the finite state space case ($|\mathcal{S}| < \infty$). Instead of maximizing the ℓ_2 norm-based objective $\left\|\mathbf{\Sigma}_{1,k}^{-1/2}\boldsymbol{\psi}_f(s_h^k, a_h^k)\right\|_2$, we write $\boldsymbol{\psi}_f(s,a) = \mathbf{\Phi}(s,a)\mathbf{f}$ with $\mathbf{\Phi}(s,a) = (\boldsymbol{\phi}(s,a,S_1), \cdots, \boldsymbol{\phi}(s,a,S_{|\mathcal{S}|}))$ and $\mathbf{f} = (f(S_1), \cdots, f(S_{|\mathcal{S}|}))^\top$, relax the ℓ_2 norm into ℓ_1 norm since $\|\mathbf{x}\|_2 \geq \|\mathbf{x}_1\|_1/\sqrt{d}$ for any $\mathbf{x} \in \mathbb{R}^d$, and maximize the following ℓ_1 norm-based objective

$$\max_{\mathbf{f}} \left\|\mathbf{\Sigma}_{1,k}^{-1/2}\mathbf{\Phi}(s,a)\mathbf{f}\right\|_1 \text{ subject to } \|\mathbf{f}\|_\infty \leq H - h. \tag{4.5}$$

(4.5) can be formulated as a linear programming, which can be solved by interior method [11] or simplex method [5] efficiently. Since $\|\mathbf{x}\|_1/\sqrt{d} \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$, the performance of this approximate solution is guaranteed. For the case where the state space is infinite, we can use state aggregation methods such as soft state aggregation [15] to reduce the infinite state space to finite state space and then apply the above approximate solution to solve it.

4.4 Sample complexity

Now we provide the sample complexity for Algorithm 2.

Theorem 4.2 (Sample complexity of UCRL-RFE)**.** For Algorithm 2, setting parameter $\beta = H\sqrt{d\log(3(1+KH^3B^2)/\delta)} + 1, \quad \lambda = B^{-2}$, then for any $0 < \epsilon < 1$, if $K = \widetilde{\mathcal{O}}(H^5 d^2 \epsilon^{-2})$, we have with probability at least $1-\delta$ that, $\mathbb{E}_{s\sim\mu}[V_1^*(s;r) - V_1^\pi(s;r)] \leq \epsilon$.

Remark 4.3. Theorem 4.2 shows that UCRL-RFE only needs $\text{poly}(d, H, \epsilon^{-1})$ sample complexity to find an ϵ-optimal policy, which suggests that model-based reward-free algorithm is sample-efficient. Thanks to linear function approximation, the sample complexity only depends on the dimension of the feature mapping d and the length of the episode and does not depend on the cardinalities of the state and action spaces.

Corollary 4.4. Under the same conditions as in Theorem 4.2, if solving the relaxed optimization problem in (4.5), Algorithm 2 has $K = \widetilde{\mathcal{O}}(H^5 d^3 \epsilon^{-2})$ sample complexity.

5 Improved Algorithm with Bernstein Bonus

Theorem 4.2 suggests that UCRL-RFE in Algorithm 2 enjoys an $\widetilde{\mathcal{O}}(H^5 d^2 \epsilon^{-2})$ sample complexity to find an ϵ-optimal policy. In this section, we seek to further improve the sample complexity.

A key observation is that for any given reward functions $\{r_h\}_h$, the error between the exploration policy $\{\pi_h\}_h$ and the optimal policy can be decomposed into two parts: *the exploration error* which is the difference between $\{r_h\}_h$ and the exploration driven reward function $\{r_h^k\}_h$, and *the approximation error* which is the difference between the optimal value function $V_1^*(\cdot; r_h^k)$ and our estimated value function $V_1^{\pi_h^k}(\cdot; r_h^k)$ with respect to $\{r_h^k\}_h$. For the latter one, our exploration strategy adapted from VTR is often too conservative since it does not distinguish different value functions and state-action pairs from different episodes and steps. Therefore, inspired by [28], we propose a variant of UCRL-RFE called UCRL-RFE+, which adopts a Bernstein-type bonus for exploration and achieves a better sample complexity.

Algorithm 3 UCRL-RFE+ (Bernstein Bonus)

Input: Parameter $\beta, \widehat{\beta}, \widetilde{\beta}, \check{\beta}$, regularization parameter λ

Stage I: Exploration Phase
Initialize $\mathbf{\Sigma}_{1,1} = \widehat{\mathbf{\Sigma}}_{1,1} = \widetilde{\mathbf{\Sigma}}_{1,1} = \lambda \mathbf{I}, \mathbf{b}_1 = \widehat{\mathbf{b}}_1 = \widetilde{\mathbf{b}}_1 = \boldsymbol{\theta}_1 = \widehat{\boldsymbol{\theta}}_1 = \widetilde{\boldsymbol{\theta}}_1 = \mathbf{0}$
for $k = 1, 2, \cdots, K$ **do**
  Set $\{r_h^k(\cdot,\cdot)\}_{h=1}^H$ to (4.2).
  Compute exploration policy and value function as $(\{\pi_h^k\}_{h=1}^H, \{V_h^k\}_{h=1}^H) \leftarrow \texttt{PLAN}(\widehat{\boldsymbol{\theta}}_k, \widehat{\mathbf{\Sigma}}_{1,k}, \{r_h^k\}_{h=1}^H, \widehat{\beta})$
  Receive the initial state $s_1^k \sim \mu$.
  for $h = 1, 2, \cdots, H$ **do**
    Take action $a_h^k = \pi_h^k(s_h^k)$ and receive s_{h+1}^k
    Calculate u_h^k, ν_h^k for s_h^k, a_h^k according to (4.3) and (5.2) separately
    Set $\mathbf{\Sigma}_{h+1,k} \leftarrow \mathbf{\Sigma}_{h,k} + \boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)^\top$
    Set $\widehat{\mathbf{\Sigma}}_{h+1,k}, \widetilde{\mathbf{\Sigma}}_{h+1,k}, \widehat{\mathbf{b}}_{h+1,k}, \widetilde{\mathbf{b}}_{h+1,k}$ using (5.4)
  end for
  Set $\mathbf{\Sigma}_{1,k+1} \leftarrow \mathbf{\Sigma}_{H+1,k}$
  Set $\widehat{\mathbf{\Sigma}}_{1,k+1} \leftarrow \widehat{\mathbf{\Sigma}}_{H+1,k}, \widehat{\mathbf{b}}_{1,k+1} \leftarrow \widehat{\mathbf{b}}_{H+1,k}, \widehat{\boldsymbol{\theta}}_{k+1} \leftarrow \widehat{\mathbf{\Sigma}}_{1,k+1}^{-1}\widehat{\mathbf{b}}_{1,k+1}$
  Set $\widetilde{\mathbf{\Sigma}}_{1,k+1} \leftarrow \widetilde{\mathbf{\Sigma}}_{H+1,k}, \widetilde{\mathbf{b}}_{1,k+1} \leftarrow \widetilde{\mathbf{b}}_{H+1,k}, \widetilde{\boldsymbol{\theta}}_{k+1} \leftarrow \widetilde{\mathbf{\Sigma}}_{1,k+1}^{-1}\widetilde{\mathbf{b}}_{1,k+1}$
end for
Set $\boldsymbol{\theta}_{K+1} \leftarrow \mathbf{\Sigma}_{1,K+1}^{-1} \sum_{k=1}^K \sum_{h=1}^H \boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k) u_h^k(s_{h+1}^k)$
Stage II: Planning Phase
Receive target reward function $\{r_h\}_{h=1}^H$
Compute exploration policy as $(\{\pi_h\}_{h=1}^H, \{V_h\}_{h=1}^H) \leftarrow \texttt{PLAN}(\boldsymbol{\theta}_{K+1}, \mathbf{\Sigma}_{1,K+1}, \{r_h\}_{h=1}^H, \beta)$

Output: Policy $\{\pi_h\}_{h=1}^H$

5.1 Exploration phase algorithm with Bernstein bonus

UCRL-RFE+ in presented in Algorithm 3. The algorithm structure is similar to that of UCRL-RFE, which can be decomposed into the exploration phase and planning phase. There are two main differences. First, in contrast to UCRL-RFE which uses $\boldsymbol{\theta}_k$ for the PLAN function in both exploration and planning phases, UCRL-RFE+ only uses $\boldsymbol{\theta}_{K+1}$ for the PLAN function in the planning phase. For the exploration phase, UCRL-RFE+ constructs a new estimator $\widehat{\boldsymbol{\theta}}_k$ based on $\{V_{h+1}^{k'}\}_{k' \le k-1,h}$, which are the value functions of the exploration driven rewards. Second, to build $\widehat{\boldsymbol{\theta}}_k$, one way is to choose it as the solution to the ridge regression problem with contexts $\boldsymbol{\psi}_{V_{h+1}^{k'}}(s_h^{k'}, a_h^{k'})$ and targets $V_{h+1}^{k'}(s_{h+1}^{k'})$, similar to (4.4). However, since the targets $V_{h+1}^{k'}(s_{h+1}^{k'})$ have different variances at different steps and episodes, we are actually facing a *heteroscedastic linear regression* problem. Therefore, inspired by a recent line of work [28, 19] which use Bernstein inequality for vector-valued self-normalized

martingale to construct a tighter confidence ball for exploration, we also incorporate the variance to build choose $\widehat{\boldsymbol{\theta}}_k$ as the solution to the following *weighted ridge regression* problem, which is an enhanced estimator for the heteroscedastic case:

$$\widehat{\boldsymbol{\theta}}_k \leftarrow \operatorname*{argmin}_{\boldsymbol{\theta}} \lambda\|\boldsymbol{\theta}\|_2^2 + \sum_{k'=1}^{k-1}\sum_{h=1}^{H}\Big(\big\langle \boldsymbol{\theta}, \boldsymbol{\psi}_{V_{h+1}^{k'}}(s_h^{k'}, a_h^{k'})\big\rangle - V_{h+1}^{k'}(s_{h+1}^{k'})\Big)^2/[\sigma_h^{k'}]^2, \tag{5.1}$$

where $[\sigma_h^{k'}]^2$ is the variance of $V_{h+1}^{k'}(s_{h+1}^{k'})$. The idea to use variances to improve the sample complexity is closely related to the use of "Bernstein bonus" in reward-free RL for the tabular MDPs [12, 27, 14]. Since $\sigma_h^{k'}$ is unknown, we will use $\nu_h^{k'} = [\bar{\sigma}_h^{k'}]^2$ as a plug-in estimator to replace $[\sigma_h^{k'}]^2$ in (5.1). After obtaining $\widehat{\boldsymbol{\theta}}_k$, UCRL-RFE+ sets the $\widehat{\boldsymbol{\Sigma}}_{1,k}$ as the covariance matrix of the features $\boldsymbol{\psi}_{V_{h+1}^k}(s_h^k, a_h^k)/\bar{\sigma}_h^k$, and feeds it into the `PLAN` function with the exploration-driven reward functions and the confidence radius $\widehat{\beta}$. UCRL-RFE+ takes the output $\{\pi_h^k\}_h$ as the exploration policy, and $\{V_h^k\}_h$ as the value functions to construct the estimator $\widehat{\boldsymbol{\theta}}_{k+1}$ for next episode. In the end, when it comes to the planning phase, after receiving reward functions $\{r_h\}_h$, UCRL-RFE+ takes $\boldsymbol{\theta}_{K+1}$ as the solution to the ridge regression problem with contexts $\{\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\}_{k,h}$ and targets $\{u_h^k(s_{h+1}^k)\}_{k,h}$, and the covariance matrix $\boldsymbol{\Sigma}_{1,K+1}$ as input, and uses `PLAN` to find the near optimal policy $\{\pi_h\}_h$ with confidence radius β. It remains to specify ν_h^k in the weighted ridge regression. On the one hand, we need ν_h^k to be an upper bound of $[\sigma_h^k]^2$. On the other hand, we require ν_h^k to have a strictly positive lower bound to let (5.1) be valid. Therefore, we construct ν_h^k as follows:

$$\nu_h^k = \max\{\alpha, \bar{\mathbb{V}}_h^k(s_h^k, a_h^k) + E_k^h(s_h^k, a_h^k)\}, \tag{5.2}$$

where $\bar{\mathbb{V}}_h^k$ is the estimated variance of value function V_h^k and E_h^k is a correction term to calibrate the estimated variance, and $\alpha > 0$ is a positive constant. To compute $\bar{\mathbb{V}}_h^k(s_h^k, a_h^k)$, considering the following fact:

$$[\mathbb{V}V_{h+1}^k](s,a) = [\mathbb{P}[V_{h+1}^k]^2](s,a) - [\mathbb{P}V_{h+1}^k](s,a)^2 = \langle \boldsymbol{\theta}^*, \boldsymbol{\psi}_{[V_{h+1}^k]^2}(s,a)\rangle - \langle \boldsymbol{\theta}^*, \boldsymbol{\psi}_{V_{h+1}^k}(s,a)\rangle^2,$$

it suffices to estimate $\langle \boldsymbol{\theta}^*, \boldsymbol{\psi}_{[V_{h+1}^k]^2}(s,a)\rangle$ and $\langle \boldsymbol{\theta}^*, \boldsymbol{\psi}_{V_{h+1}^k}(s,a)\rangle$ separately. For the first term, $\boldsymbol{\theta}^*$ can be regarded as the unknown parameter of a regression problem w.r.t. contexts $\boldsymbol{\psi}_{[V_{h+1}^{k'}]^2}(s_h^{k'}, a_h^{k'})$ and targets $\boldsymbol{\psi}_{[V_{h+1}^{k'}]^2}(s_h^{k'}, a_h^{k'})$. Therefore, the first term can be estimated by $\big\langle \boldsymbol{\psi}_{[V_{h+1}^k]^2}(s,a), \widetilde{\boldsymbol{\theta}}_k\big\rangle$, where

$$\widetilde{\boldsymbol{\theta}}_k \leftarrow \operatorname*{argmin}_{\boldsymbol{\theta}} \lambda\|\boldsymbol{\theta}\|_2^2 + \sum_{k'=1}^{k-1}\sum_{h=1}^{H}\Big(\big\langle \boldsymbol{\theta}, \boldsymbol{\psi}_{[V_{h+1}^{k'}]^2}(s_h^{k'}, a_h^{k'})\big\rangle - [V_{h+1}^{k'}(s_{h+1}^{k'})]^2\Big)^2.$$

In addition, the second term $\langle \boldsymbol{\theta}^*, \boldsymbol{\psi}_{V_{h+1}^k}(s,a)\rangle$ can be approximated by $\langle \boldsymbol{\psi}_{V_{h+1}^k}(s,a), \widehat{\boldsymbol{\theta}}_k\rangle$. Therefore, the final estimator $[\bar{\mathbb{V}}V_{h+1}^k](s,a)$ is defined as

$$\bar{\mathbb{V}}_h^k(s,a) = \Big[\big\langle \boldsymbol{\psi}_{[V_{h+1}^k]^2}(s,a), \widetilde{\boldsymbol{\theta}}_k\big\rangle\Big]_{(0,H^2)} - \Big[\big\langle \boldsymbol{\psi}_{V_{h+1}^k}(s,a), \widehat{\boldsymbol{\theta}}_k\big\rangle\Big]_{(0,H)}^2. \tag{5.3}$$

For the correction terms E_h^k, we define it as follows:

$$E_h^k(s,a) = \min\Big\{H^2, \widetilde{\beta}\big\|\boldsymbol{\psi}_{[V_{h+1}^k]^2}(s,a)\big\|_{\widetilde{\boldsymbol{\Sigma}}_{1,k}^{-1}}\Big\} + \min\Big\{H^2, 2H\check{\beta}\big\|\boldsymbol{\psi}_{V_{h+1}^k}(s,a)\big\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}^{-1}}\Big\},$$

where $\widetilde{\boldsymbol{\Sigma}}_{1,k}$ is the covariance matrix of the features $\boldsymbol{\psi}_{[V_{h+1}^{k'}]^2}(s_h^{k'}, a_h^{k'})$, $\widetilde{\beta}$, $\check{\beta}$ are two confidence radius. It can be shown that, with these definitions, $\bar{\mathbb{V}}_h^k(s,a) + E_h^k(s,a)$ is an upper bound of $[\sigma_h^k]^2$.

Finally, to enable online update, UCRL-RFE+ updates its covariance matrices recursively as follows, along with sequences $\widehat{\mathbf{b}}_h^k, \widetilde{\mathbf{b}}_h^k$:

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{h+1,k} &\leftarrow \widehat{\boldsymbol{\Sigma}}_{h,k} + \boldsymbol{\psi}_{V_{h+1}^k}(s_h^k, a_h^k)\boldsymbol{\psi}_{V_{h+1}^k}(s_h^k, a_h^k)^\top/\nu_h^k \\
\widetilde{\boldsymbol{\Sigma}}_{h+1,k} &\leftarrow \widetilde{\boldsymbol{\Sigma}}_{h,k} + \boldsymbol{\psi}_{[V_{h+1}^k]^2}(s_h^k, a_h^k)\boldsymbol{\psi}_{[V_{h+1}^k]^2}(s_h^k, a_h^k)^\top \\
\widehat{\mathbf{b}}_{h+1,k} &\leftarrow \widehat{\mathbf{b}}_{h,k} + \boldsymbol{\psi}_{V_{h+1}^k}(s_h^k, a_h^k)V_{h+1}^k(s_{h+1}^k)/\nu_h^k \\
\widetilde{\mathbf{b}}_{h+1,k} &\leftarrow \widetilde{\mathbf{b}}_{h,k} + \boldsymbol{\psi}_{[V_{h+1}^k]^2}(s_h^k, a_h^k)[V_{h+1}^k(s_{h+1}^k)]^2,
\end{aligned} \tag{5.4}$$

where u_h^k is the pseudo value function in (4.3) and ν_h^k is defined in (5.2). Then UCRL-RFE+ computes $\widehat{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_k$ as in Line 14 to Line 15 of Algorithm 3.

5.2 Sample complexity

Now we present the sample complexity for Algorithm 3.

Theorem 5.1 (Sample complexity of UCRL-RFE+)**.** For Algorithm 3, setting $\lambda = B^{-2}$, $\alpha = H^2/d$ in (5.2), and the confidence radius as

$$\begin{aligned}
\widehat{\beta} &= 8\sqrt{d\log(1+KHB^2)\log(48K^2H^2/\delta)} + 4\sqrt{d}\log(48K^2H^2/\delta) + 1 \\
\check{\beta} &= 8d\sqrt{\log(1+KHB^2)\log(48K^2H^2/\delta)} + 4\sqrt{d}\log(48K^2H^2/\delta) + 1 \\
\widetilde{\beta} &= 8H^2\sqrt{d\log(1+KHB^2)\log(48K^2H^2/\delta)} + 4H^2\log(48K^2H^2/\delta) + 1 \\
\beta &= H\sqrt{d\log(12(1+KH^3B^2)/\delta)} + 1,
\end{aligned}$$

then for any $0 < \epsilon < 1$, if $K = \widetilde{\mathcal{O}}(H^4 d(H+d)\epsilon^{-2})$, then with probability at least $1-\delta$, we have $\mathbb{E}_{s\sim\mu}[V_1^*(s;r) - V_1^\pi(s;r)] \le \epsilon$.

Remark 5.2. Theorem 5.1 suggests that when $d \ge H$, the sample complexity of UCRL-RFE+ is $\widetilde{\mathcal{O}}(H^4 d^2 \epsilon^{-2})$, which improves the sample complexity of UCRL-RFE by a factor of H. On the other hand, when $H \ge d$, the sample complexity of UCRL-RFE+ reduces to $\widetilde{\mathcal{O}}(H^5 d\epsilon^{-2})$, which is better than that of UCRL-RFE by a factor of d. At a high-level, the sample complexity improvement is attributed to the Bernstein-type bonus.

Corollary 5.3. Under the same conditions as in Theorem 5.1, if solving the relaxed optimization problem in (4.5), Algorithm 3 has $K = \widetilde{\mathcal{O}}(H^5 d^3 \epsilon^{-2})$ sample complexity.

6 Lower Bound of Sample Complexity

In this section, we will provide a lower bound of sample complexity for reward-free RL under linear mixture MDP setting.

The proof is by construction. Given $d \ge 2$, we first define a binary vector set $\mathcal{M} = \{\mathbf{x} | \mathbf{x} \in \mathbb{R}^{d-1}, [\mathbf{x}]_i \in \{-1, 1\}\}$. We index each vector in $\mathcal{M}$ as $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{|\mathcal{M}|}$. Equipped with the set $\mathcal{M}$, we construct a class of MDPs. As shown in Figure 1, there are in total three states $S_1, S_{2,1}, S_{2,2}$ and $|\mathcal{A}| = |\mathcal{M}|$ actions $a_1, a_2, \cdots a_{|\mathcal{A}|}$. We define the feature mapping $\boldsymbol{\phi}(s'|s, a_i) \in \mathbb{R}^d$ as follows:

$$\boldsymbol{\phi}(S_{2,1}|S_1, a_j) = \begin{pmatrix} \frac{\sqrt{2}}{4} & \frac{\mathbf{a}_j^\top}{\sqrt{2d}} \end{pmatrix}^\top,$$

$$\boldsymbol{\phi}(S_{2,2}|S_1, a_j) = \begin{pmatrix} \frac{\sqrt{2}}{4} & -\frac{\mathbf{a}_j^\top}{\sqrt{2d}} \end{pmatrix}^\top,$$

$\boldsymbol{\phi}(S_{2,j}|S_{2,j}, a_i) = \begin{pmatrix} 1/\sqrt{2} & \mathbf{0}^\top \end{pmatrix}^\top$ for $j = 1, 2$, and $\boldsymbol{\phi}(s'|s, a) = \mathbf{0}$ for all the remaining cases. Furthermore, we define a d-dimensional parameter set $\boldsymbol{\Theta} \subseteq \mathbb{R}^{d+1}$ by $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_i | \boldsymbol{\theta}_i = (\sqrt{2}, \alpha\widetilde{\boldsymbol{\theta}}_i^\top/\sqrt{d})^\top\}$ where $\widetilde{\boldsymbol{\theta}}_i = \mathbf{x}_i \in \mathcal{M}$ and α is a small absolute constant. Therefore, for each parameter $\boldsymbol{\theta}_i$, we define the transition probability of the linear mixture MDP as $\mathbb{P}(\cdot|\cdot,\cdot) = \langle \boldsymbol{\phi}(\cdot|\cdot,\cdot), \boldsymbol{\theta}_i \rangle$. Specifically, the transition between S_1 and $\{S_{2,1}, S_{2,2}\}$ is represented as

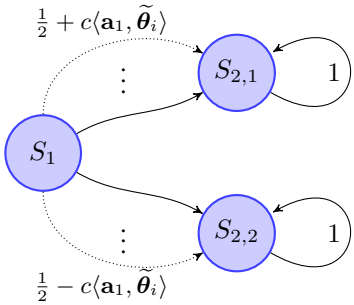

Figure 1: The transition kernel $\mathbb{P}$ of the class of hard-to-learn linear mixture MDPs. The kernel $\mathbb{P}$ is parameterized by $\boldsymbol{\theta}_i = (\sqrt{2}, \alpha\widetilde{\boldsymbol{\theta}}_i^\top/\sqrt{d})^\top$ for some small α. $c = \alpha/(\sqrt{2}d)$. The learner knows the MDP structure, but does not know the parameter $\boldsymbol{\theta}_i$ (or $\widetilde{\boldsymbol{\theta}}_i \in \mathcal{M}$).

$$\mathbb{P}_{\boldsymbol{\theta}_i}(S_{2,1}|S_1, a_j) = \frac{1}{2} + \frac{\alpha}{\sqrt{2}d}\langle \widetilde{\boldsymbol{\theta}}_i, \mathbf{a}_j \rangle, \quad \mathbb{P}_{\boldsymbol{\theta}_i}(S_{2,2}|S_1, a_j) = \frac{1}{2} - \frac{\alpha}{\sqrt{2}d}\langle \widetilde{\boldsymbol{\theta}}_i, \mathbf{a}_j \rangle.$$

Meanwhile, we have $S_{2,1}$ and $S_{2,2}$ are both absorbing states. With the constructed hard-to-learn MDP class, we can prove the lower bound of sample complexity as follows:

Theorem 6.1. Given dimension $d \geq 50$ and $H \geq 2$, set $\epsilon \leq (H-1)/(2\sqrt{2})$ and $\delta \in (0, 1/2)$, then there exists a class of linear mixture MDPs, such that any reward-free RL algorithm that (ϵ, δ)-learns the problem $(\mathcal{P}, \mathcal{R})$ where $\mathcal{R} = \{\{r_h\}_{h=1}^H, 0 \leq r_h \leq 1\}$, needs to collect at least $K = C(1-\delta)dH^2\epsilon^{-2}$ episodes during exploration, where C is an absolute constant.

Remark 6.2. When $d \leq H$, the sample complexity of UCRL-RFE+ is $\widetilde{\mathcal{O}}(H^5 d\epsilon^{-2})$, which matches the lower bound in terms of both ϵ and d, ignoring the logarithmic terms. When $d > H$, the sample complexity of UCRL-RFE+ is $\widetilde{\mathcal{O}}(H^4 d^2\epsilon^{-2})$, which matches the lower bound only in terms of ϵ. The factor of d gap between the upper and lower bounds is due to the fact that our upper bound holds for the arbitrary number of actions. Such a gap also appears in best-arm identification in the linear bandits problem (See Eq. (3) in Tao et al. [16] with $N = \mathcal{O}(2^d)$). There is also a factor of H^2 gap between the upper and lower bounds, and we leave it as future work to remove this gap.

7 Conclusion

We studied model-based reward-free exploration for learning the linear mixture MDPs. We proposed two algorithms, UCRL-RFE, and UCRL-RFE+, which are guaranteed to have polynomial sample complexities in exploration to find a near-optimal policy in the planning phase for any given reward function. To our knowledge, these are the first algorithms and theoretical guarantees for model-based reward-free RL with function approximation. We also give a sample complexity lower bound for any reward-free algorithm to learn linear mixture MDPs. The sample complexity of our algorithm UCRL-RFE+ matches the lower bound in terms of the dependence on accuracy ϵ and feature dimension d when $H \geq d$.

References

[1] ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*.

[2] AUER, P., JAKSCH, T. and ORTNER, R. (2009). Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*.

[3] AYOUB, A., JIA, Z., SZEPESVARI, C., WANG, M. and YANG, L. (2020). Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*. PMLR.

[4] AZUMA, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series* **19** 357–367.

[5] DANTZIG, G. B. (1965). *Linear programming and extensions*, vol. 48. Princeton university press.

[6] FANO, R. M. (1961). Transmission of information: A statistical theory of communications. *American Journal of Physics* **29** 793–794.

[7] JIA, Z., YANG, L., SZEPESVARI, C. and WANG, M. (2020). Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*. PMLR.

[8] JIN, C., ALLEN-ZHU, Z., BUBECK, S. and JORDAN, M. I. (2018). Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*.

[9] JIN, C., KRISHNAMURTHY, A., SIMCHOWITZ, M. and YU, T. (2020). Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*. PMLR.

[10] JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*.

[11] KARMARKAR, N. (1984). A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*.

[12] KAUFMANN, E., MÉNARD, P., DOMINGUES, O. D., JONSSON, A., LEURENT, E. and VALKO, M. (2021). Adaptive reward-free exploration. In *Algorithmic Learning Theory*. PMLR.

[13] LIU, Q., YU, T., BAI, Y. and JIN, C. (2020). A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604* .

[14] MÉNARD, P., DOMINGUES, O. D., JONSSON, A., KAUFMANN, E., LEURENT, E. and VALKO, M. (2020). Fast active learning for pure exploration in reinforcement learning. *arXiv preprint arXiv:2007.13442* .

[15] MICHAEL, S. P. S. T. J. and JORDAN, I. (1995). Reinforcement learning with soft state aggregation. *Advances in neural information processing systems 7* **7** 361.

[16] TAO, C., BLANCO, S. and ZHOU, Y. (2018). Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*.

[17] WANG, R., DU, S. S., YANG, L. F. and SALAKHUTDINOV, R. (2020). On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems* .

[18] WANG, Y., WANG, R., DU, S. S. and KRISHNAMURTHY, A. (2019). Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*.

[19] WU, Y., ZHOU, D. and GU, Q. (2021). Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. *arXiv preprint arXiv:2102.07301* .

[20] YANG, L. and WANG, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*. PMLR.

[21] YANG, L. and WANG, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*. PMLR.

[22] YAO, A. C.-C. (1977). Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*. IEEE.

[23] ZANETTE, A., BRANDFONBRENER, D., BRUNSKILL, E., PIROTTA, M. and LAZARIC, A. (2020). Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*.

[24] ZANETTE, A. and BRUNSKILL, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*. PMLR.

[25] ZANETTE, A., LAZARIC, A., KOCHENDERFER, M. and BRUNSKILL, E. (2020). Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*. PMLR.

[26] ZANETTE, A., LAZARIC, A., KOCHENDERFER, M. J. and BRUNSKILL, E. (2020). Provably efficient reward-agnostic navigation with linear value iteration. *Advances in Neural Information Processing Systems* .

[27] ZHANG, Z., DU, S. S. and JI, X. (2020). Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901* .

[28] ZHOU, D., GU, Q. and SZEPESVARI, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*. PMLR.

[29] ZHOU, D., HE, J. and GU, Q. (2021). Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*. PMLR.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Remark 6.2 addressing the gap between the lower bound and the upper bound
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We are focusing on the theoretical analysis for Reinforcement Learning, no negative social impact can be found at this point
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs of Upper Bounds

In this section, we provide the proofs of sample complexity upper bounds.

A.1 Proof of Theorem 4.2

We will first introduce a lemma to show that for the planning module Algorithm 1, if it is guaranteed that the estimation $\boldsymbol{\theta}$ is close to the true parameter $\boldsymbol{\theta}^*$, then the estimated value function is optimistic. Also the gap between the optimal value function and the value function of the output policy $\{\pi_h\}_{h=1}^H$ could be controlled by the summation of UCB bonus term.

Lemma A.1. Let $\boldsymbol{\theta}, \boldsymbol{\Sigma}, \beta$ be as defined in Algorithm 1. Suppose there exists some event $\boldsymbol{\xi}$ such that $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_{\boldsymbol{\Sigma}} \leq \beta$ on this event. Then on this event, for all $s \in \mathcal{S}$, $V_1(s) \geq V_1^*(s; r)$, where V_1 is the output value function for Algorithm 1. We also have that

$$V_1(s) - V_1^\pi(s) \leq \mathbb{E}\Big[\sum_{h=1}^H \min\{H, 2\beta\|\boldsymbol{\psi}_{V_{h+1}}(s_h, \pi_h(s_h))\|_{\boldsymbol{\Sigma}^{-1}}\}\Big| s, \pi\Big],$$

where the policy $\pi = \{\pi_h\}_{h=1}^H$ is generated by the planning module Algorithm 1 and V_h is the value function calculated on Line 5 in Algorithm 1.

Next we will give the lemmas on how to guarantee the condition of Lemma A.1 and how to utilize the result of that lemma to control the final policy error $V_1^*(s_1; r) - V_1^\pi(s_1; r)$ where the policy π is output of the planning phase. We start with Algorithm 2, which uses the Hoeffding bonus.

Firstly, the next lemma shows how to guarantee the condition in Lemma A.1.

Lemma A.2 (Confidence interval, Hoeffding)**.** For Algorithm 2, let λ, β be as defined in Theorem 4.2, then with probability at least $1 - \delta/3$, $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_k\|_{\boldsymbol{\Sigma}_{1,k}} \leq \beta$ for any $k \in [K+1]$.

Secondly, based on the lemma above, we find that the policy error during the planning phase is controlled by a summation of the UCB terms. Since from the intuition, the exploration driven reward function (4.2) is the UCB term divided by H, the policy error during the planning phase can be converted to the value function V_1^k in the exploration phase. The next lemma shows that the summation of V_1^k over K iterations is sub-linear to K, thus the policy error during the planning phase should be small.

Lemma A.3 (Summation, Hoeffding)**.** Set the parameters of Algorithm 2 as that of Theorem 4.2. If the condition in Lemma A.2 holds, then with probability at least $1 - \delta/3$, the summation of the value function $V_1^k(s_1^k)$ during the exploration phase is controlled by

$$\sum_{k=1}^K V_1^k(s_1^k) \leq 8\beta\sqrt{HKd\log(1 + KH^3B^2/d)} + 8\beta Hd\log(1 + KH^3B^2) + 2H\sqrt{2HK\log(1/\delta)}.$$

Equipped with these lemmas, we are about to prove Theorem 4.2.

Proof of Theorem 4.2. In the following proof, we condition on the events in Lemma A.2 and Lemma A.3 which holds with probability at least $1 - 2\delta/3$ by taking the union bound. Applying Lemma A.1 to the final planning phase, we have

$$V_1^*(s; r) - V_1^\pi(s; r) \leq V_1(s; r) - V_1^\pi(s; r) \leq \underbrace{\mathbb{E}\Big[\sum_{h=1}^H \min\{H, 2\beta\|\boldsymbol{\psi}_{V_{h+1}}(s_h, \pi_h(s_h))\|_{\boldsymbol{\Sigma}_{1,K+1}^{-1}}\}\Big]}_{I_1}, \tag{A.1}$$

where the expectation is taken condition on initial state s and policy π generated by the planning phase. Since $\boldsymbol{\Sigma}_{1,k} \preceq \boldsymbol{\Sigma}_{1,K+1}$ for all $k \in [K]$, we can guarantee that $\|\boldsymbol{\psi}_{V_{h+1}}(s_h, \pi_h(s_h))\|_{\boldsymbol{\Sigma}_{1,K+1}^{-1}} \leq \|\boldsymbol{\psi}_{V_{h+1}}(s_h, \pi_h(s_h))\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}$. Recall the exploration driven reward function is defined by

$$r_h^k(s, a) = \min\Big\{1, \frac{2\beta}{H}\sqrt{\max_{f \in \mathcal{S} \mapsto [0, H-h]} \|\boldsymbol{\psi}_f(s, a)\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}}\Big\}, \text{(A.3)} \tag{A.2}$$

one can easily verify that $\min\{H, 2\beta\|\boldsymbol{\psi}_{V_{h+1}}(s_h, \pi_h(s_h))\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}\} \leq H r_h^k(s_h, \pi_h(s_h))$. Therefore for any $k \in [K]$ episode, we can bound the term I_1 using the value function $V_1^\pi(s; \{r_h^k\}_{h=1}^H)$ of the output policy π in the planning phase given the $\{r_h^k\}_{h=1}^H$ as the reward function, i.e.

$$I_1 \leq \mathbb{E}\bigg[\sum_{h=1}^H H r_h^k(s_h, \pi_h(s_h))\bigg] = H V_1^\pi(s; \{r_h^k\}_{h=1}^k). \tag{A.3}$$

Plugging the bound of I_1 back into (A.1) then taking the expectation over the initial state distribution μ, we have for any $k \in [K]$,

$$\begin{aligned}\mathbb{E}_{s\sim\mu}[V_1^*(s;r) - V_1^\pi(s;r)] &\leq H\mathbb{E}_{s\sim\mu}[V_1^\pi(s; \{r_h^k\}_{h=1}^k)] \\ &= H\Big(V_1^\pi(s_1^k; \{r_h^k\}_{h=1}^k) - V_1^\pi(s_1^k; \{r_h^k\}_{h=1}^k)\Big) \\ &\quad + H\mathbb{E}_{s\sim\mu}[V_1^\pi(s; \{r_h^k\}_{h=1}^k)].\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E}_{s\sim\mu}[V_1^*(s;r) - V_1^\pi(s;r)] \leq \frac{H}{K}\sum_{k=1}^K \Big(&V_1^\pi(s_1^k; \{r_h^k\}_{h=1}^k) - V_1^\pi(s_1^k; \{r_h^k\}_{h=1}^k) \\ &+ \mathbb{E}_{s\sim\mu}[V_1^\pi(s; \{r_h^k\}_{h=1}^k)]\Big).\end{aligned} \tag{A.4}$$

Since $V_1^\pi(s; \{r_h^k\}_{h=1}^k) \leq H$ for all $k \in [K], s \in \mathcal{S}$, by Azuma-Hoeffding's inequality, with probability at least $1 - \delta/3$,

$$\sum_{k=1}^K \Big(\mathbb{E}_{s\sim\mu}[V_1^\pi(s; \{r_h^k\}_{h=1}^k)] - V_1^\pi(s_1^k; \{r_h^k\}_{h=1}^k)\Big) \leq H\sqrt{2K\log(3/\delta)}. \tag{A.5}$$

By plugging (A.5) into (A.4), we have

$$\mathbb{E}_{s\sim\mu}[V_1^*(s;r) - V_1^\pi(s;r)] \leq \frac{H}{K}\sum_{k=1}^K V_1^\pi(s_1^k; \{r_h^k\}_{h=1}^k) + H^2\sqrt{2\log(3/\delta)/K}.$$

Applying Lemma A.1 to the exploration phase, for any k-th episode, $V_1^\pi(s_1^k; \{r_h^k\}_{h=1}^k) \leq V_1^*(s_1^k; \{r_h^k\}_{h=1}^k) \leq V_1^k(s_1^k)$, thus replacing the value function V_1^π with the estimated value function V_1^k, we have

$$\mathbb{E}_{s\sim\mu}[V_1^*(s;r) - V_1^\pi(s;r)] \leq \frac{H}{K}\sum_{k=1}^K V_1^k(s_1^k) + H^2\sqrt{2\log(3/\delta)/K}. \tag{A.6}$$

Finally by Lemma A.3 we can bound the summation over V_1^k, hence

$$\begin{aligned}\mathbb{E}_{s\sim\mu}[V_1^*(s;r) - V_1^\pi(s;r)] \leq\ & H^2\sqrt{2\log(3/\delta)/K} + 8\beta\sqrt{H^3 d\log(1 + KH^3B^2/d)/K} \\ &+ 8\beta d H^2\log(1 + KH^3B^2)/K + 2H^2\sqrt{2H\log(1/\delta)/K}\end{aligned}$$

and by taking union bound, the result holds with probability at least $1 - \delta$. Recall the setting of $\beta \sim \widetilde{\mathcal{O}}(H\sqrt{d})$ as in Theorem 4.2, let $K = \widetilde{\mathcal{O}}(H^5 d^2 \epsilon^{-2})$, the policy error $\mathbb{E}_{s\sim\mu}[V_1^*(s;r) - V_1^\pi(s;r)]$ is bounded by ϵ. □

A.2 Proof of Corollary 4.4

Proof of Corollary 4.4. Following the proof of Theorem 4.2, since for all $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_2 \leq \sqrt{d}\|\mathbf{x}\|_1$ it follows that

$$\begin{aligned}\|\boldsymbol{\psi}_{V_{h+1}}(s_h, \pi_h(s_h))\|_{\boldsymbol{\Sigma}_{1,K+1}^{-1}} &= \|\boldsymbol{\Sigma}_{1,K+1}^{-1/2}\boldsymbol{\psi}_{V_{h+1}}(s_h, \pi_h(s_h))\|_2 \\ &\leq \sqrt{d}\|\boldsymbol{\Sigma}_{1,K+1}^{-1/2}\boldsymbol{\psi}_{V_{h+1}}(s_h, \pi_h(s_h))\|_1.\end{aligned} \tag{A.7}$$

We denote $\widetilde{u}_h^k$ as the result using the ℓ_1 norm as the surrogate objective function in this optimization problem (4.5), i.e.

$$\widetilde{u}_h^k := \operatorname*{argmax}_{f \in \mathcal{S} \mapsto [0, H-h]} \|\boldsymbol{\Sigma}_{1,k}^{-1/2} \boldsymbol{\psi}_f(s_h^k, a_h^k)\|_1,$$

then (A.7) yields

$$\begin{aligned}
\|\boldsymbol{\psi}_{V_{h+1}}(s_h, \pi_h(s_h))\|_{\boldsymbol{\Sigma}_{1,K+1}^{-1}} &\leq \sqrt{d} \|\boldsymbol{\Sigma}_{1,K+1}^{-1/2} \boldsymbol{\psi}_{V_{h+1}}(s_h, \pi_h(s_h))\|_1 \\
&\leq \sqrt{d} \|\boldsymbol{\Sigma}_{1,K+1}^{-1/2} \boldsymbol{\psi}_{\widetilde{u}_h^k}(s_h, \pi_h(s_h))\|_1 \\
&\leq \sqrt{d} \|\boldsymbol{\Sigma}_{1,K+1}^{-1/2} \boldsymbol{\psi}_{\widetilde{u}_h^k}(s_h, \pi_h(s_h))\|_2 \\
&\leq \sqrt{d} \|\boldsymbol{\Sigma}_{1,K+1}^{-1/2} \boldsymbol{\psi}_{u_h^k}(s_h, \pi_h(s_h))\|_2,
\end{aligned}$$

where the second inequality comes from $\widetilde{u}_h^k$ is the solution in (4.5), the third inequality comes from the fact that $\|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_2$ and the forth inequality comes from the definition that u_h^k. Then (A.3) is changed to be

$$I_1 \leq H\sqrt{d} V_1^\pi(s, \{r_h^k\}_{h=1}^k).$$

Noticing that comparing to the original result, there's an additional $\sqrt{d}$ factor which yields (A.7)

$$\mathbb{E}_{s\sim\mu}[V_1^*(s;r) - V_1^\pi(s;r)] \leq \frac{H\sqrt{d}}{K} \sum_{k=1}^K V_1^k(s_1^k) + H^2\sqrt{2d\log(3/\delta)/K}.$$

Then it is easy to show that using ℓ_1 as the surrogate objective function, the sample complexity of Algorithm 2 turns out to be $\widetilde{\mathcal{O}}(H^5 d^3 \epsilon^{-2})$ □

A.3 Proof of Theorem 5.1

We are going to analyze Algorithm 3 and provide the proof of Theorem 5.1. Following the proof of Theorem 4.2, we only need to revise Lemmas A.2 and A.3 to continue the proof of Theorem 5.1.

Lemma A.4 (Confidence interval, Bernstein)**.** Let $\beta, \widehat{\beta}, \widetilde{\beta}, \check{\beta}$ and λ be defined as Theorem 5.1, then with probability at least $1 - \delta/3$, for all $k \in [K+1]$,

$$\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}} \leq \widehat{\beta},\ \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\check{\boldsymbol{\Sigma}}_{1,k}} \leq \check{\beta},\ \|\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}_k\|_{\widetilde{\boldsymbol{\Sigma}}_{1,k}} \leq \widetilde{\beta},\ \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{K+1}\|_{\boldsymbol{\Sigma}_{1,K+1}} \leq \beta, \tag{A.8}$$

and $|[\mathbb{V}_h V_{h+1}^k](s,a) - \bar{\mathbb{V}}_h^k(s,a)| \leq E_h^k(s,a)$.

Lemma A.5 (Summation, Bernstein)**.** For Algorithm 2, setting its parameters as in Lemma A.2, with probability at least $1 - \delta/3$, the summation of the value function during exploration phase is controlled by

$$\sum_{k=1}^K V_1^k(s_1^k) \leq \widetilde{\mathcal{O}}(\sqrt{H^3 K d} + H d\sqrt{K}) + o(\sqrt{K}).$$

Proof of Theorem 5.1. The proof is almost the same as the proof of Theorem 4.2 by replacing Lemma A.2 with Lemma A.4, Lemma A.3 with Lemma A.5. In detail, following the same method, (A.6) works for Algorithm 3 under the condition in Lemma A.4 holds. Therefore, by using Lemma A.5 instead of Lemma A.3, with probability at least $1 - \delta$,

$$\begin{aligned}
\mathbb{E}_{s\sim\mu}[V_1^*(s;r) - V_1^\pi(s;r)] &\leq \frac{H}{K} \sum_{k=1}^K V_1^k(s_1^k) + H^2\sqrt{2\log(3/\delta)/K} \\
&\leq \widetilde{\mathcal{O}}\Big((\sqrt{H^4 d^2} + \sqrt{H^5 d})/\sqrt{K}\Big).
\end{aligned}$$

Letting $K = \widetilde{\mathcal{O}}(H^4 d(H+d)\epsilon^{-2})$, the policy error for the planning phase could be controlled by $\mathbb{E}_{s\sim\mu}[V_1^*(s;r) - V_1^\pi(s;r)] \leq \epsilon$. □

A.4 Proof of Corollary 5.3

Proof of Corollary 5.3. The proof is almost the same as proof of Corollary 4.4, by adding the additional dependency d into the regret bound achieved by Theorem 5.1, it's easy to verify that the sample complexity using the ℓ_1 norm as the surrogate function (4.5) is $\widetilde{\mathcal{O}}(H^4 d^2 (H+d)\epsilon^{-2})$ $\square$

B Missing Proofs in Appendix A

B.1 Filtration

For the simplicity of further proof, we define the event filtration here as

$$\mathcal{G}_{h,k} = \big\{\{s_i^\kappa, a_i^\kappa\}_{i=1,\kappa=1}^{H,k-1}, \{s_i^k, a_i^k\}_{i=1}^{h-1}\big\},$$

it is easy to verify that s_h^k is $\mathcal{G}_{h+1,k}$-measurable. Also, since π^k is $\mathcal{G}_{h,k}$-measurable for all $h \in [H]$, $a_h^k = \pi_h^k(s_h^k)$ is also $\mathcal{G}_{h+1,k}$-measurable. Also, for any function $f \leq R$ built on $\mathcal{G}_{h+1,k}$, such as V_{h+1}^k, u_h^k, $f(s_{h+1}^k) - [\mathbb{P}f](s_h^k, a_h^k)$ is $\mathcal{G}_{h+1,k}$-measurable and it is also a zero-mean R-sub-Gaussian conditioned on $\mathcal{G}_{h+1,k}$.

Since $\mathcal{G}_{H+1,k} = \mathcal{G}_{1,k+1}$, we could arrange the filtration as

$$\mathcal{G} = \{\mathcal{G}_{1,1}, \cdots, \mathcal{G}_{H,1}, \cdots, \mathcal{G}_{1,k}, \cdots, \mathcal{G}_{h,k}, \cdots \mathcal{G}_{H,k}, \cdots, \mathcal{G}_{1,k+1}, \cdots, \mathcal{G}_{H,K}, \mathcal{G}_{1,K+1}\},$$

and we will use $\mathcal{G}$ as the filtration set for all of the proofs in the following section and it is obvious that $\mathcal{G}_{1,K+1}$ contains all information we collect during the exploration phase.

B.2 Proof of Lemma A.1

Proof of Lemma A.1. We prove this lemma by induction on time step h. Indeed, when $h = H+1$, $V_{H+1}(s) = V_{H+1}^*(s;r) = 0$ by definition. Suppose for $h \in [H]$, $V_{h+1}(s) \geq V_{h+1}^*(s;r)$, then following the update rule of Q function in Algorithm 1, we have

$$\begin{aligned}
&Q_h(s,a) - Q_h^*(s,a;r) \\
&\quad = \min\big\{H, r_h(s,a) + \langle \boldsymbol{\psi}_{V_{h+1}}(s,a), \boldsymbol{\theta}\rangle + \beta\|\boldsymbol{\psi}_{V_{h+1}}(s,a)\|_{\boldsymbol{\Sigma}^{-1}}\big\} - r_h(s,a) - [\mathbb{P}V_{h+1}^*](s,a;r) \\
&\quad \geq \min\big\{H - Q_h^*(s,a;r), \langle \boldsymbol{\psi}_{V_{h+1}}(s,a), \boldsymbol{\theta}\rangle + \beta\|\boldsymbol{\psi}_{V_{h+1}}(s,a)\|_{\boldsymbol{\Sigma}^{-1}} - [\mathbb{P}V_{h+1}^*](s,a;r)\big\}.
\end{aligned}$$

We need to show that $Q_h(s,a) \geq Q_h^*(s,a;r)$. Since it is obvious that the first term $H - Q_h^*(s,a;r)$ in $\min$ operator is greater than zero, we only need to verify that the second term is also positive where

$$\begin{aligned}
&\langle \boldsymbol{\psi}_{V_{h+1}}(s,a), \boldsymbol{\theta}\rangle + \beta\|\boldsymbol{\psi}_{V_{h+1}}(s,a)\|_{\boldsymbol{\Sigma}^{-1}} - [\mathbb{P}V_{h+1}^*](s,a;r) \\
&\quad \geq \langle \boldsymbol{\psi}_{V_{h+1}}(s,a), \boldsymbol{\theta}\rangle + \beta\|\boldsymbol{\psi}_{V_{h+1}}(s,a)\|_{\boldsymbol{\Sigma}^{-1}} - [\mathbb{P}V_{h+1}](s,a;r) \\
&\quad = \langle \boldsymbol{\psi}_{V_{h+1}}(s,a), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle + \beta\|\boldsymbol{\psi}_{V_{h+1}}(s,a)\|_{\boldsymbol{\Sigma}^{-1}} \\
&\quad \geq \beta\|\boldsymbol{\psi}_{V_{h+1}}(s,a)\|_{\boldsymbol{\Sigma}^{-1}} - \|\boldsymbol{\psi}_{V_{h+1}}(s,a)\|_{\boldsymbol{\Sigma}^{-1}}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}},
\end{aligned}$$

where the first inequality is from the induction assumption that $V_{h+1}^*(s;r) \leq V_{h+1}(s)$. The second equality is from the expectation of value function is a linear function of $\boldsymbol{\psi}_{V_{h+1}}$ shown in (3.2). Then the inequality on the third line is utilizing the fact that $\langle \mathbf{x}, \mathbf{y}\rangle \geq -\|\mathbf{x}\|_{\mathbf{A}^{-1}}\|\mathbf{y}\|_{\mathbf{A}}$. Since it is guaranteed that $\beta \geq \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}$ from the statement of this lemma, $Q_h(s,a) - Q_h^*(s,a;r) \geq 0$, which from induction we get our conclusion.

For the second part controlling $V_1(s) - V_1^\pi(s)$, since aforementioned proof has shown that $V_h^*(s;r) \leq V_h(s)$ for all $h \in [H]$, we have $V_h^*(s;r) - V_h^\pi(s;r) \leq V_h(s) - V_h^\pi(s;r)$ and

$$\begin{aligned}
V_h(s) - V_h^\pi(s;r) &= \min\{H, r_h(s,\pi_h(s)) + \langle \boldsymbol{\psi}_{V_{h+1}}, \boldsymbol{\theta}\rangle + \beta\|\boldsymbol{\psi}_{V_{h+1}}(s,\pi_h(s))\|_{\boldsymbol{\Sigma}^{-1}}\} \\
&\quad - r_h(s,\pi_h(s)) - [\mathbb{P}V_{h+1}^\pi](s,\pi_h(s);r) \\
&\leq \min\{H, \langle \boldsymbol{\psi}_{V_{h+1}}, \boldsymbol{\theta}\rangle + \beta\|\boldsymbol{\psi}_{V_{h+1}}(s,\pi_h(s))\|_{\boldsymbol{\Sigma}^{-1}} - [\mathbb{P}V_{h+1}](s,\pi_h(s))\} \\
&\quad + [\mathbb{P}V_{h+1}](s,\pi_h(s))\} - [\mathbb{P}V_{h+1}^\pi](s,\pi_h;r) \\
&= \min\{H, \langle \boldsymbol{\psi}_{V_{h+1}}, \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle + \beta\|\boldsymbol{\psi}_{V_{h+1}}(s,\pi_h(s))\|_{\boldsymbol{\Sigma}^{-1}}\} \\
&\quad + [\mathbb{P}V_{h+1}](s,\pi_h(s))\} - [\mathbb{P}V_{h+1}^\pi](s,\pi_h(s);r) \\
&\leq \min\{H, 2\beta\|\boldsymbol{\psi}_{V_{h+1}}(s,\pi_h(s))\|_{\boldsymbol{\Sigma}^{-1}}\} \\
&\quad + [\mathbb{P}V_{h+1}](s,\pi_h(s))\} - [\mathbb{P}V_{h+1}^\pi](s,\pi_h(s);r),
\end{aligned}$$

where the first inequality is directly from moving term $-r_h(s,\pi_h(s)) - [\mathbb{P}V_{h+1}](s,\pi_h(s))$ into the $\min$ operator, the second inequality uses the condition that $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}} \le \beta$ and $\langle \mathbf{x},\mathbf{y}\rangle \le \|\mathbf{x}\|_{\mathbf{A}^{-1}}\|\mathbf{y}\|_{\mathbf{A}}$. Considering the first step $h=1$, we have

$$
\begin{aligned}
V_1(s_1) - V_1^{\pi}(s_1;r) &\le \min\{H, 2\beta\|\boldsymbol{\psi}_{V_2}(s_1,\pi_1(s_1))\|_{\boldsymbol{\Sigma}^{-1}}\} + \mathbb{E}_{s_2\sim\mathbb{P}(\cdot|s_1,\pi_1(s_1))}[V_2(s_2)-V_2^{\pi}(s_2)] \\
&\le \min\{H, 2\beta\|\boldsymbol{\psi}_{V_2}(s_1,\pi_1(s_1))\|_{\boldsymbol{\Sigma}^{-1}}\} \\
&\quad + \mathbb{E}_{s_2\sim\mathbb{P}(\cdot|s_1,\pi_1(s_1))}\Big[\min\{H, 2\beta\|\boldsymbol{\psi}_{V_3}(s_2,\pi_2(s_2))\|_{\boldsymbol{\Sigma}^{-1}}\} \\
&\quad + \mathbb{E}_{s_3\sim\mathbb{P}(\cdot|s_2,\pi_2(s_2))}[V_3(s_3)-V_3^{\pi}(s_3)]\Big] \\
&\le \cdots \\
&\le \mathbb{E}\Big[\sum_{h=1}^{H}\min\{H, 2\beta\|\boldsymbol{\psi}_{V_{h+1}}(s_h,\pi_h(s_h))\|_{\boldsymbol{\Sigma}^{-1}}\}\Big| s_1,\pi\Big],
\end{aligned}
$$

which concludes our proof. □

B.3 Proof of Lemma A.2

We introduce the classical confidence set lemma from [1].

Lemma B.1 (Theorem 2, [1])**.** Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration and $\{\eta_t\}$ is a real-valued stochastic process which is F_t-measurable and conditionally R-sub-Gaussian. Set $y_t = \langle \mathbf{x}_t, \boldsymbol{\psi}^*\rangle + \eta_t$, $\mathbf{V}_t = \lambda\mathbf{I} + \sum_{i=1}^{t}\mathbf{x}_i\mathbf{x}_i^{\top}$ where $\mathbf{x}\in\mathbb{R}^d$. Denote the estimation of $\boldsymbol{\psi}^*$ as $\boldsymbol{\psi}_t = \mathbf{V}_t^{-1}\sum_{i=1}^{t} y_i\mathbf{x}_i$. If $\|\boldsymbol{\psi}^*\|_2 \le S$, $\|\mathbf{x}_t\|_2 \le L$, then with probability at least $1-\delta$, for all $t\ge 0$

$$
\|\boldsymbol{\psi}^* - \boldsymbol{\psi}_t\|_{\mathbf{V}_t} \le R\sqrt{d\log\left(\frac{1+tL^2/\lambda}{\delta}\right)} + S\sqrt{\lambda}.
$$

Equipped with this lemma, we begin our proof.

Proof of Lemma A.2. Since $[\mathbb{P}u_h^k](s_h^k,a_h^k) = \langle \boldsymbol{\psi}_{u_h^k}(s_h^k,a_h^k), \boldsymbol{\theta}^*\rangle$ due to (3.2) and $u_h^k(s) \le H$, $u_h^k(s) - \langle \boldsymbol{\psi}_{u_h^k}(s_h^k,a_h^k), \boldsymbol{\theta}^*\rangle$ is $\mathcal{G}_{h,k}$-measurable and it is also a zero mean H-sub-Gaussian random variable conditioned on $\mathcal{G}_{h,k}$. Also from Definition 3.1, $\|\boldsymbol{\theta}^*\|_2 \le B$, $\|\boldsymbol{\psi}_{u_h^k}(s_h^k,a_h^k)\|_2 \le H$. Therefore, recall the calculation of $\boldsymbol{\theta}_k$, according to Lemma B.1, let $t=(k-1)H$ we have

$$
\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_{1,k}} \le H\sqrt{d\log\left(\frac{1+(k-1)H^3/\lambda}{\delta}\right)} + B\sqrt{\lambda}.
$$

Let $\lambda = B^{-2}$, $\delta = \delta/3$ and relax k with $k=K+1$, we can get the β claimed in Theorem 4.2. □

B.4 Proof of Lemma A.3

We provide the proof to control the summation of the value function during the exploration phase. To start with, since rather than immediately updating the parameter after each time step, we can only update the estimation $\boldsymbol{\theta}$ and its 'covariance matrix' $\boldsymbol{\Sigma}$ once after each episode. As a result, this 'batched update rule' make the UCB bonus term at step (h,k) be $\|\boldsymbol{\psi}_{u_h^k}(s_h^k,a_h^k)\|_{\mathbf{U}_{1,k}^{-1}}$ instead of $\|\boldsymbol{\psi}_{u_h^k}(s_h^k,a_h^k)\|_{\mathbf{U}_{h,k}^{-1}}$ in the vanilla linear bandit setting. Therefore, we need lemmas showing that these two UCB terms are close to each other.

Lemma B.2. For any $\{\mathbf{x}_{h,k}\}_{h=1,k=1}^{H,K} \subset \mathbb{R}^d$ satisfying that $\|\mathbf{x}_{h,k}\|_2 \le L, \forall (h,k)\in[H]\times[K]$, let $\mathbf{U}_{h,k} = \lambda\mathbf{I} + \sum_{\kappa=1}^{k-1}\sum_{i=1}^{H}\mathbf{x}_{i,\kappa}\mathbf{x}_{i,\kappa}^{\top} + \sum_{i=1}^{h-1}\mathbf{x}_{i,k}\mathbf{x}_{i,k}^{\top}$, there exists at most $2Hd\log(1+KHL^2/\lambda)$ pairs of (h,k) tuple such that $\det\mathbf{U}_{h,k} \le 2\det\mathbf{U}_{1,k}$.

Lemma B.3 (Lemma 12, [1])**.** Suppose $\mathbf{A},\mathbf{B}\in\mathbb{R}^{d\times d}$ are two positive definite matrices satisfying that $\mathbf{A}\succeq\mathbf{B}$, then for any $\mathbf{x}\in\mathbb{R}^d$, we have $\|\mathbf{x}\|_{\mathbf{A}} \le \|\mathbf{x}\|_{\mathbf{B}}\sqrt{\det(\mathbf{A})/\det(\mathbf{B})}$.

Following that, we also need to introduce the classical lemma to control the summation of the UCB bonus terms in vanilla linear bandit setting.

Lemma B.4 (Lemma 11, [1]). For any $\{\mathbf{x}_t\}_{t=1}^T \subset \mathbb{R}^d$ satisfying that $\|\mathbf{x}_t\|_2 \leq L, \forall t \in [T]$, let $\mathbf{U}_t = \lambda \mathbf{I} + \sum_{\tau=1}^{t-1} \mathbf{x}_\tau \mathbf{x}_\tau^\top$, we have

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

We also need to introduce the Azuma-Hoeffding's inequality to build the concentration bound for martingale difference sequences.

Lemma B.5 (Azuma-Hoeffding's inequality, [4]). Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}_{i=1}^n$ (i.e. $\mathbb{E}[x_i|\mathcal{G}_i] = 0$ a.s. and x_i is $\mathcal{G}_{i+1}$ measurable) such that $|x_i| \leq M$ a.s.. Then for any $0 < \delta < 1$, with probability at least $1-\delta$, $\sum_{i=1}^n x_i \leq M\sqrt{2n\log(1/\delta)}$.

Proof of Lemma A.3. By Lemma A.1, for the k-th episode, we have

$$\begin{aligned} V_1^k(s_1^k) - V^{\pi^k}(s_1^k) &= \mathbb{E}\left[\sum_{h=1}^H \min\{H, 2\beta\|\boldsymbol{\psi}_{V_{h+1}^k}(s_h, \pi_h^k(s_h))\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}\}\middle| s_1^k, \pi^k\right] \\ &\leq \mathbb{E}\left[\sum_{h=1}^H \min\{H, 2\beta\|\boldsymbol{\psi}_{u_h^k}(s_h, \pi_h^k(s_h))\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}\}\middle| s_1^k, \pi^k\right] \end{aligned} \tag{B.1}$$

where the inequality comes from that the pseudo value function u_h^k defined in (4.3) is from maximizing the UCB term $\|\boldsymbol{\psi}_{V_{h+1}^k}(s_h, \pi_h^k(s_h))\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}$ and we denote $\{\pi_h^k\}_{h=1}^H$ by π^k in short. By the definition of r_h^k, we have

$$\begin{aligned} V^{\pi^k}(s_1^k) &= \mathbb{E}[\sum_{h=1}^H r_h^k(s_h, \pi_h^k(s_h))|s_1^k, \pi^k] \\ &= \mathbb{E}\left[\sum_{h=1}^H \min\{1, 2\beta\|\boldsymbol{\psi}_{u_h^k}(s_h, \pi_h^k(s_h))\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}/H\}\middle| s_1^k, \pi^k\right]. \end{aligned} \tag{B.2}$$

Adding (B.1) and (B.2) together and taking summation over k, we have

$$\sum_{k=1}^K V_1^k(s_1^k) \leq \frac{H+1}{H} \underbrace{\sum_{k=1}^K \mathbb{E}\left[\sum_{h=1}^H \min\{H, 2\beta\|\boldsymbol{\psi}_{u_h^k}(s_h, \pi_h^k(s_h))\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}\}\middle| s_1^k, \pi^k\right]}_{I_1} \leq 2I_1, \tag{B.3}$$

where the last inequality is due to $(H+1)/H \leq 2$. Next we are going to control the expectation of summation I_1. Consider the filtration $\{\mathcal{G}_{h,k}\}_{h=1,k=1}^{H,K}$ defined in Section B.1, denote $x_{h,k}$ as follows:

$$x_{h,k} = \min\{H, 2\beta\|\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}\} - \mathbb{E}_{s_h}\left[\min\{H, 2\beta\|\boldsymbol{\psi}_{u_h^k}(s_h, \pi_h^k(s_h))\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}\}\right],$$

then $x_{h,k}$ is obviously a martingale difference sequence bounded by H w.r.t. $\{\mathcal{G}_{h,k}\}_{h=1,k=1}^{H,K}$. Thus by Azuma-Hoeffding's inequality in Lemma B.5, we have with probability at least $1-\delta$, $\sum_{k=1}^K \sum_{h=1}^H x_h \leq H\sqrt{2HK\log(1/\delta)}$. Therefore,

$$\begin{aligned} I_1 &= \sum_{k=1}^K \sum_{h=1}^H \min\{H, 2\beta\|\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}\} + \sum_{k=1}^K \sum_{h=1}^H x_h \\ &\leq 2\beta \sum_{k=1}^K \sum_{h=1}^H \min\{1, \|\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}\} + H\sqrt{2HK\log(1/\delta)} \\ &\leq 2\sqrt{2}\beta \underbrace{\sum_{k=1}^K \sum_{h=1}^H \min\{1, \|\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{h,k}^{-1}}\}}_{I_2} + 4\beta H d\log(1 + KH^3/\lambda) + H\sqrt{2HK\log(1/\delta)}, \end{aligned}$$

where the inequality on the second line is due to $2\beta \geq 2H\sqrt{d\log 3} \geq H$ and the last inequality uses Lemma B.3 with $\mathbf{\Sigma}_{1,k}^{-1} \succeq \mathbf{\Sigma}_{h,k}^{-1}$ and $\det \mathbf{\Sigma}_{1,k}^{-1} \leq 2\det \mathbf{\Sigma}_{1,k}^{-1}$ expect for $\widetilde{\mathcal{O}}(Hd)$ cases by Lemma B.2. By $\min\{1, \|\boldsymbol{\psi}_{u_h^k}(s_h, \pi_h^k(s_h))\|_{\mathbf{\Sigma}_{h,k}^{-1}}\} \leq 1$ and $\|\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\|_2 \leq H$ since $u_h^k \leq H$, we can further bound the $\widetilde{\mathcal{O}}(Hd)$ terms where $\det \mathbf{\Sigma}_{1,k}^{-1} > 2\det \mathbf{\Sigma}_{1,k}^{-1}$. To bound I_2, by Lemma B.4, using Cauchy-Schwarz inequality we have

$$I_2 \leq \sqrt{KH}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1, \|\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\|^2_{\mathbf{\Sigma}_{h,k}^{-1}}\}} \leq \sqrt{2KHd\log(1+KH^3/(d\lambda))},$$

Plugging I_2 into I_1 then plugging I_1 into (B.3). Let $\lambda = B^{-2}$, the summation of the value function $V_1^k(s_1^k)$ is bounded by

$$\sum_{k=1}^{K} V_1^k(s_1^k) \leq 8\beta\Big(\sqrt{HKd\log(1+KH^3B^2/d)} + dH\log(1+KH^3B^2)\Big) + 2H\sqrt{2HK\log(1/\delta)}.$$

Taking $\delta = \delta/3$, we can finalize the proof of Lemma A.3. □

B.5 Proof of Lemma A.4

The proof of this lemma is similar to the proof of Lemma 5.2 in [28]. We extend their proof to a *time varying* reward and *homogeneous* setting, where the rewards (i.e., the exploration-driven reward function r_h^k) are different in different episode k. To prove this lemma, we need to introduce the Bernstein inequality for vector-valued martingales.

Lemma B.6 (Theorem 4.1, [28])**.** Let $\{\mathcal{G}_t\}_{t=1}^{\infty}$ be a filtration, $\{\mathbf{x}_t, \eta_t\}_{t\geq 1}$ a stochastic process so that $\mathbf{x}_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable and η_t is $\mathcal{G}_{t+1}$-measurable. Fix $R, L, \sigma, \lambda > 0$, $\boldsymbol{\mu}^* \in \mathbb{R}^d$. For $t \geq 1$, let $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t\rangle + \eta_t$. Suppose $\eta_t, \mathbf{x}_t$ satisfy

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

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

$$\forall t > 0,\ \Big\|\sum_{\tau=1}^{t}\mathbf{x}_\tau\eta_\tau\Big\|_{\mathbf{U}_\tau^{-1}} \leq \beta_t,\ \|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{U}_t} \leq \beta_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2,$$

where $\boldsymbol{\mu}_t = \mathbf{U}_t^{-1}\mathbf{b}_t, \mathbf{U}_t = \lambda\mathbf{I} + \sum_{\tau=1}^{t}\mathbf{x}_\tau\mathbf{x}_\tau^\top, \mathbf{b}_t = \sum_{\tau=i}^{t} y_\tau\mathbf{x}_\tau$, and

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

We also introduce the following lemma to analyze the error between the estimated variance $\bar{\mathbb{V}}_h^k$ and the true variance $\mathbb{V}_h^k$.

Lemma B.7 (Lemma C.1, [28])**.** Let $\mathbb{V}_h^k(s,a)$ be as defined in (3.1) and $\bar{\mathbb{V}}_h^k(s,a)$ be as defined in (5.3), then

$$|\mathbb{V}_h^k(s,a) - \bar{\mathbb{V}}_h^k(s,a)| \leq \min\Big\{H^2, \|\boldsymbol{\psi}_{[V_{h+1}^k]^2}(s,a)\|_{\widetilde{\mathbf{\Sigma}}_{1,k}^{-1}}\|\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\|_{\widetilde{\mathbf{\Sigma}}_{1,k}}\Big\} + \min\Big\{H^2, 2H\|\boldsymbol{\psi}_{V_{h+1}^k}(s,a)\|_{\widehat{\mathbf{\Sigma}}_{1,k}^{-1}}\|\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\|_{\widehat{\mathbf{\Sigma}}_{1,k}}\Big\}.$$

Equipped with these lemmas, we can start the proof of Lemma A.4.

Proof of Lemma A.4. Recall the regression in (5.4). For the regression on $\widehat{\mathbf{\Sigma}}, \widehat{\boldsymbol{\theta}}$, let $\mathbf{x}_h^k = \boldsymbol{\psi}_{V_{h+1}^k}(s_h^k, a_h^k)/\bar{\sigma}_h^k$, and $\eta_h^k = V_{h+1}^k(s_{h+1}^k)/\bar{\sigma}_h^k - \langle\boldsymbol{\theta}^*, \mathbf{x}_h^k\rangle$. Since $\bar{\sigma}_h^k \geq H/\sqrt{d}$ defined in (5.2), we get $\|\mathbf{x}_h^k\|_2 \leq \sqrt{d}, |\eta_h^k| \leq \sqrt{d}$, thus one could verify that $\mathbb{E}[[\eta_h^k]^2|\mathcal{G}_{h,k}] \leq d$, $\mathbb{E}[\eta_h^k|\mathcal{G}_{h,k}] = 0$, from Lemma B.6, taking $t = (k-1)H$ we have

$$\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\mathbf{\Sigma}}_{1,k}} \leq 8d\sqrt{\log(1+(k-1)H/\lambda)\log(4(k-1)^2H^2/\delta)} + 4\sqrt{d}\log(4(k-1)^2H^2/\delta) + \sqrt{\lambda}B.$$

For the regression of $\widetilde{\boldsymbol{\Sigma}}, \widetilde{\boldsymbol{\theta}}$, $\mathbf{x}_h^k = \boldsymbol{\psi}_{[V_{h+1}^k]^2}(s_h^k, a_h^k)$ which directly implies $\|\mathbf{x}_h^k\|_2 \leq H^2$. Let $\eta_h^k = V_{h+1}^k(s_{h+1}^k)^2 - \langle \boldsymbol{\theta}^*, \mathbf{x}_h^k \rangle$, one can easily verify that $|\eta_h^k| \leq H^2$ and $\mathbb{E}[\eta_h^k | \mathcal{G}_{h,k}] = 0, \mathbb{E}[[\eta_h^k]^2 | \mathcal{G}_{h,k}] \leq H^4$, thus using Lemma B.6 again we have

$$\|\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}_k\|_{\widetilde{\boldsymbol{\Sigma}}_{1,k}} \leq 8H^2\sqrt{d\log(1+(k-1)H/\lambda)\log(4(k-1)^2H^2/\delta)} + 4H^2\log(4(k-1)^2H^2/\delta) + \sqrt{\lambda}B.$$

Since $\lambda = B^{-2}$, if we select $\check{\beta}$ and $\widetilde{\beta}$ as

$$\check{\beta} = 8d\sqrt{\log(1+KHB^2/)\log(4K^2H^2/\delta)} + 4\sqrt{d}\log(4(k-1)^2H^2/\delta) + 1$$
$$\widetilde{\beta} = 8H^2\sqrt{d\log(1+KHB^2)\log(4K^2H^2/\delta)} + 4H^2\log(4K^2H^2/\delta) + 1,$$

then with probability at least $1-2\delta$, for all $k \in [K+1]$, $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}} \leq \check{\beta}$, $\|\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}_k\|_{\widetilde{\boldsymbol{\Sigma}}_{1,k}} \leq \widetilde{\beta}$.

Next we are going to give the choice of $\widehat{\beta}$ to make sure that $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}} \leq \check{\beta}$ holds with high probability. The following proof is conditioned on that the aforementioned event $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}} \leq \check{\beta}$, $\|\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}_k\|_{\widetilde{\boldsymbol{\Sigma}}_{1,k}} \leq \widetilde{\beta}$ holds, then from Lemma B.7 we have

$$\begin{aligned}
&|\mathbb{V}_h^k(s,a) - \bar{\mathbb{V}}_h^k(s,a)| \\
&\quad \leq \min\left\{H^2, \widetilde{\beta}\|\boldsymbol{\psi}_{[V_{h+1}^k]^2}(s,a)\|_{\widetilde{\boldsymbol{\Sigma}}_{1,k}^{-1}}\right\} + \min\left\{H^2, 2\check{\beta}H\|\boldsymbol{\psi}_{V_{h+1}^k}(s,a)\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}^{-1}}\right\} \\
&\quad = E_h^k(s,a)
\end{aligned} \tag{B.4}$$

Again, let $\mathbf{x}_h^k = \boldsymbol{\psi}_{V_{h+1}^k}(s_h^k, a_h^k)/\bar{\sigma}_h^k$ to denote the context vector and $\eta_h^k = V_{h+1}^k(s_{h+1}^k)/\bar{\sigma}_h^k - \langle \boldsymbol{\theta}^*, \mathbf{x}_h^k \rangle$ to denote the noise term, since $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}} \leq \check{\beta}$, we have

$$\mathbb{E}[[\eta_h^k]^2 | \mathcal{G}_{h,k}] = \mathbb{V}_h^k(s_h^k, a_h^k)/\nu_h^k \leq (E_h^k(s_h^k, a_h^k) + \bar{\mathbb{V}}_h^k(s_h^k, a_h^k))/\nu_h^k \leq 1,$$

where the first inequality is from (B.4), the second inequality holds because the definition of ν_h^k in (5.2).

Therefore we have verified that the noise term η_h^k is a zero-mean random variable conditioned on $\mathcal{G}_{h,k}$ and $\mathbb{E}[[\eta_h^k]^2 | \mathcal{G}_{h,k}] \leq 1$. In that case, using Lemma B.6 again we could get with probability at least $1-\delta$,

$$\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}} \leq 8\sqrt{d(1+(k-1)H/\lambda)\log(4(k-1)^2H^2/\delta)} \tag{B.5}$$
$$+ 4\sqrt{d}\log(4(k-1)^2H^2/\delta) + \sqrt{\lambda}B, \tag{B.6}$$

again, since $\lambda = B^{-2}$, if we select $\widehat{\beta}$ as

$$\widehat{\beta} = 8\sqrt{d(1+KHB^2)\log(4K^2H^2/\delta)} + 4\sqrt{d}\log(4K^2H^2/\delta) + 1,$$

then $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}} \leq \widehat{\beta}$ with probability at least $1-\delta$ for all $k \in [K+1]$.

Next, for the regression of $\boldsymbol{\theta}_{K+1}, \boldsymbol{\Sigma}_{1,K+1}$, by Lemma A.2, we obtain the same result with the selection of β as

$$\beta = H\sqrt{d\log\left(\frac{1+KH^3/\lambda}{\delta}\right)} + B\sqrt{\lambda},$$

which suggests that with probability at least $1-\delta$, $\|\boldsymbol{\theta}_{K+1} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_{1,K+1}} \leq \beta$. Then taking union bound with all aforementioned event $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}} \leq \check{\beta}$, $\|\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}_k\|_{\widetilde{\boldsymbol{\Sigma}}_{1,k}} \leq \widetilde{\beta}$, $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_k\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}} \leq \widehat{\beta}$, we have all these events mentioned in this proof holds with probability at least $1-4\delta$. Replace δ with $\delta/12$, we obtain our final results.

Next, for the regression of $\boldsymbol{\theta}_{K+1}, \boldsymbol{\Sigma}_{1,K+1}$, by Lemma A.2, we obtain the same result with the selection of β as

$$\beta = H\sqrt{d\log\left(\frac{1+KH^3/\lambda}{\delta}\right)} + B\sqrt{\lambda},$$

which suggests that with probability at least $1-\delta$, $\|\boldsymbol{\theta}_{K+1}-\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_{1,K+1}} \leq \beta$. Again, taking an additional union bound, with probability at least $1-4\delta$, all events mentioned in this proof hold. Replace δ with $\delta/12$, we obtain our final results. □

B.6 Proof of Lemma A.5

The proof of this lemma borrows some intuition from the proof of Theorem 5.3 in [28]. Unlike Zhou et al. [28] that deals the fixed reward and time-inhomogeneous setting, we need to extend their proof in order to deal with the time-varying reward and time-homogeneous setting.

The next lemmas shows the relationship between the summation of ν_h^k and the difference between $V_h^k(s)$ calculated in Algorithm 3 and $V_h^{\pi^k}(s;\{r_h^k\}_{h=1,k=1}^{H,K})$

Lemma B.8. Let V_h^k, ν_h^k be defined in Algorithm 3. Then if the condition in Lemma A.4 holds, the following inequality holds with probability at least $1-2\delta$,

$$\begin{aligned}\sum_{k=1}^{K}[V_1^k(s_1^k)-V_1^{\pi^k}(s_1^k)] &\leq 4\sqrt{d}\widehat{\beta}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k}\sqrt{\log(1+KHB^2)} \\ &\quad + 2H^2 d\log(1+KHB^2 d) + H\sqrt{2KH\log(1/\delta)} \\ \sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi^k})](s_h^k,a_h^k) &\leq 4\sqrt{d}H\widehat{\beta}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k}\sqrt{\log(1+KHB^2)} \\ &\quad + 2H^3 d\log(1+KHB^2 d) + 2H^2\sqrt{2KH\log(1/\delta)},\end{aligned}$$

Lemma B.9. Let V_h^k, ν_h^k be defined in Algorithm 3. Then if the condition in Lemma A.4 holds, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k \leq \frac{H^3K}{d} + 3H^2K + 3H^3\log(1/\delta) + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi^k}](s_h^k,a_h^k) \\ + 2\widetilde{\beta}\sqrt{KHd\log(1+KH^5B^2/d)} + 4\widetilde{\beta}Hd\log(1+KH^5B^2/d) \\ + 8H^2\check{\beta}\sqrt{KHd\log(1+KHB^2)} + 8H^3d\check{\beta}\log(1+KHdB^2).\end{aligned}$$

Equipped with these two lemmas, we can start to prove Lemma A.5.

Proof of Lemma A.5. In this proof, we use $\widetilde{\mathcal{O}}(\cdot)$ to ignore all constant and log terms to simplify the results. Recall the selection of $\beta, \widehat{\beta}, \check{\beta}, \widetilde{\beta}$, we have $\beta = \widetilde{\mathcal{O}}(H\sqrt{d})$, $\widehat{\beta} = \widetilde{\mathcal{O}}(\sqrt{d})$, $\check{\beta} = \widetilde{\mathcal{O}}(d)$, $\widetilde{\beta} = \widetilde{\mathcal{O}}(H^2\sqrt{d})$. Therefore Lemma B.8 could be simplified as

$$\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi^k})](s_h^k,a_h^k) \leq \widetilde{\mathcal{O}}\left(Hd\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k} + H^3d + \sqrt{KH^5}\right). \tag{B.7}$$

Lemma B.9 could also be simplified as

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k \leq \widetilde{\mathcal{O}}\left(\frac{H^3K}{d} + H^2K + H\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi^k})](s_h^k,a_h^k) + \sqrt{KH^5d^3} + H^3d^2\right). \tag{B.8}$$

Let $\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k} = x$, plugging (B.7) into (B.8), we have

$$x^2 \leq \widetilde{\mathcal{O}}(H^3Kd^{-1} + H^2K + H^2dx + H^4d + \sqrt{KH^7} + \sqrt{KH^5d^3} + H^3d^2),$$

Since the quadratic inequality $x^2 \leq \widetilde{\mathcal{O}}(bx+c)$ indicates that $x \leq \mathcal{O}(b+\sqrt{c})$, setting

$$b = \widetilde{\mathcal{O}}(H^2d), c = \widetilde{\mathcal{O}}(H^3Kd^{-1} + H^2K + H^4d + \sqrt{KH^7} + \sqrt{KH^5d^3} + H^3d^2),$$

hence

$$\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k} \leq \widetilde{\mathcal{O}}(H^2 d + \sqrt{H^3K/d} + H\sqrt{K} + H^2\sqrt{d} + d\sqrt{H^3} + (KH^7)^{1/4} + (KH^5d^3)^{1/4}) \tag{B.9}$$

$$= \widetilde{\mathcal{O}}(\sqrt{H^3K/d} + H\sqrt{K}) + o(\sqrt{K}). \tag{B.10}$$

Plugging (B.10) back to Lemma B.8, we have

$$\sum_{k=1}^{K}[V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)] \leq \widetilde{\mathcal{O}}(\sqrt{H^3Kd} + Hd\sqrt{K}) + o(\sqrt{K}). \tag{B.11}$$

Next we are going to show the bound of the summation over $V_1^{\pi^k}(s_1^k)$, note that this value function is bounded by H and from Bellman equality, we have

$$V_h^{\pi^k}(s_1^k) = r_h^k(s_1^k, a_1^k) + [\mathbb{P}V_{h+1}^{\pi^k}](s_h^k, a_h^k),$$

taking summation over $h \in [H], k \in [K]$ then

$$\begin{aligned}\sum_{k=1}^{K} V_1^{\pi^k}(s_1^k) &= \sum_{k=1}^{K}\sum_{h=1}^{H} r_h^k(s_1^k, a_1^k) + \sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}V_{h+1}^{\pi^k}](s_h^k, a_h^k) - V_{h+1}^{\pi^k}(s_{h+1}^k) \\ &\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1, 2\beta\|\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}/H\} + H\sqrt{HK\log(1/\delta)},\end{aligned}$$

where the last inequality holds due to Azuma-Hoeffding's inequality i.e. Lemma B.5. For the first term,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1, 2\beta\|\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}/H\} \leq \frac{2\beta}{H}\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1, \|\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}\}}_{I_1},$$

where the inequality is due to $\beta \geq H\sqrt{\log(12)} \geq H/2$. Using Lemma B.2 and Lemma B.3 with $\boldsymbol{\Sigma}_{1,k}^{-1} \succeq \boldsymbol{\Sigma}_{h,k}^{-1}$ and $\det \boldsymbol{\Sigma}_{1,k}^{-1} \leq 2\det \boldsymbol{\Sigma}_{h,k}^{-1}$ except for $\widetilde{\mathcal{O}}(Hd)$ steps mentioned in Lemma B.2, setting $\lambda = B^{-2}$, we have

$$\begin{aligned}I_1 &\leq 2Hd\log(1 + KH^3B^2) + \sqrt{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1, \|\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\|_{\boldsymbol{\Sigma}_{h,k}^{-1}}\} \\ &\leq 2Hd\log(1 + KH^3B^2) + \sqrt{2HK}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1, \|\boldsymbol{\psi}_{u_h^k}(s_h^k, a_h^k)\|^2_{\boldsymbol{\Sigma}_{h,k}^{-1}}\}} \\ &\leq 2Hd\log(1 + KH^3B^2) + 2\sqrt{HKd\log(1 + KH^3B^2/d)}.\end{aligned}$$

Therefore, since $\beta = \widetilde{\mathcal{O}}(H\sqrt{d})$, then

$$\sum_{k=1}^{K} V_1^{\pi^k}(s_1^k) \leq 4\beta d\log(1 + KH^3B^2) + 4\beta\sqrt{Kd\log(1 + KH^3B^2/d)/H} + \sqrt{H^3K\log(1/\delta)} \tag{B.12}$$

$$\leq \widetilde{\mathcal{O}}(d\sqrt{KH} + \sqrt{KH^3}) + o(\sqrt{K}). \tag{B.13}$$

Adding (B.11) and (B.13) together, we have the following result,

$$\sum_{k=1}^{K} V_1^k(s_1^k) \leq \widetilde{\mathcal{O}}(\sqrt{H^3Kd} + Hd\sqrt{K}) + o(\sqrt{K}).$$

By taking the union bound, this inequality holds with probability at least $1 - 4\delta$. Since δ only appears in the logarithmic terms, thus changing δ to $\delta/12$ will not affect the result. □

C Proof of Auxiliary Lemmas in Appendix B

C.1 Proof of Lemma B.2

Proof of Lemma B.2. We want to know how many pairs of (h, k) exists such that $\det \mathbf{U}_{h,k} \geq 2 \det \mathbf{U}_{1,k}$.

Furthermore, we have if there exists $k \in [K]$ such that $\det \mathbf{U}_{1,k+1} \leq 2 \det \mathbf{U}_{1,k}$, then it is obvious that for all $h \in [H]$, we have $\det \mathbf{U}_{h,k} \leq \det \mathbf{U}_{1,k+1} \leq 2 \det \mathbf{U}_{1,k}$.

Therefore, suppose there exists a set $\mathcal{K} \subset [K]$ such that for all $k \notin \mathcal{K}$, $\det \mathbf{U}_{1,k+1} \leq 2 \det \mathbf{U}_{1,k}$ and for all $k \in \mathcal{K}$, $\det \mathbf{U}_{1,k+1} > 2 \det \mathbf{U}_{1,k}$, then the pair of (h, k) such that $\det \mathbf{U}_{h,k} \geq 2 \det \mathbf{U}_{1,k}$ is upper bounded by $H|\mathcal{K}|$.

Notice that for all $k \in \mathcal{K}$, $\det \mathbf{U}_{1,k+1} > 2 \det \mathbf{U}_{1,k}$, it is easy to show that

$$\det \mathbf{U}_{1,K+1} > 2^{|\mathcal{K}|} \det \mathbf{U}_{1,1} = 2^{|\mathcal{K}|} \lambda^d,$$

where the last inequality comes from $\mathbf{U}_{1,1} = \lambda \mathbf{I} \in \mathbb{R}^{d \times d}$. Notice that $\det \mathbf{U} \leq \|\mathbf{U}\|_2^d$, taking log we have

$$d \log(\|\mathbf{U}_{1,K+1}\|_2) \geq \log \det \mathbf{U}_{1,K+1} > |\mathcal{K}| \log 2 + d \log \lambda. \tag{C.1}$$

From the definition of $\mathbf{U}_{1,K+1}$, by triangle inequality,

$$\|\mathbf{U}_{1,K+1}\|_2 \leq \lambda + \sum_{k=1}^{K} \sum_{h=1}^{H} \|\mathbf{x}_h^k \mathbf{x}_h^{k\top}\|_2 \leq \lambda + KH \|\mathbf{x}_h^k\|_2^2 \leq \lambda + KHL^2, \tag{C.2}$$

where the last inequality is due to $\|\mathbf{x}\|_2 \leq L$ from the statement of the lemma. Therefore we conclude our proof by merging (C.1) and (C.2) together to get

$$|\mathcal{K}| \log 2 < d \log(1 + HKL^2/\lambda),$$

noticing $\log 2 \geq 1/2$ we can get the result claimed in the lemma. □

C.2 Proof of Lemma B.8

Proof of Lemma B.8. Assume that the condition in Lemma A.4 holds, then

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

where the first inequality holds due to the definition of V_h^k, the second inequality holds due to Cauchy-Schwarz inequality and the third one holds due to the condition (A.8) in Lemma A.4. Notice that $V_h^k - V_h^{\pi^k} \leq H$, we have

$$V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k) \leq \min\{H, 2\widehat{\beta} \|\boldsymbol{\psi}_{V_{h+1}^k}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}^{-1}}\} + [\mathbb{P}V_{h+1}^k - V_{h+1}^{\pi^k}](s_h^k, a_h^k)$$

Taking summation over $k \in [K]$ and $h \in [H]$, we have

$$\begin{aligned}
\sum_{k=1}^{K} [V_1^k(s_1^k) - V_1^{\pi_k}(s_1^k)] &\leq \sum_{k=1}^{K} \sum_{h=1}^{H} \min\{H, 2\widehat{\beta} \|\boldsymbol{\psi}_{V_{h+1}^k}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}^{-1}}\} \\
&\quad + \sum_{k=1}^{K} \sum_{h=1}^{H} \Big[[\mathbb{P}V_{h+1}^k - V_{h+1}^{\pi^k}](s_h^k, a_h^k) - [V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)] \Big] \\
&\leq \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} \min\{H, 2\widehat{\beta} \|\boldsymbol{\psi}_{V_{h+1}^k}(s_h^k, a_h^k)\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}^{-1}}\}}_{I_1} + H\sqrt{2KH \log(1/\delta)},
\end{aligned} \tag{C.3}$$

where the second inequality is a direct result of Azuma-Hoeffding's inequality as in Lemma B.5.

Next we bound I_1. Recall the update rule of $\widehat{\mathbf{\Sigma}}_{h,k}$, notice that $\bar{\sigma}_h^k \geq H/\sqrt{d}$ and $\|\boldsymbol{\psi}_{V_{h+1}^k}(s_h^K, a_h^K)\|_2 \leq H$ from $V_{h+1}^k \leq H$, it is easy to verify that $\|\boldsymbol{\psi}_{V_{h+1}^k}(s_h^K, a_h^K)/\bar{\sigma}_h^k\|_2 \leq \sqrt{d}$. Hence

$$
\begin{aligned}
I_1 &\leq \sqrt{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{H, 2\widehat{\beta}\|\boldsymbol{\psi}_{V_{h+1}^k}(s_h^k, a_h^k)\|_{\widehat{\mathbf{\Sigma}}_{h,k}^{-1}}\} + 2H^2 d\log(1+KHd/\lambda) \\
&\leq \sqrt{2}\max\{\sqrt{d}, 2\widehat{\beta}\}\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_h^k\min\{1, \|\boldsymbol{\psi}_{V_{h+1}^k}(s_h^k, a_h^k)/\bar{\sigma}_h^k\|_{\widehat{\mathbf{\Sigma}}_{h,k}^{-1}}\} + 2H^2 d\log(1+KHd/\lambda) \\
&\leq 2\sqrt{2}\widehat{\beta}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1, \|\boldsymbol{\psi}_{V_{h+1}^k}(s_h^k, a_h^k)/\bar{\sigma}_h^k\|^2_{\widehat{\mathbf{\Sigma}}_{h,k}^{-1}}\}} + 2H^2 d\log(1+KHd/\lambda) \\
&\leq 4\widehat{\beta}\sqrt{d}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k}\sqrt{\log(1+KH/\lambda)} + 2H^2 d\log(1+KHd/\lambda),
\end{aligned}
$$

where the first inequality, similar to the corresponding proof in Lemma A.3, is a direct implication of Lemma B.2 and Lemma B.3 with $\widehat{\mathbf{\Sigma}}_{1,k}^{-1} \succeq \widehat{\mathbf{\Sigma}}_{h,k}^{-1}$ and $\det \mathbf{\Sigma}_{1,k}^{-1} \leq 2\det \widehat{\mathbf{\Sigma}}_{1,k}^{-1}$ except for $\widetilde{\mathcal{O}}(Hd)$ cases mentioned in Lemma B.2, the second inequality moves $\bar{\sigma}_h^k$ outside, the third inequality holds because $\widehat{\beta} \geq 4\sqrt{d}\log 12 \geq \sqrt{d}$ and Cauchy-Schwarz inequality, and the forth inequality holds due to Lemma B.4. Plugging I_1 into (C.3) and let $h' = 1, \lambda = B^{-2}$, we have

$$
\begin{aligned}
\sum_{k=1}^{K}[V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)] \leq 4\sqrt{d}\widehat{\beta}&\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k}\sqrt{\log(1+KHB^2)} \\
&+ 2H^2 d\log(1+KHB^2 d) + H\sqrt{2KH\log(1/\delta)}.
\end{aligned}
$$

Furthermore, by Azuma-Hoeffding's inequality as in Lemma B.5,

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{P}[V_{h+1}^k - V_{h+1}^{\pi^k}](s_h^k, a_h^k) &= \sum_{k=1}^{K}\sum_{h=2}^{H}[V_h^k - V_h^{\pi^k}](s_h^k) \\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H}\Big[\mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi^k}](s_h^k, a_h^k) - [V_{h+1}^k - V_{h+1}^{\pi^k})](s_{h+1}^k) \\
&\leq 4\sqrt{d}H\widehat{\beta}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k}\sqrt{\log(1+KHB^2)} \\
&\quad + 2H^3 d\log(1+KHB^2 d) + (H+1)H\sqrt{2KH\log(1/\delta)},
\end{aligned}
$$

which becomes the second part of the statement in the lemma. Using $H+1 \leq 2H$ we can get the result claimed in the lemma. □

C.3 Proof of Lemma B.9

To begin with, we will first show the total variance lemma originally introduced in [8].

Lemma C.1 (Total variance lemma, Lemma C.5, [8]). [1] With probability at least $1-\delta$, we have

$$
\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}V_h^{\pi^k}(\cdot; \{r_h^k\}_{h=1}^H)](s,a) \leq 3H^2K + 3H^3\log(1/\delta).
$$

[1] The original Lemma C.5 in Jin et al. [8] holds for the identical reward functions, i.e., $r_h^1 = \cdots = r_h^K$. Their lemma also holds for the general case $r_h^1 \neq \cdots \neq r_h^K$ without changing their proof.

Proof of Lemma B.9. Assume the condition in Lemma A.4 holds, we have with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k &\le \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\frac{H^2}{d}+\bar{\mathbb{V}}_h^k(s_h^k,a_h^k)+E_h^k(s_h^k,a_h^k)\right)\\
&= \frac{H^3K}{d}+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\left([\mathbb{V}_hV_{h+1}^k](s_h^k,a_h^k)-[\mathbb{V}_hV_{h+1}^{\pi^k}](s_h^k,a_h^k)\right)}_{I_1}+2\underbrace{\sum_{k=1}^{H}\sum_{h=1}^{H}E_h^k(s_h^k,a_h^k)}_{I_2}\\
&\quad+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_hV_{h+1}^{\pi^k}](s_h^k,a_h^k)}_{I_3}+\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\bar{\mathbb{V}}_h^k(s_h^k,a_h^k)-[\mathbb{V}_hV_{h+1}^k](s_h^k,a_h^k)-E_h^k\right]}_{I_4}\\
&\le \frac{H^3K}{d}+I_1+I_2+3H^2K+3H^3\log(1/\delta), \qquad \text{(C.4)}
\end{aligned}$$

where the value function $V_h^{\pi^k}(s)$ is short for $V_h^{\pi^k}(s;\{r_h^k\}_{h=1}^H)$ for simplicity. The first inequality is from the definition of ν_h^k in (5.2), while the last inequality is from Lemma C.1 to control I_3. $I_4\le 0$ is due to Lemma A.4. Next we are about to bound I_1 and I_2 separately.

Since the estimated value function V_{h+1}^k and the real value function $V_{h+1}^{\pi^k}$ are both bounded by $[0,H]$, we have

$$I_1\le\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{P}([V_{h+1}^k]^2-[V_{h+1}^{\pi^k}]^2)\right](s_h^k,a_h^k)\le 2H\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}(V_{h+1}^k-V_{h+1}^{\pi^k})](s_h^k,a_h^k).$$

For term I_2, we have

$$\begin{aligned}
I_2 &\le \sum_{k=1}^{K}\sum_{h=1}^{H}\min\{H^2,\widetilde{\beta}\|\boldsymbol{\psi}_{[V_{h+1}^k]^2}(s_h^k,a_h^k)\|_{\widetilde{\boldsymbol{\Sigma}}_{1,k}^{-1}}\}+\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{H^2,2H\check{\beta}\|\boldsymbol{\psi}_{V_{h+1}^k}(s,a)\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}^{-1}}\}\\
&\le \max\{H^2,\widetilde{\beta}\}\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1,\|\boldsymbol{\psi}_{[V_{h+1}^k]^2}(s_h^k,a_h^k)\|_{\widetilde{\boldsymbol{\Sigma}}_{1,k}^{-1}}\}\\
&\quad+\sum_{k=1}^{K}\sum_{h=1}^{H}\max\{H^2,2H\check{\beta}\bar{\sigma}_h^k\}\min\left\{1,\left\|\boldsymbol{\psi}_{V_{h+1}^k}(s,a)/\bar{\sigma}_h^k\right\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}^{-1}}\right\}.
\end{aligned}$$

Noticing that from the definition of ν_h^k,

$$\nu_k^h=\max\{H^2/d,\bar{\mathbb{V}}_h^k(s_h^k,a_h^k)+E_h^k(s_h^k,a_h^k)\}\le\max\{H^2/d,H^2+2H^2\}=3H^2,$$

thus $\bar{\sigma}_h^k=\sqrt{\nu_h^k}\le 2H$. Recall that $\widetilde{\beta}\ge 4H^2\log(12)\ge H^2$ and $\check{\beta}\ge 1$, we have

$$I_2\le\widetilde{\beta}\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1,\|\boldsymbol{\psi}_{[V_{h+1}^k]^2}(s_h^k,a_h^k)\|_{\widetilde{\boldsymbol{\Sigma}}_{1,k}^{-1}}\}}_{I_5}+4H^2\check{\beta}\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{1,\left\|\boldsymbol{\psi}_{V_{h+1}^k}(s,a)/\bar{\sigma}_h^k\right\|_{\widehat{\boldsymbol{\Sigma}}_{1,k}^{-1}}\right\}}_{I_6}.$$

For I_5, using Lemmas B.2 and B.3 with $\widetilde{\boldsymbol{\Sigma}}_{1,k}^{-1}\succeq\widetilde{\boldsymbol{\Sigma}}_{h,k}^{-1}$ and $\det\widetilde{\boldsymbol{\Sigma}}_{1,k}^{-1}\le 2\det\widetilde{\boldsymbol{\Sigma}}_{1,k}^{-1}$ except for $\widetilde{\mathcal{O}}(Hd)$ cases mentioned in Lemma B.2, we have

$$\begin{aligned}
I_5 &\le \sqrt{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1,\|\boldsymbol{\psi}_{[V_{h+1}^k]^2}(s_h^k,a_h^k)\|_{\widetilde{\boldsymbol{\Sigma}}_{h,k}^{-1}}\}+2Hd\log(1+KH^5/d\lambda)\\
&\le \sqrt{2KH}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1,\|\boldsymbol{\psi}_{[V_{h+1}^k]^2}(s_h^k,a_h^k)\|_{\widetilde{\boldsymbol{\Sigma}}_{h,k}^{-1}}^2\}}+2Hd\log(1+KH^5/d\lambda)\\
&\le 2\sqrt{KHd\log(1+KH^5/d\lambda)}+2Hd\log(1+KH^5/d\lambda),
\end{aligned}$$

where the first inequality is a direct implication from Lemma B.2 and the second inequality is due to Cauchy-Schwarz inequality. The third inequality utilizes Lemma B.4. As for I_6, we have

$$
\begin{aligned}
I_6 &\leq \sqrt{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{1, \left\|\boldsymbol{\psi}_{V_{h+1}^k}(s,a)/\bar{\sigma}_h^k\right\|_{\widehat{\boldsymbol{\Sigma}}_{h,k}^{-1}}\right\} + 2Hd\log(1+KHd/\lambda) \\
&\leq \sqrt{2KH}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{1, \left\|\boldsymbol{\psi}_{V_{h+1}^k}(s,a)/\bar{\sigma}_h^k\right\|_{\widehat{\boldsymbol{\Sigma}}_{h,k}^{-1}}\right\}} + 2Hd\log(1+KHd/\lambda) \\
&\leq 2\sqrt{KHd\log(1+KH/\lambda)} + 2Hd\log(1+KHd/\lambda).
\end{aligned}
$$

Finally, plugging I_5, I_6 into I_2 and I_1, I_2 into (C.4) we have

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\nu_h^k \leq \frac{H^3K}{d} + 3H^2K + 3H^3\log(1/\delta) + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi^k}](s_h^k, a_h^k) \\
+ 2\widetilde{\beta}\sqrt{KHd\log(1+KH^5/d\lambda)} + 4\widetilde{\beta}Hd\log(1+KH^5/d\lambda) \\
+ 8H^2\check{\beta}\sqrt{KHd\log(1+KH/\lambda)} + 8H^3d\check{\beta}\log(1+KHd/\lambda).
\end{aligned}
$$

Let $\lambda = B^{-2}$ we could get the result in the statement of the lemma. □

D Proof of Lower Bound

In this section, we will give the detailed proof of the sample complexity lower bound. We start with verifying that the MDP structure as shown in Figure 1 is a linear mixture MDP satisfying Definition 3.1.

D.1 Verification of the MDP structure

We will first show that the ℓ_2 norm of $\boldsymbol{\theta}$ is controlled. Recall the $\boldsymbol{\theta}_i$ is set by $\boldsymbol{\Theta} = \left\{\boldsymbol{\theta}_i | \boldsymbol{\theta}_i = \left(\sqrt{2}, \alpha\widetilde{\boldsymbol{\theta}}_i^\top/\sqrt{d}\right)^\top\right\}$ where $\widetilde{\boldsymbol{\theta}}_i = \mathbf{x}_i \in \mathcal{M}$, we can have that $\|\boldsymbol{\theta}_i\|_2 = \sqrt{2+\alpha^2}$, therefore, as long as the parameter α is an absolute constant, the ℓ_2 norm of $\boldsymbol{\theta}$ is controlled. Next, considering a function $V \leq D$, we have

$$
\|\boldsymbol{\psi}_V(S_1, a_i)\|_2 = \left\|\left(\tfrac{V(S_{2,1})+V(S_{2,2})}{2\sqrt{2}} \quad (V(S_{2,1}) - V(S_{2,2}))\tfrac{\mathbf{a}_i^\top}{\sqrt{2d}}\right)\right\|_2 \leq \sqrt{\frac{D^2}{2} + \frac{D^2 d}{2d}} = D,
$$

which shows that the MDP structure satisfies Definition 3.1.

D.2 Proof of Theorem 6.1

We denote the $d-1$-dimension binary vector set as $\mathcal{B}_{d-1} = \{\mathbf{x} | \mathbf{x} \in \mathbb{R}^{d-1}, [\mathbf{x}]_i \in \{-1, 1\}\}$. The next lemma shows that the binary set $\mathcal{M}$ exists.

Lemma D.1. Given $\gamma \in (0,1)$, there exists a $\mathcal{M} \subset \mathcal{B}_{d-1}$ such that for any two different vector $\mathbf{x}, \mathbf{x}' \in \mathcal{M}$, $\langle \mathbf{x}, \mathbf{x}' \rangle \leq (d-1)\gamma$, and the log-cardinality of the proposed set is bounded as $\log(|\mathcal{M}|) < (d-1)\gamma^2/4$.

With this lemma, we can construct a set $\mathcal{M}$ with $|\mathcal{M}| = \lceil \exp(d\gamma^2/4) \rceil - 1$ where $\gamma = \frac{1}{2}$. It is easy to verify that $|\mathcal{M}| < \exp(d\gamma^2/4)$ and

$$
\begin{aligned}
\log(|\mathcal{M}|) &\geq \log(\exp(d\gamma^2/4) - 1) \\
&\geq \frac{(d-1)\gamma^2}{4} + \log(1 - \exp(-(d-1)\gamma^2/4)) \\
&\geq \frac{(d-1)\gamma^2}{4} - 3, \qquad \text{(D.1)}
\end{aligned}
$$

where the last inequality holds since $\gamma = \frac{1}{2}$ and $d \geq 2$, we have $\log(1 - \exp(-(d-1)\gamma^2/4) \geq -3$. From Lemma D.1, we know that for any two different vectors $\mathbf{x}, \mathbf{x}' \in \mathcal{M}$, $\langle \mathbf{x}, \mathbf{x}' \rangle \leq (d-1)/2$.

Next lemma establishes the lower bound of sample complexity for any algorithm to estimate the true parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ of the proposed linear mixture MDP, from the sampled state-action pairs of this MDP.

Lemma D.2. Suppose an algorithm estimates the underlying parameter $\boldsymbol{\theta}$ by building an estimator $\widehat{\boldsymbol{\theta}}$ from K sampled trajectories. If the algorithm guarantees that $\mathbb{E}_{\boldsymbol{\theta} \sim \mathrm{Unif}_{\boldsymbol{\Theta}}}[\mathbb{1}(\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta})] \geq 1 - \delta$, we have

$$\delta \geq 1 - \left(\frac{d-1}{16} - 3\right)^{-1} \left(\log 2 + \frac{4K\alpha^2}{2 - \alpha^2}\right).$$

Finally, the next lemma suggests that if α is selected properly, then any (ϵ, δ)-reward free algorithm can be converted into an algorithm that provides the exact estimator with a probability of at least $1 - \delta$.

Lemma D.3. Suppose $\alpha \geq \frac{2\sqrt{2}\epsilon}{H-1}$, then any (ϵ, δ)-reward free algorithm could be converted to an algorithm which outputs an estimator $\widehat{\boldsymbol{\theta}}$, satisfying $\mathbb{E}_{\boldsymbol{\theta} \sim \mathrm{Unif}(\boldsymbol{\Theta})}[\mathbb{1}(\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta})] \geq 1 - \delta$.

Equipped with these lemmas, we can provide the proof for Theorem 6.1.

Proof of Theorem 6.1. Set $\alpha = \frac{2\sqrt{2}\epsilon}{H-1}$ and $\epsilon \leq (H-1)/(2\sqrt{2})$, then by Lemma D.3, any (ϵ, δ)-reward free algorithm could be converted to an estimation algorithm with successful rate at least $1 - \delta$. Thus from Lemma D.2, the sample complexity K of these reward free algorithms is bounded by

$$\begin{aligned} K &\geq \frac{2 - \alpha^2}{4\alpha^2}\left(\frac{d-1}{16} - 3\right)(1 - \delta) - \frac{2 - \alpha^2}{4\alpha^2}\log 2 \\ &\geq \frac{(H-1)^2}{128\epsilon^2}\left(\frac{d-1}{16} - 3\right)(1 - \delta) - \frac{(H-1)^2 \log 2}{128\epsilon^2}. \end{aligned}$$

Suppose $H \geq 2, d \geq 50, \delta \geq 1/2$ to simplify the result, we conclude that there exists an absolute positive constant C such that $K \geq C(1 - \delta)H^2 d\epsilon^{-2}$, which leads to our final conclusion. □

E Missing Proofs in Appendix D

We provide detailed proofs for lemmas in Appendix D. For simplicity, we denote by $d' = d - 1$ the dimension of the binary set $\mathcal{M}$.

E.1 Proof of Lemma D.1

Proof of Lemma D.1. To begin with, we assume that $\mathbf{x} \sim \mathrm{Unif}(\mathcal{B}_{d'})$, i.e. $[\mathbf{x}]_i \sim \{-1, 1\}$. Thus given any $\mathbf{x}, \mathbf{x}' \sim \mathrm{Unif}(\mathcal{B}_{d'})$, we have

$$\begin{aligned} \mathbb{P}(\langle \mathbf{x}, \mathbf{x}' \rangle \geq d'\gamma) &= \mathbb{P}_{z_i \sim \mathrm{Unif}\{-1,1\}}\left(\sum_{i=1}^{d'} z_i \geq d'\gamma\right) \\ &= \mathbb{P}_{z_i \sim \mathrm{Unif}\{-1,1\}}\left(\frac{1}{d'}\sum_{i=1}^{d'} z_i \geq \gamma\right) \\ &\leq \exp(-d'\gamma^2/2), \end{aligned}$$

where the last inequality holds by utilizing the Azuma-Hoeffding's inequality with the fact that $z_i \sim \mathrm{Unif}\{-1, 1\}$ is a bounded random variable. Consider a set $\mathcal{M}$ with cardinality $|\mathcal{M}|$, then there is at most $|\mathcal{M}|^2$ pair of $(\mathbf{x}, \mathbf{x}')$. Thus taking a union bound over all vector pairs $(\mathbf{x}, \mathbf{x}')$, we have

$$\mathbb{P}(\exists \mathbf{x}, \mathbf{x}' \in \mathcal{M}, \mathbf{x} \neq \mathbf{x}', \langle \mathbf{x}, \mathbf{x}' \rangle \geq d'\gamma) \leq |\mathcal{M}|^2 \exp(-d'\gamma^2/2),$$

thus

$$\mathbb{P}(\forall \mathbf{x}, \mathbf{x}' \in \mathcal{M}, \mathbf{x} \neq \mathbf{x}', \langle \mathbf{x}, \mathbf{x}' \rangle \leq d'\gamma) \geq 1 - |\mathcal{M}|^2 \exp(-d'\gamma^2/2),$$

Once we have that $|\mathcal{M}|^2 \exp(-d'\gamma^2/2) < 1$, there exists a set $\mathcal{M}$ such that for any two different vector $\mathbf{x}, \mathbf{x}' \in \mathcal{M}$, $\langle \mathbf{x}, \mathbf{x}' \rangle \leq d\gamma$. □

E.2 Proof of Lemma D.2

We start our lower bound proof from Fano's inequality.

Lemma E.1 (Fano's inequality, [6])**.** Consider probability measures $\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}$ on space Ω parameterized by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then for any estimator $\widehat{\boldsymbol{\theta}}$ on Ω and any comparison law $\mathbb{P}_0$ on Ω

$$\frac{1}{|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{P}_{\boldsymbol{\theta}}[\widehat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}] \geq 1 - \frac{\log 2 + \frac{1}{|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_0)}{\log|\boldsymbol{\Theta}|}.$$

Then we will start our proof from the deterministic algorithms, which could be further extended to random algorithms using Yao's principle [22].

Proof of Lemma D.2. We denote Y_k as such a trajectory at episode k and $Y_{1:k}$ for the trajectories $Y_1, \cdots, Y_k$. We have for the KL divergence over joint distribution $Y_{1:k}$,

$$\begin{aligned}
&\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K}), \mathbb{P}_0(Y_{1:K})) \\
&= \sum_{Y_{1:K}} \mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K}) \log(\mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K})/\mathbb{P}_0(Y_{1:K})) \\
&= \sum_{Y_{1:K}} \mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K-1})\mathbb{P}_{\boldsymbol{\theta}}(Y_K|Y_{1:K-1}) \log(\mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K-1})/\mathbb{P}_0(Y_{1:K-1})) \\
&\quad + \sum_{Y_{1:K}} \mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K-1})\mathbb{P}_{\boldsymbol{\theta}}(Y_K|Y_{1:K-1}) \log(\mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K-1})/\mathbb{P}_0(Y_{1:K-1})) \\
&= \sum_{Y_{1:K-1}} \mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K-1}) \log(\mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K-1})/\mathbb{P}_0(Y_{1:K-1})) \sum_{Y_K} \mathbb{P}_{\boldsymbol{\theta}}(Y_K|Y_{1:K-1}) \\
&\quad + \sum_{Y_{1:K-1}} \mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K-1}) \sum_{Y_k} \mathbb{P}_{\boldsymbol{\theta}}(Y_K|Y_{1:K-1}) \log(\mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K-1})/\mathbb{P}_0(Y_{1:K-1})) \\
&= \mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K-1}), \mathbb{P}_0(Y_{1:K-1})) + \mathbb{E}_{Y_{1:K-1}}[\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}(Y_K|Y_{1:K-1}, \mathbb{P}_0(Y_K|Y_{1:K-1})].
\end{aligned}$$

Thus by further expanding the above equations, we have

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}(Y_{1:K}), \mathbb{P}_0(Y_{1:K})) = \sum_{k=1}^{K} \mathbb{E}_{Y_{1:k-1}}[\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}(Y_k|Y_{1:k-1}), \mathbb{P}_0(Y_k|Y_{1:k-1})],$$

where we denote $\mathbb{E}_{Y_{1:0}}[\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}(Y_1|1:0), \mathbb{P}_0(Y_1|1:0)] := \mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}(Y_1), \mathbb{P}_0(Y_0))$ for consistency.

Since for any deterministic algorithm, in any episode, the trajectory $s_1, a_1, \cdots, s_H, a_H$ is determined after the algorithm goes into $S_{2,1}$ or $S_{2,2}$, furthermore, for these deterministic algorithms, the first action a at k-th trajectory is fixed given previous knowledge $Y_{1:k}$. Therefore, the distribution of the whole trajectory could be replaced by the distribution of $S_{2,1}$ and $S_{2,2}$. We have there are at most two possible value for Y_k, we denote the trajectory $S_1, a_1, S_{2,1}, \cdots, S_{2,1}$ by $Y_k = 0$ and the other trajectory $S_1, a_1, S_{2,2}, \cdots, S_{2,2}$ by $Y_k = 1$. We define the comparison distribution $\mathbb{P}_0$ as

$$\begin{aligned}
\mathbb{P}_0(Y_k = 0|Y_{1:k-1}) &= \frac{1}{|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}_i\in\boldsymbol{\Theta}} \mathbb{P}_{\boldsymbol{\theta}_i}(Y_k = 0|Y_{1:k-1}) := \frac{1}{2} + \frac{\alpha}{\sqrt{2}d'}\langle \mathbf{a}, \bar{\boldsymbol{\theta}}\rangle \\
\mathbb{P}_0(Y_k = 1|Y_{1:k-1}) &= \frac{1}{|\boldsymbol{\Theta}|}\sum_{\boldsymbol{\theta}_i\in\boldsymbol{\Theta}} \mathbb{P}_{\boldsymbol{\theta}_i}(Y_k = 1|Y_{1:k-1}) := \frac{1}{2} - \frac{\alpha}{\sqrt{2}d'}\langle \mathbf{a}, \bar{\boldsymbol{\theta}}\rangle,
\end{aligned}$$

where we denote $\bar{\boldsymbol{\theta}}$ is the mean value of $\widetilde{\boldsymbol{\theta}}_i \in \mathcal{M}$. (Recall that $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_i | \boldsymbol{\theta}_i = (\sqrt{2}, \alpha\widetilde{\boldsymbol{\theta}}_i^\top/\sqrt{d'})^\top\}$). For simplicity, we use $\mathbb{P}_{\boldsymbol{\theta}_i}$ and $\mathbb{P}_0$ to denote the distributions for the whole trajectory defined above.

Then we could bound the KL divergence between $\mathbb{P}_0$ and $\mathbb{P}_{\boldsymbol{\theta}_i}$ as

$$\begin{aligned}
&\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_i}, \mathbb{P}_0) \\
&= \left(\frac{1}{2} + \frac{\alpha}{\sqrt{2}d'}\langle \mathbf{a}, \widetilde{\boldsymbol{\theta}}_i\rangle\right) \log\left(\frac{\sqrt{2}d' + \alpha\langle \mathbf{a}, \widetilde{\boldsymbol{\theta}}_i\rangle}{\sqrt{2}d' + \alpha\langle \mathbf{a}, \bar{\boldsymbol{\theta}}\rangle}\right) \\
&+ \left(\frac{1}{2} - \frac{\alpha}{\sqrt{2}d'}\langle \mathbf{a}, \widetilde{\boldsymbol{\theta}}_i\rangle\right) \log\left(\frac{\sqrt{2}d' - \alpha\langle \mathbf{a}, \widetilde{\boldsymbol{\theta}}_i\rangle}{\sqrt{2}d' - \alpha\langle \mathbf{a}, \bar{\boldsymbol{\theta}}\rangle}\right) \\
&\leq \left(\frac{1}{2} + \frac{\alpha}{\sqrt{2}d'}\langle \mathbf{a}, \widetilde{\boldsymbol{\theta}}_i\rangle\right) \frac{\alpha\langle \mathbf{a}, \widetilde{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}}\rangle}{\sqrt{2}d' + \alpha\langle \mathbf{a}, \bar{\boldsymbol{\theta}}\rangle} - \left(\frac{1}{2} - \frac{\alpha}{\sqrt{2}d'}\langle \mathbf{a}, \widetilde{\boldsymbol{\theta}}_i\rangle\right) \frac{\alpha\langle \mathbf{a}, \widetilde{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}}\rangle}{\sqrt{2}d' - \alpha\langle \mathbf{a}, \bar{\boldsymbol{\theta}}\rangle}.
\end{aligned}$$

Taking summation over $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$, recall that $\bar{\boldsymbol{\theta}}$ is the mean value of $\widetilde{\boldsymbol{\theta}}_i \in \mathcal{M}$, we have

$$\sum_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}} \mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_i}, \mathbb{P}_0) = \frac{2\alpha^2}{2d'^2 - \alpha^2\langle \mathbf{a}, \bar{\boldsymbol{\theta}}\rangle^2} \sum_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}} \langle \mathbf{a}, \widetilde{\boldsymbol{\theta}}_i\rangle\langle \mathbf{a}, \widetilde{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}}\rangle.$$

Given the fact that $\langle \mathbf{x}, \mathbf{x}'\rangle \leq d'$ for any $\mathbf{x}, \mathbf{x}' \leq \mathcal{A}$, one can easily get that

$$\frac{1}{|\boldsymbol{\Theta}|} \sum_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}} \mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_i}, \mathbb{P}_0) \leq \frac{4\alpha^2}{2 - \alpha^2}.$$

Plugging the above inequality into the decomposition of KL divergence, from Fano's inequality Lemma E.1, we have

$$\delta \geq \frac{1}{|\boldsymbol{\Theta}|} \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{P}_{\boldsymbol{\theta}}[\widehat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}] \geq 1 - \left(\frac{d'}{16} - 3\right)^{-1} \left(\log 2 + \frac{4K\alpha^2}{2 - \alpha^2}\right).$$

Replacing d' by $d - 1$, we can get the same result as the statement of the lemma. □

E.3 Proof of Lemma D.3

We show the proof for Lemma D.3 by establishing different reward functions for this MDP structure.

Proof of Lemma D.3. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we build the reward sequence as $r(S_1) = r(S_{2,1}) = 0, r(S_{2,2}) = 1$, then any (ϵ, δ)-correct algorithm guarantees that

$$\mathbb{P}(V^*(S_1, r; \boldsymbol{\theta}) - V^\pi(S_1, r; \boldsymbol{\theta}) \leq \epsilon) \geq 1 - \delta, \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Our proof is to show that, as long as $\alpha \geq \frac{2\sqrt{2}\epsilon}{H-1}$, we can build up the estimation of $\boldsymbol{\theta}$ using $\boldsymbol{\theta}_i = \left(\sqrt{2}, \alpha\mathbf{a}_i^\top/\sqrt{d}\right)^\top$ where $\mathbf{a}_i$ is determined by $a_i = \pi(S_1)$. It is guaranteed that $\mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i = \boldsymbol{\theta}] \geq 1 - \delta$ and furthermore, $\mathbb{E}_{\boldsymbol{\theta} \sim \mathrm{Unif}(\boldsymbol{\Theta})}(\mathbb{1}[\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}]) \geq 1 - \delta$.

Suppose for the MDP parameter $\boldsymbol{\theta}_i$, it is easy to find that the optimal policy for the first step is $\pi^*(S_1) = a_i$. Suppose that for any policy $\pi(S_1) = a_j$ where $j \neq i$, then from the MDP structure, the gap between policy and the optimal policy is

$$\begin{aligned}
V^*(S_1, r; \boldsymbol{\theta}_i) - V^\pi(S_1, r; \boldsymbol{\theta}_i) &= \frac{(H-1)\alpha}{\sqrt{2}d}(\langle \mathbf{a}_i, \mathbf{a}_i\rangle - \langle \mathbf{a}_i, \mathbf{a}_j\rangle) \\
&\geq \frac{(H-1)\alpha}{\sqrt{2}d'}(d' - d'/2) \\
&= \frac{(H-1)\alpha}{2\sqrt{2}},
\end{aligned}$$

as long as we have $\alpha \geq \frac{2\sqrt{2}\epsilon}{H-1}$, we can get the policy gap $V^*(S_1, r; \boldsymbol{\theta}_i) - V^\pi(S_1, r; \boldsymbol{\theta}_i) \geq \epsilon$.

Therefore, it is easy to show that the estimation using the policy $\pi(S_1)$ is guaranteed with successful rate at least $1 - \delta$ for any MDP parameter $\boldsymbol{\theta}_i$, thus we can further conclude that $\mathbb{E}_{\boldsymbol{\theta} \sim \mathrm{Unif}(\boldsymbol{\Theta})}(\mathbb{1}[\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}]) \geq 1 - \delta$. □