Optimal Horizon-Free Reward-Free Exploration for Linear Mixture MDPs

Junkai Zhang [1] **Weitong Zhang** [1] **Quanquan Gu** [1]

Abstract

We study reward-free reinforcement learning (RL) with linear function approximation, where the agent works in two phases: (1) in the exploration phase, the agent interacts with the environment but cannot access the reward; and (2) in the planning phase, the agent is given a reward function and is expected to find a near-optimal policy based on samples collected in the exploration phase. The sample complexities of existing reward-free algorithms have a polynomial dependence on the planning horizon, which makes them intractable for long planning horizon RL problems. In this paper, we propose a new reward-free algorithm for learning linear mixture Markov decision processes (MDPs), where the transition probability can be parameterized as a linear combination of known feature mappings. At the core of our algorithm is uncertainty-weighted value-targeted regression with exploration-driven pseudo-reward and a high-order moment estimator for the aleatoric and epistemic uncertainties. When the total reward is bounded by 1, we show that our algorithm only needs to explore $\widetilde{O}\left(d^2\varepsilon^{-2}\right)$ episodes to find an ε-optimal policy, where d is the dimension of the feature mapping. The sample complexity of our algorithm only has a polylogarithmic dependence on the planning horizon and therefore is "horizon-free". In addition, we provide an $\Omega\left(d^2\varepsilon^{-2}\right)$ sample complexity lower bound, which matches the sample complexity of our algorithm up to logarithmic factors, suggesting that our algorithm is optimal.

1. Introduction

In Reinforcement Learning (RL), the agent sequentially interacts with the environment by executing policies and receiving observations, including states and rewards. The goal of the agent is to maximize the total reward. To achieve this goal, the agent needs to explore the environment and exploit the collected information to find the optimal policy. The exploration has long been considered as a central challenge for RL, for which the agent needs to strategically visit states to learn transition dynamics and the value of different states. RL algorithms are often designed to exploit the transition and the reward information to achieve efficient exploration.

Unfortunately, in many real-world RL problems, reward functions are manually designed to incentive the agent to learn specific tasks, and they may change over time (Achiam et al., 2017; Tessler et al., 2018; Miryoosefi et al., 2019). To avoid learning the transition dynamics repeatedly, Jin et al. (2020a) proposed a new RL paradigm, *Reward-free Exploration*, which separates exploration and planning into two different phases. In the exploration phase, the agent cannot access the real reward function. It can only learn the transition dynamics based on the collected episodes. While in the planning phase, the agent can no longer interact with the environment but has access to the reward function. The goal is to find the optimal policy based on the reward function and previous exploration. A series of work (Jin et al., 2020a; Kaufmann et al., 2021b; Ménard et al., 2021; Zhang et al., 2020) have achieved the optimal sample complexity of $\widetilde{O}(H^2S^2A\varepsilon^{-2})$, where H is the planning horizon, S is the number of states, and A is the number of actions.

The sample complexity for learning tabular MDPs shows that learning becomes intractable when the sizes of the state and action spaces increase without further structural assumptions. Linear function approximation is a classical approach to deal with this challenge, which approximates the transition dynamic or the value function by linear functions on compact feature mappings. To this end, Wang et al. (2020b); Zanette et al. (2020c) studied reward-free RL in linear MDPs (Yang & Wang, 2019; Jin et al., 2020b). Zhang et al. (2021a) studied reward-free exploration for linear mixture MDPs (Ayoub et al., 2020; Zhou et al., 2021c), where the transition probability is a linear combination of feature mappings. The subsequent work Chen et al. (2021) has achieved near optimal sample complexity $\widetilde{O}\left(H^3d(H+d)\varepsilon^{-2}\right)$. However, this sample complexity depends on the planning horizon, which will blow up for long planning horizon problems. Therefore, a natural question arises:

[1]Department of Computer Science, University of California, Los Angeles, California, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Can we design horizon-free, minimax optimal reward-free RL algorithms with linear function approximation?

In this paper, we answer this question affirmatively for linear mixture MDPs. In detail, our contributions are highlighted as follows.

- We propose an algorithm for reward-free exploration in the linear mixture MDP setting. The algorithm guides the agent to collect samples using a well-designed exploration-driven pseudo-reward function. With a novel analysis based on high-order moment estimation that precisely controls the aleatoric and epistemic uncertainties, our algorithm can achieve an $\widetilde{O}(d^2\varepsilon^{-2})$ sample complexity. This complexity only has polylogarithmic dependence on H.
- We show that any reward-free algorithm needs to explore at least $\Omega(d^2\varepsilon^{-2})$ episodes, to achieve an ε-optimal policy for any reward function, by constructing a special class of linear mixture MDPs. This lower bound matches the upper bound of our algorithm up to logarithmic factors, which indicates that our algorithm is optimal.
- When rescaling the reward to satisfy $\sum_{h=1}^{H} r_h(s_h, a_h) \leq H$, our algorithm achieves an $\widetilde{O}(d^2H^2\varepsilon^{-2})$ sample complexity. This improves the sample complexity in the previous work Chen et al. (2021), which requires the $d > H$ condition to match the lower bound.

Notation We use the lowercase letter to denote scalars and lower and uppercase boldface letters to denote vectors and matrices, respectively. We denote by $[n]$ the set $\{1, \cdots, n\}$, and by $\overline{[n]}$ the set $\{0, \cdots, n-1\}$. For a vector $\mathbf{x}$ and a positive semi-definite matrix $\mathbf{\Sigma}$, we denote by $\|\mathbf{x}\|_2$ the vector's Euclidean norm and define $\|\boldsymbol{x}\|_{\mathbf{\Sigma}} = \sqrt{\boldsymbol{x}^\top \mathbf{\Sigma} \boldsymbol{x}}$. For two positive sequences $\{a_n\}$ and $\{b_n\}$ with $n = 1, 2, \cdots$, we write $a_n = O(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \leq Cb_n$ holds for all $n \geq 1$, write $a_n = \Omega(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \geq Cb_n$ holds for all $n \geq 1$, and write $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. We use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to further hide the polylogarithmic factors.

2. Related Work

RL with Linear Function Approximation In recent years, a series of works have been devoted to the study of RL with linear function approximation (Jiang et al., 2017; Dann et al., 2018; Yang & Wang, 2019; Wang et al., 2019; Du et al., 2019; Sun et al., 2019; Jin et al., 2020b; Zanette et al., 2020a;b; Yang & Wang, 2020; Modi et al., 2020; Ayoub et al., 2020; Jia et al., 2020; Cai et al., 2020; Weisz et al., 2021; Zhou et al., 2021d;a; He et al., 2022; Agarwal et al., 2022). Our work belongs to the linear mixture MDP setting (Yang & Wang, 2019; Modi et al., 2020; Ayoub et al., 2020; Jia et al., 2020; Zhou et al., 2021a;d), where the transition kernel can be parameterized as a linear combination of some basic transition probability functions. Zhou et al. (2021a) firstly achieved minimax regret $\widetilde{O}\left(dH\sqrt{T}\right)$ in linear mixture MDPs by proposing a Bernstein-type concentration inequality for self-normalized martingales. Another kind of popular linearly parameterized MDP is linear MDP (Wang et al., 2019; Du et al., 2019; Yang & Wang, 2020; Jin et al., 2020b; Zanette et al., 2020a; Wang et al., 2020c; He et al., 2021), which assumes both transition probability and reward function are linear functions of known feature mappings on state-action pairs. Under this setting, Jin et al. (2020b) firstly proposed statistically and computationally efficient algorithm LSVI-UCB and achieved $\widetilde{O}\left(\sqrt{d^3H^3T}\right)$ regret bound. Recent works (He et al., 2022) further achieved nearly minimax optimal regret $\widetilde{O}(d\sqrt{H^3K})$ by proposing computationally efficient algorithm LSVI-UCB++. Its concurrent work (Agarwal et al., 2022) achieves similar result under assumption $\sum_{h=1}^{H} r_h(s_h, a_h) \leq 1$ with regret upper bound of $\widetilde{O}(d\sqrt{HT} + d^6H^5)$.

Reward-free RL Unlike standard RL settings, exploration and planning in reward-free RL are separated into two different phases. Jin et al. (2020a) achieved $\widetilde{O}(H^5S^2A/\varepsilon^2)$ sample complexity in tabular MDPs by executing exploratory policy visiting states with probability proportional to its maximum visitation probability under any possible policy. Subsequent works (Kaufmann et al., 2021b; Ménard et al., 2021) proposed algorithms RF-UCRL and RF-Express to gradually improve the result to $\widetilde{O}\left(H^3S^2A\varepsilon^{-2}\right)$. The optimal sample complexity bound $\widetilde{O}(H^2S^2A\varepsilon^{-2})$ was achieved by algorithm SSTP proposed in Zhang et al. (2020), which matched the lower bound provided in Jin et al. (2020a) up to logarithmic factors.

Recent years have witnessed a trend of reward-free exploration in RL with linear function approximation (Wang et al., 2020b; Zanette et al., 2020c; Zhang et al., 2021a; Chen et al., 2021; Huang et al., 2022; Wagenmaker et al., 2022). The near minimax optimal sample complexity of reward-free exploration in linear mixture MDP was achieved by Chen et al. (2021) when $d > H$ using the well-designed exploration-driven pseudo reward function. On the other hand, in the linear MDP setting, Wang et al. (2020b) proposed the first efficient algorithm, which only required $\widetilde{O}(H^6d^3\varepsilon^{-2})$ sample complexity. The subsequent works, Chen et al. (2021) and Wagenmaker et al. (2022), gave sample complexity of $\widetilde{O}(H^4d^3\varepsilon^{-2})$ and $\widetilde{O}(H^5d^2\varepsilon^{-2})$, which are the sharpest for H and d, respectively. Some significant works are summarized in Table 1.

Setting	Algorithm	Rewards Scale	Time Homo.	Sample Complexity
Tabular MDP	Jin et al. (2020a)	$r_h(s_h, a_h) \in [0, 1]$	$\times$	$\widetilde{O}(H^5S^2A\varepsilon^{-2})$
	RF-UCRL (Kaufmann et al., 2021a)	$r_h(s_h, a_h) \in [0, 1]$	$\times$	$\widetilde{O}(H^4S^2A\varepsilon^{-2})$
	RF-Express (Ménard et al., 2021)	$r_h(s_h, a_h) \in [0, 1]$	$\times$	$\widetilde{O}(H^3S^2A\varepsilon^{-2})$
	SSTP (Zhang et al., 2020)	$\sum_{h=1}^H r_h(s_h, a_h) \le 1$	✓	$\widetilde{O}(S^2A\varepsilon^{-2})$
	Lower bound (Jin et al., 2020a)	$r_h(s_h, a_h) \in [0, 1]$	$\times$	$\Omega(H^2S^2A\varepsilon^{-2})$
	Lower bound (Zhang et al., 2020)	$\sum_{h=1}^H r_h(s_h, a_h) \le 1$	✓	$\Omega(S^2A\varepsilon^{-2})$
Linear MDP	Wang et al. (2020b)	$r_h(s_h, a_h) \in [0, 1]$	$\times$	$\widetilde{O}\left(H^6d^3\varepsilon^{-2}\right)$
	FRANCIS (Zanette et al., 2020c)	$r_h(s_h, a_h) \in [0, 1]$	$\times$	$\widetilde{O}\left(H^5d^3\varepsilon^{-2}\right)$
	RFLIN (Wagenmaker et al., 2022)	$r_h(s_h, a_h) \in [0, 1]$	$\times$	$\widetilde{O}\left(H^5d^2\varepsilon^{-2}\right)$
Linear Mixture MDP	UCRL-RFE+ (Zhang et al., 2021a)	$r_h(s_h, a_h) \in [0, 1]$	✓	$\widetilde{O}\left(H^4d(H+d)\varepsilon^{-2}\right)$
	Chen et al. (2021)	$r_h(s_h, a_h) \in [0, 1]$	$\times$	$\widetilde{O}\left(H^3d(H+d)\varepsilon^{-2}\right)$
	Our work (Cor. 5.2)	$\sum_{h=1}^H r_h(s_h, a_h) \le 1$	✓	$\widetilde{O}(d^2\varepsilon^{-2})$
	Our work (Cor. 5.4)	$\sum_{h=1}^H r_h(s_h, a_h) \le H$	✓	$\widetilde{O}(H^2d^2\varepsilon^{-2})$
	Lower bound (Thm. 5.6)	$\sum_{h=1}^H r_h(s_h, a_h) \le 1$	✓	$\Omega\left(d^2\varepsilon^{-2}\right)$
	Lower bound (Cor. 5.8)	$r_h(s_h, a_h) \in [0, 1]$	✓	$\Omega(H^2d^2\varepsilon^{-2})$

Table 1. Comparison of episodic reward-free RL algorithms. Column **Time Homo.** means if the algorithm is time-homogeneous (✓) or not ($\times$), rows with light blue background indicates our results

Horizon-free RL The long planning horizon has long been viewed as RL's main challenge. However, a series of works has shown that RL is no more difficult than contextual bandits by removing the influence of the total reward scale. In tabular MDPs, the algorithm proposed in Wang et al. (2020a) firstly achieved polylogarithmic H dependency sample complexity bound $\widetilde{O}(S^5A^4\varepsilon^{-2})$ by carefully reusing samples and avoid unnecessary sampling. Zhang et al. (2021b) further proposed an improved algorithm `MVP` to achieve near-optimal regret bound $\widetilde{O}(\sqrt{SAK} + S^2A)$ based on new Bernstein-type bonus. Similar polylogarithmic H dependency bounds had been established by Ren et al. (2021) for linear MDP with anchor points, Tarbouriech et al. (2021) for the stochastic shortest path. Li et al. (2022) achieved the surprising H independent sample complexity bound $O((SA)^{O(S)}\varepsilon^{-5})$ by building a connection between discounted MDPs and episodic MDPs and a novel perturbation analysis in MDPs. The algorithm proposed by Zhang et al. (2022) further improved the sample complexity to $O(S^9A^3\varepsilon^{-2}\text{polylog}(S, A, \varepsilon^{-1}))$ only depending on state and action spaces size polynomially by exploiting the power of stationary policy. Thanks to the linear function approximation, Zhou & Gu (2022) firstly achieve horizon-free regret bound $\widetilde{O}(d\sqrt{K} + d^2)$ independent of state and action spaces size. However, all the above works are limited to standard RL settings. In the paradigm of reward-free exploration, the only horizon-free result was achieved by Zhang et al. (2021a) with sample complexity bound of $\widetilde{O}(S^2A\varepsilon^{-2})$, where the polynomial dependency on S and A is still unacceptable when the states and actions spaces are large. Our algorithm HF-UCRL-RFE++ establishes the first horizon-free sample complexity bound independent of state and action spaces size in reward-free exploration.

3. Preliminaries

We consider an episodic finite horizon Markov Decision Process (MDP) $\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \mathbb{P}\right)$, where $\mathcal{S}$ is the countable (and maybe infinite) state space, $\mathcal{A}$ is the action space, H is the length of the episode, $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the deterministic reward function, and $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the time-homogeneous transition probability.

Based on the current state $s \in \mathcal{S}$ and the time step $h \in [H]$, a policy π chooses the action $a \in \mathcal{A}$ to be executed by the agent. Formally, we denote a policy as $\pi = \{\pi_h\}_{h=1}^H$, where $\pi_h : \mathcal{S} \to \mathcal{A}$ is a function which maps a state s to an action a. For any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$, we define the

action-value function $Q_h^\pi(s,a)$ and value function $V_h^\pi(s)$ as follows:

$$Q_h^\pi(s,a) = \mathbb{E}\Big[\sum_{h'=h}^{H} r(s_{h'},a_{h'})\Big| s_h = s, a_h = a, \\ s_{h'} \sim \mathbb{P}(\cdot|s_{h'-1},a_{h'-1}), a_{h'} = \pi_{h'}(s_{h'})\Big],$$
$$V_h^\pi(s) = Q_h^\pi\big(s,\pi_h(s)\big).$$

We define the optimal value function $V_h^*(\cdot)$ and optimal action-value function $Q_h^*(\cdot,\cdot)$ as $V_h^*(\cdot) = \sup_\pi V_h^\pi(\cdot)$ and $Q_h^*(\cdot,\cdot) = \sup_\pi Q_h^\pi(\cdot,\cdot)$, respectively. For any function $V : \mathcal{S} \to [0,1]$, we introduce the following short-hands to denote the conditional expectation and variance of V:

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

For any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $h \in [H]$, the Bellman equation and Bellman optimality equation can be defined recursively as

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

In this paper, we make the structural assumption that the transition dynamic has a linear structure, which has been considered in prior works as below:

Definition 3.1 (Linear Mixture MDPs, Jia et al. 2020; Ayoub et al. 2020; Zhou et al. 2021c)**.** The unknown transition probability $\mathbb{P}$ is a linear combination of d signed basis measures $\boldsymbol{\phi}_i(s'|s,a)$, i.e., $\mathbb{P}(s'|s,a) = \sum_{i=1}^d \boldsymbol{\phi}_i(s'|s,a)\boldsymbol{\theta}_i^*$. Meanwhile, for any $V : \mathcal{S} \to [0,1]$, $i \in [d]$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, the summation $\sum_{s'\in\mathcal{S}} \boldsymbol{\phi}_i(s'|s,a)V(s')$ can be calculated efficiently within $\mathcal{O}$ time. For simplicity, let $\boldsymbol{\phi} = [\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_d]^\top$, $\boldsymbol{\theta}^* = [\boldsymbol{\theta}_1^*,\ldots,\boldsymbol{\theta}_d^*]^\top$ and $\boldsymbol{\phi}_V(s,a) = \sum_{s'\in\mathcal{S}} \boldsymbol{\phi}(s'|s,a)V(s')$. W.l.o.g., we assume $\|\boldsymbol{\theta}^*\|_2 \le B, \|\boldsymbol{\phi}_V(s,a)\|_2 \le 1$ for all $V : \mathcal{S} \to [0,1]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$.

We further assume that the accumulated reward of an episode for any trajectory is upper bounded by 1, which ensures that the only factors affecting the final statistical complexity are difficulties brought by exploration and long planning horizon rather than the scale of the total reward.

Assumption 3.2. (Bounded total reward) For any trajectory $\{s_h,a_h\}_{h=1}^H$, we have $0 \le \sum_{h=1}^H r_h(s_h,a_h) \le 1$.

We denote the set of reward functions satisfying Assumption 3.2 by $\mathcal{R}$.

Reward-free RL In reward-free RL, the real reward function is accessible only after the agent finishes the interactions with the environment. Specifically, the algorithm can be separated into two phases: (i) *Exploration phase*: the algorithm can't access the reward function but collect K episodes of samples by interacting with the environment. (ii) *Planning phase*: The algorithm is given reward function $\{r_h\}_{h=1}^H$ and is expected to find the optimal policy. (ε,δ)-learn and sample complexity of the algorithm is formally defined as follows.

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 $K(\varepsilon,\delta)$, if for any transition kernel $P \in \mathcal{P}$, after receiving $K(\varepsilon,\delta)$ 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 $1-\delta$, $V_1^*(s_1;r) - V_1^\pi(s_1;r) \le \varepsilon$.

4. Algorithms

In this section, we propose our reward-free exploration algorithm HF-UCRL-RFE++. This algorithm consists of two phases. In the exploration phase, it builds an estimator $\boldsymbol{\theta}$ for the linear mixture MDP transition kernel parameter $\boldsymbol{\theta}^*$ based on the sampled episodes. At a high level, the estimation follows the *value-targeted regression* (VTR) framework proposed by Jia et al. (2020). The VTR is basically a ridge regression with value functions as responses and feature mappings as predictors. However, value functions have no estimates since the reward function is not accessible. Therefore, the value functions and reward functions are replaced by well-designed exploration-driven pseudo-value functions and pseudo-reward functions. To achieve a better estimation, we further apply the *high-order moment estimation* (HOME) technique proposed by Zhou & Gu (2022). Then, during the planning phase, the algorithm uses the estimator $\boldsymbol{\theta}$ acquired in the exploration phase to find the optimal policy π for the given reward functions. Our algorithm is described in Algorithm 1.

Exploration-driven Pseudo Value Function As mentioned above, in the paradigm of reward-free exploration, we have to construct the pseudo-reward function to guide the agent in taking actions in the absence of the real reward function. As we adopt in this work, the most natural idea is to construct the pseudo-reward function related to uncertainty, which urges the agent to collect information about the most uncertain states and actions. Two approaches follow this idea: one is constructing the pseudo reward function directly measuring and maximizing the uncertainty of each stage, and the other is constructing the pseudo reward function maximizing the overall uncertainty along trajectories.

Algorithm 1 HF-UCRL-RFE++

Input: Confidence radius $\{\beta_k\}$, regularization parameter λ, number of the high-order estimator M

Phase I: Exploration Phase
Initialize $\widehat{\boldsymbol{\Sigma}}_{1,1,m} \leftarrow \lambda\mathbf{I}$, $\widetilde{\boldsymbol{\Sigma}}_{1,1,m} \leftarrow \lambda\mathbf{I}$, $\widetilde{\mathbf{b}}_{1,1,m} \leftarrow \mathbf{0}$, $\widehat{\mathbf{b}}_{1,1,m} \leftarrow \mathbf{0}$ for all $m \in \overline{[M]}$, $\mathcal{U}_1 = \big\{\boldsymbol{\theta} | \boldsymbol{\theta} \in \mathbb{R}^d\big\}$.
Set $\widehat{\boldsymbol{\theta}}_{1,m} \leftarrow \widehat{\boldsymbol{\Sigma}}_{1,1,m}^{-1}\widehat{\mathbf{b}}_{1,1,m}$, $\widetilde{\boldsymbol{\theta}}_{1,m} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{1,1,m}^{-1}\widetilde{\mathbf{b}}_{1,1,m}$ for all $m \in \overline{[M]}$.
for $k = 1, 2, \cdots, K$ **do**
  Calculate $\pi_k, \boldsymbol{\theta}_k, r_k = \operatorname{argmax}_{\pi, \boldsymbol{\theta} \in \mathcal{U}_k, r \in R} \widehat{V}_{k,1}(s_1; \boldsymbol{\theta}, \pi, r)$, where $\widehat{V}_{k,1}$ is defined in (4.2). Denote $\big\{\widetilde{V}_{k,h}(\cdot)\big\}_{h=1}^{H} = \{V_h(\cdot; \boldsymbol{\theta}_k, \pi_k, r_k)\}_{h=1}^{H}$.
  Receive initial state $s_1^k = s_1$.
  for $h = 1, 2, \cdots, H$ **do**
    Execute $a_h^k = \pi_h^k\left(s_h^k\right)$, receive $s_{h+1}^k \sim \mathbb{P}\left(\cdot | s_h^k, a_h^k\right)$.
    For $m \in \overline{[M]}$, denote $\widehat{\boldsymbol{\phi}}_{k,h,m} = \boldsymbol{\phi}_{\widehat{V}_{k,h+1}^{2^m}}(s_h^k, a_h^k)$, $\widetilde{\boldsymbol{\phi}}_{k,h,m} = \boldsymbol{\phi}_{\widetilde{V}_{k,h+1}^{2^m}}(s_h^k, a_h^k)$
    Set $\big\{\widehat{\sigma}_{k,h,m}\big\} \leftarrow \text{HOME}_{\text{Alg. 2}}\Big(\big\{\widehat{\boldsymbol{\phi}}_{k,h,m}, \widehat{\boldsymbol{\theta}}_{k,m}, \widehat{\boldsymbol{\Sigma}}_{k,h,m}, \dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}\big\}, \beta_k, \alpha, \gamma\Big)$
    Set $\big\{\widetilde{\sigma}_{k,h,m}\big\} \leftarrow \text{HOME}_{\text{Alg. 2}}\Big(\big\{\widetilde{\boldsymbol{\phi}}_{k,h,m}, \widetilde{\boldsymbol{\theta}}_{k,m}, \widetilde{\boldsymbol{\Sigma}}_{k,h,m}, \dot{\widetilde{\boldsymbol{\Sigma}}}_{k,m}\big\}, \beta_k, \alpha, \gamma\Big)$.
    Set $\widetilde{\boldsymbol{\Sigma}}_{k,h+1,m} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{k,h,m} + \widetilde{\boldsymbol{\phi}}_{k,h,m}\widetilde{\boldsymbol{\phi}}_{k,h,m}^{\top}\widetilde{\sigma}_{k,h,m}^{-2}$ for $m \in \overline{[M]}$
    Set $\widehat{\boldsymbol{\Sigma}}_{k,h+1,m} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k,h,m} + \widehat{\boldsymbol{\phi}}_{k,h,m}\widehat{\boldsymbol{\phi}}_{k,h,m}^{\top}\widehat{\sigma}_{k,h,m}^{-2}$ for $m \in \overline{[M]}$
    Set $\widetilde{\mathbf{b}}_{k,h+1,m} \leftarrow \widetilde{\mathbf{b}}_{k,h,m} + \widetilde{\boldsymbol{\phi}}_{k,h,m}\widetilde{V}_{k,h+1}^{2^m}(s_{h+1}^k)\widetilde{\sigma}_{k,h,m}^{-2}$ for $m \in \overline{[M]}$
    Set $\widehat{\mathbf{b}}_{k,h+1,m} \leftarrow \widehat{\mathbf{b}}_{k,h,m} + \widehat{\boldsymbol{\phi}}_{k,h,m}\widehat{V}_{k,h+1}^{2^m}(s_{h+1}^k)\widehat{\sigma}_{k,h,m}^{-2}$ for $m \in \overline{[M]}$
  end for
  $\dot{\widetilde{\boldsymbol{\Sigma}}}_{k+1,m} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{k,H+1,m}$, $\dot{\widehat{\boldsymbol{\Sigma}}}_{k+1,m} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k,H+1,m}$
  Set $\widetilde{\boldsymbol{\Sigma}}_{k+1,1,m} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{k,H+1,m}$, $\widetilde{\mathbf{b}}_{k+1,1,m} \leftarrow \widetilde{\mathbf{b}}_{k,H+1,m}$, $\widetilde{\boldsymbol{\theta}}_{k+1,m} = \widetilde{\boldsymbol{\Sigma}}_{k+1,1,m}^{-1}\widetilde{\mathbf{b}}_{k+1,1,m}$.
  Set $\widehat{\boldsymbol{\Sigma}}_{k+1,1,m} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k,H+1,m}$, $\widehat{\mathbf{b}}_{k+1,1,m} \leftarrow \widehat{\mathbf{b}}_{k,H+1,m}$, $\widehat{\boldsymbol{\theta}}_{k+1,m} = \widehat{\boldsymbol{\Sigma}}_{k+1,1,m}^{-1}\widehat{\mathbf{b}}_{k+1,1,m}$.
  Update the confidence set $\mathcal{U}_k$ to $\mathcal{U}_{k+1}$ by adding constraints (4.5), (4.6).
end for
Phase II: Planning Phase
Receive reward function r.
$\widehat{\pi}_r = \arg\max_{\pi} V_1(\cdot; \boldsymbol{\theta}_K, \pi, r)$.
Return policy $\widehat{\pi}_r$.

Zhang et al. (2021d) took the first approach, constructing the pseudo-reward function in the form of

$$r_h^k(s,a) = \min\Big\{1, \frac{2\beta}{H}\sqrt{\max_{V \in \mathcal{S} \to [0, H-h]} \|\boldsymbol{\phi}_V(s,a)\|_{\boldsymbol{\Sigma}_{1,k}^{-1}}}\Big\},$$

and the pseudo-value function to be the argument of the maxima for the above uncertainty measure. Under this construction, the suboptimality in the planning phase can be bounded by the accumulation of uncertainty. This approach is straightforward but has the following two drawbacks. Firstly, without the truncation for accumulation of uncertainty, the upper bound of overall suboptimality in the planning phase will be in the scale of $O(H)$, which is meaningless since the value function lies in the interval of $[0, 1]$ under our assumption. Second, since VTR utilizes value functions' variance information for $\boldsymbol{\theta}$ estimation, it requires a Bellman-equation-type equality between two consecutive stages h and $h+1$. However, the first approach does not satisfy this requirement, preventing us from acquiring a more accurate estimate.

To address the above issues, we follow the design of pseudo value function proposed in Chen et al. (2021). In particular, we are constructing the pseudo-reward function aiming to maximize the overall uncertainty along trajectories. We view the uncertainty of states and actions as a function of (pseudo) reward function r, policy π, and transition kernel parameter $\boldsymbol{\theta}$ defined as follows

$$u_{k,h}(s,a;\boldsymbol{\theta},\pi,r) = \min\Big\{1, \beta\big\|\boldsymbol{\phi}_{V_h(\cdot;\boldsymbol{\theta},\pi,r)}(s,a)\big\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\Big\}, \tag{4.1}$$

where $V_h(\cdot; \boldsymbol{\theta}, \pi, r)$ is the the value function of policy π for linear mixture MDP with transition kernel parameter $\boldsymbol{\theta}$ and the reward function r, and the overall uncertainty along the trajectory is the truncated sum of each step uncertainty defined as

$$\bar{V}_{k,h}(s;\boldsymbol{\theta},\pi,r) = \min\Big\{1, u_{k,h}(s,\pi(s);\boldsymbol{\theta},\pi,r) + \boldsymbol{\phi}_{\widetilde{V}_{k,h+1}(\cdot;\boldsymbol{\theta},\pi,r)}^{\top}(s,\pi(s))\boldsymbol{\theta}^*\Big\}.$$

However, the definition of $\bar{V}_{k,h}(s;\boldsymbol{\theta},\pi,r)$ involves $\boldsymbol{\theta}^*$, which is unknown to the agent. Hence, we construct the optimistic estimation of $\bar{V}_{k,h}(s;\boldsymbol{\theta},\pi,r)$ as $\widehat{V}_{k,h}(s;\boldsymbol{\theta},\pi,r)$ defined as

$$\begin{aligned}\widehat{V}_{k,h}(s;\boldsymbol{\theta},\pi,r) = \min\Big\{1, u_{k,h}(s,\pi(s);\boldsymbol{\theta},\pi,r) \\ + 2\beta\big\|\boldsymbol{\phi}_{\widehat{V}_{k,h+1}(\cdot;\boldsymbol{\theta},\pi,r)}(s,\pi(s))\big\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}^{-1}} \\ + \boldsymbol{\phi}^{\top}_{\widehat{V}_{k,h+1}(\cdot;\boldsymbol{\theta},\pi,r)}(s,\pi(s))\boldsymbol{\theta}\Big\}.\end{aligned} \tag{4.2}$$

Notable, the definitions of $u_{k,h}$ and $\widehat{V}_{k,h}$ involve the covariance matrices $\dot{\bar{\boldsymbol{\Sigma}}}_{k,0}$ and $\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}$, which are computed at the end of the preceding episode at Line 17 of Algorithm 1. In the following content, when there is no confusion, we may write $\widehat{V}_{k,h}(\cdot) = \widehat{V}_{k,h}(\cdot;\boldsymbol{\theta}_k,\pi_k,r_k)$, $u_{k,h}(\cdot,\cdot) = u_{k,h}(\cdot,\cdot;\boldsymbol{\theta}_k,\pi_k,r_k)$. In order to collect more information, the agent is expected to transit through the trajectory with the largest uncertainty $\widehat{V}_{k,h}$. It is notable that $\widehat{V}_{k,h}$ is a function of (pseudo) reward function r_k, policy π_k, and transition kernel parameter $\boldsymbol{\theta}_k$. Thus, at the beginning of each episode, we set r_k, π_k, and $\boldsymbol{\theta}_k$ to be arguments of the maxima, as presented in Line 5 in Algorithm 1. Through this process, we acquire the pseudo value function r_k, which is essential for reward-free exploration. Afterward, the algorithm collects samples along trajectories induced by policy π_k and improves the estimation of $\boldsymbol{\theta}_k$ in Line 6 to Line 21. In this stage, Algorithm 1 encounters two series of functions in the form of Bellman equations; one is the sum of pseudo rewards r, $\widetilde{V}_{k,h}(\cdot) = V_h(\cdot;\boldsymbol{\theta}_k,\pi_k,r_k)$, which we refer as pseudo value function, and one is the uncertainty along the trajectory, $\widehat{V}_{k,h}$. These two series of functions are both eligible for refined VTR and thus help estimate $\boldsymbol{\theta}$, as we will explain in the following.

High-order Moment Estimation The key technique used in our algorithm consists of two series of high-order estimations for the transition kernel parameter $\boldsymbol{\theta}$. The algorithm for high-order moment estimation is stated in Algorithm 2. In the exploration phase, the agent learns the environment with the help of two series of value functions $\widetilde{V}_{k,h}$ and $\widehat{V}_{k,h}$. They serve to characterize different aspects of the model, one for pseudo values and one for trajectory uncertainty. And thus, they rely on different estimations of transition kernel parameter $\boldsymbol{\theta}$. Two independent series of higher-order moment estimations are necessary for achieving accurate estimation. In the Algorithm 1, both estimations of $\boldsymbol{\theta}$ are the solutions to the weighted regression problem in the following form:

$$\begin{aligned}\operatorname*{argmin}_{\boldsymbol{\theta}} \Bigg(\lambda\|\boldsymbol{\theta}\|_2^2 \\ + \sum_{j=1}^{k-1}\sum_{h=1}^{H} \big(\boldsymbol{\phi}_{j,h,0}^{\top}\boldsymbol{\theta} - V_{j,h}(s_{h+1}^j)\big)^2/\bar{\sigma}_{j,h,0}^2 \Bigg),\end{aligned} \tag{4.3}$$

where the regression weight $\bar{\sigma}_{j,h,0}$ is set as Equation (4.4).

$$\begin{aligned}\overline{\sigma}_{k,h,0}^2 \leftarrow \max\Big\{\gamma^2 \left\|\boldsymbol{\phi}_{k,h,0}\right\|_{\boldsymbol{\Sigma}_{k,h,0}^{-1}}, \\ \left[\overline{\mathbb{V}}_{k,0}V_{k,h+1}\right](s_h^k,a_h^k) + E_{k,h,0}, \alpha^2\Big\}.\end{aligned} \tag{4.4}$$

$\bar{\sigma}_{j,h,0}$ can be considered as an combination of *aleatoric uncertainty* and *epistemic uncertainty* (Kendall & Gal, 2017; Mai et al., 2022). The first term $\gamma^2\left\|\boldsymbol{\phi}_{k,h,m}\right\|_{\boldsymbol{\Sigma}_{k,h,0}^{-1}}$ in (4.4) is the *epistemic uncertainty* caused by limited available data. And the second term in Equation (4.4) is supposed to be the *aleatoric uncertainty* $\mathbb{V}_{k,0}V_{k,h+1}$ characterizing the inherent non-determinism of the transition kernel, which is irreducible. Here the $\mathbb{V}_{k,m}V_{k,h+1}$ is the variance of $V_{k,h+1}$ to 2^m defined as $[\mathbb{P}V_{k,h+1}^{2^{m+1}}](s_h^k,a_h^k) - [\mathbb{P}V_{k,h+1}^{2^m}](s_h^k,a_h^k)^2$. Then, $\left[\mathbb{V}_{k,0}V_{k,h+1}\right](s_h^k,a_h^k)$ is further replaced with its estimate $\left[\overline{\mathbb{V}}_{k,0}V_{k,h+1}\right](s_h^k,a_h^k)$ plus its error bound $E_{k,h,0}$ since real variance $\left[\mathbb{V}_{k,0}V_{k,h+1}\right](s_h^k,a_h^k)$ is unknown to the agent. Because $\left[\mathbb{V}_{k,0}V_{k,h+1}\right](s_h^k,a_h^k)$ is a quadratic function of the real transition kernel parameter $\boldsymbol{\theta}^*$, its estimate can be achieved as

$$\begin{aligned}\left[\overline{\mathbb{V}}_{k,0}V_{k,h+1}\right](s_h^k,a_h^k) =& \left[\left\langle\boldsymbol{\phi}_{k,h,1},\boldsymbol{\theta}_{k,1}\right\rangle\right]_{[0,1]} \\ &- \left[\left\langle\widehat{\boldsymbol{\phi}}_{k,h,0},\boldsymbol{\theta}_{k,0}\right\rangle\right]_{[0,1]}^2,\end{aligned}$$

where $\boldsymbol{\theta}_{k,1}$ is again the solution to the weighted regression problem similar to (4.4) with predictors $\boldsymbol{\phi}_{k,h,1} = \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k)$, responses $V_{k,h+1}^2(s_{h+1}^k)$ and weight $\overline{\sigma}_{k,h,1}$. Following the above idea, the value of weight $\overline{\sigma}_{k,h,1}$ further relies on $\boldsymbol{\theta}_{k,2}$, which is the solution to a weighted regression problem involving another weight $\overline{\sigma}_{k,h,2}$. The algorithm carried out this process recursively until $\overline{\sigma}_{k,h,M-1}$, where its second term is replaced by the trivial upper bound of aleatoric uncertainty.

Applying HOME to the reward-free setting brings additional difficulties in controlling the error of our estimate for the model, as the error introduced by using the pseudo reward function instead of the real reward function and the error introduced by estimating the true transition kernel must be controlled separately. To address this problem, we carefully estimate variables indicating different kinds of error into two series of HOME in Line 10 and Line 11. Since the separation of variables deeply exploits the inner structure of the problem, the two series of HOME can be merged in the end to achieve a unified control for both kinds of error.

Previous work Chen et al. (2021) implemented the weighted value regression in a more crude way. The weights are

Algorithm 2 High-order moment estimator (HOME)

Input: Features $\{\boldsymbol{\phi}_{k,h,m}\}_{m\in\overline{[M]}}$, vector estimators $\{\boldsymbol{\theta}_{k,m}\}_{m\in\overline{[M]}}$, covariance matrix $\left\{\boldsymbol{\Sigma}_{k,h,m}, \dot{\boldsymbol{\Sigma}}_{k,m}\right\}_{m\in\overline{[M]}}$, confidence radius β_k, α, γ

1: **for** $m = 0, \ldots, M-2$ **do**

2: Set $\left[\overline{\mathbb{V}}_{k,m} V_{k,h+1}^{2^m}\right](s_h^k, a_h^k) \leftarrow \left[\boldsymbol{\phi}_{k,h,m+1}^\top \boldsymbol{\theta}_{k,m+1}\right]_{[0,1]} - \left[\boldsymbol{\phi}_{k,h,m}^\top \boldsymbol{\theta}_{k,m}\right]_{[0,1]}^2$

3: Set $E_{k,h,m} \leftarrow \left[2\beta_k \|\boldsymbol{\phi}_{k,h,m}\|_{\dot{\boldsymbol{\Sigma}}_{k,m}^{-1}}\right]_{[0,1]} + \left[\beta_k \|\boldsymbol{\phi}_{k,h,m+1}\|_{\dot{\boldsymbol{\Sigma}}_{k,m+1}^{-1}}\right]_{[0,1]}$

4: Set $\overline{\sigma}_{k,h,m}^2 \leftarrow \max\left\{\gamma^2 \|\boldsymbol{\phi}_{k,h,m}\|_{\boldsymbol{\Sigma}_{k,h,m}^{-1}}, \left[\overline{\mathbb{V}}_{k,m} V_{k,h+1}^{2^m}\right](s_h^k, a_h^k) + E_{k,h,m}, \alpha^2\right\}$

5: **end for**

6: Set $\overline{\sigma}_{k,h,M-1}^2 \leftarrow \max\left\{\gamma^2 \|\boldsymbol{\phi}_{k,h,M-1}\|_{\boldsymbol{\Sigma}_{k,h,M-1}^{-1}}, 1, \alpha^2\right\}$

Output: $\{\overline{\sigma}_{k,h,m}\}_{m\in\overline{[M]}}$

constructed only on aleatoric uncertainty, totally ignoring epistemic uncertainty. In addition, they use the same instead of different transition kernel parameters to calculate different order moments of the value function and stop target value regression at second order moment, which increased avoidable error. As a result, Chen et al. (2021) can only replace factor Hd with factor $H+d$ when trying to improve the dependency on d in the upper bound. In contrast, our work further improves factor $H+d$ to factor H through the well-designed target value regression, as we can see in Corollary 5.4.

High Confidence Set At the end of each episode, we add the following constraints into $\mathcal{U}_k$ to update the high confidence set in Line 20 of Algorithm 1.

$$\left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{k,m}\right\|_{\hat{\boldsymbol{\Sigma}}_{k,m}} \leq \beta_k,\ m \in \overline{[M]}, \tag{4.5}$$

$$\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_{k,m}\right\|_{\hat{\boldsymbol{\Sigma}}_{k,m}} \leq \beta_k,\ m \in \overline{[M]}, \tag{4.6}$$

High confidence set $\mathcal{U}_k$ ensures that the estimate $\boldsymbol{\theta}_k$ lies in a neighborhood of real transition kernel parameter $\boldsymbol{\theta}^*$. Here the algorithm adds $2M$ inequalities to constraints in each episode. These inequalities guarantee that estimations of the variance of $\widehat{V}_{k,h}$ and $\widetilde{V}_{k,h}$ up to M-th order are near the real values.

Planning Phase After finishing the exploration, the agent enters the planning phase and receives the real reward function. Depending on optimal Bellman equations, the agent is able to obtain the optimal policy backward from state H to state 1 by dynamic programming based on real reward function r and transition kernel parameter estimate $\boldsymbol{\theta}_K$. And then, the algorithm outputs the optimal policy.

Computational Complexity of HF-UCRL-RFE++ Similar with Chen et al. (2021), we assume that the optimization over $\boldsymbol{\theta}$, π, and r in Line 5 of Algorithm 1 can be accomplished with an oracle which is obvious to be called for K times. At each episode k and each stage h, **HF-UCRL-RFE++** computes $\{\widehat{\boldsymbol{\phi}}_{k,h,m}\}_{m\in\overline{[M]}}$, $\{\widetilde{\boldsymbol{\phi}}_{k,h,m}\}_{m\in\overline{[M]}}$, $\{\widehat{\sigma}_{k,h,m}\}_{m\in\overline{[M]}}$, $\{\widetilde{\sigma}_{k,h,m}\}_{m\in\overline{[M]}}$, and updates $\{\widehat{\boldsymbol{\Sigma}}_{k,h,m}\}_{m\in\overline{[M]}}$, $\{\widetilde{\boldsymbol{\Sigma}}_{k,h,m}\}_{m\in\overline{[M]}}$. The computation of $\{\widehat{\boldsymbol{\phi}}_{k,h,m}\}_{m\in\overline{[M]}}$ and $\{\widetilde{\boldsymbol{\phi}}_{k,h,m}\}_{m\in\overline{[M]}}$ require $O(\mathcal{O}M)$ times. According to Algorithm 2, calculating $\{\widehat{\sigma}_{k,h,m}\}_{m\in\overline{[M]}}$ and $\{\widetilde{\sigma}_{k,h,m}\}_{m\in\overline{[M]}}$ require $O(Md^2)$ time since the computation of the inner-product an inversion of matrix and a vector needs $O(d^2)$. The updates of $\{\widehat{\boldsymbol{\Sigma}}_{k,h,m}\}_{m\in\overline{[M]}}$ and $\{\widetilde{\boldsymbol{\Sigma}}_{k,h,m}\}_{m\in\overline{[M]}}$ further require $O(Md^2)$ time. Lastly, determining the optimal policy during the planning phase takes $O(H(SAd+\mathcal{O}))$ time. Therefore, the total time complexity of **HF-UCRL-RFE++** is $O(KH(\mathcal{O}M + Md^2) + HSAd)$.

5. Main Results

5.1. Upper Bounds

We first provide the suboptimality upper bound of our algorithm HF-UCRL-RFE++.

Theorem 5.1. For Algorithm 1, set $M = \log(7KH)/\log(2)$, $\alpha = H^{-1/2}$, $\gamma = d^{-1/4}$, $\lambda = d/B^2$, $\{\beta_k\}_{k\geq 1}$ as

$$\beta_k = 12\sqrt{d\eta\tau} + 30\tau/\gamma^2 + \sqrt{\lambda}B,$$

and denote $\beta = \beta_K$, where $\eta = \log(1 + kH/(\alpha^2 d\lambda))$ and $\tau = \log(32(\log(\gamma^2/\alpha)+1)k^2H^2/\delta)$. Then, for any $0 < \delta < 1$, we have with probability at least $1-\delta$, after collecting K episodes of samples, algorithm 1 returns a policy $\widehat{\pi}_r$ satisfying the following sub-optimality bound,

$$V_1^*(s_1; r) - V_1(s_1; \boldsymbol{\theta}^*, \widehat{\pi}_r, r) = \widetilde{O}\left(\frac{d^2}{K} + \frac{d}{\sqrt{K}}\right).$$

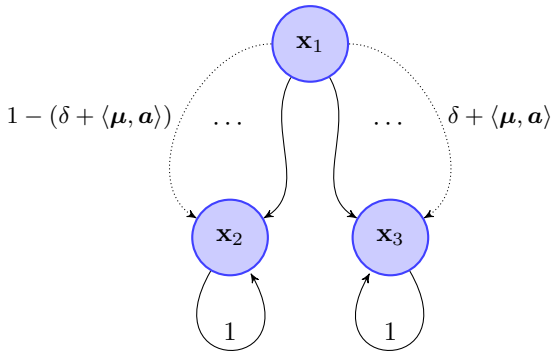

Figure 1. The transition kernel of the hard-to-learn linear mixture MDPs.

The next corollary specifies the sample complexity of our algorithm.

Corollary 5.2. Under the same conditions as in Theorem 5.1, Algorithm 1 has sample complexity of $m(\varepsilon,\delta)=\widetilde{O}(d^2\varepsilon^{-2})$ to output an ε-optimal policy in the planning phase. The exact expression of sample complexity is delayed to Appendix in Lemma A.11.

Remark 5.3. To the best of our knowledge, Corollary 5.2 provides the first horizon-free sample complexity upper bound independent of state space size S and action space size A for reward-free exploration. This result shows that long-horizon planning does not add extra difficulty to reward-free exploration.

Corollary 5.4. When rescaling the assumption $\sum_{h=1}^{H} r_h(s_h,a_h)\le 1$ to $\sum_{h=1}^{H} r_h(s_h,a_h)\le H$, under the same conditions as in Theorem 5.1, Algorithm 1 has sample complexity of $m(\varepsilon,\delta)=O(H^2d^2\varepsilon^{-2})$ to output an ε-optimal policy in the planning phase.

Remark 5.5. The assumption $\sum_{h=1}^{H} r_h(s_h,a_h)\le H$ covers the vanilla assumption $r_h(s_h,a_h)\in[0,1]$. Therefore, compared with Chen et al. (2021), our analysis does not require the $d>H$ assumption and achieves the same sample complexity bound up to logarithmic factors except for the trivial $\widetilde{O}(H)$ difference between time-homogeneous and time-inhomogeneous models with a milder assumption. This improvement can be attributed to the refined value target regression technique, high-order moment estimation (HOME), adopted in our approach. We provide a detailed analysis of this improvement in the "High-order Moment Estimation" part in the Section 4.

5.2. Lower Bounds

The following results provide lower bounds of the sample complexity and suggest that our algorithm is minimax optimal. We will consider the *hard-to-learn linear mixture MDPs* constructed in Zhou & Gu (2022). The state space is $\mathcal{S}=\{x_1,x_2,x_3\}$ and the action space is $\mathcal{A}=\{\mathbf{a}\}=\{-1,1\}^{d-1}$. The reward function satisfies $r(x_1,\cdot)=r(x_2,\cdot)=0$, and $r(x_3,\cdot)=\frac{1}{H}$. The transition probability is defined to be $\mathbb{P}(x_2\mid x_1,\boldsymbol{a})=1-(\delta+\langle\boldsymbol{\mu},\boldsymbol{a}\rangle)$ and $\mathbb{P}(x_3\mid x_1,\boldsymbol{a})=\delta+\langle\boldsymbol{\mu},\boldsymbol{a}\rangle$, where $\delta=1/6$ and $\boldsymbol{\mu}\in\{-\Delta,\Delta\}^{d-1}$ with $\Delta=\sqrt{\delta/K}/(4\sqrt{2})$.

Theorem 5.6. Suppose $B>1$. Then for any algorithm $\mathtt{ALG}_{Free}$ solving reward-free linear mixture MDP problems satisfying assumption 3.2, there exist a linear mixture MDP $\mathcal{M}$ such that $\mathtt{ALG}_{Free}$ needs to collect at least $\Omega\left(d^2\varepsilon^{-2}\right)$ episodes of samples to output an ε-optimal policy with probability at least $1-\delta$. This lower bound matches the sample complexity upper bound provided in Corollary 5.2, which shows our upper bound is optimal.

Remark 5.7. The lower bound is similar to the lower bound provided in Chen et al. (2021). The first difference is that we rescale the non-zero reward in hard-to-learn cases from 1 to $\frac{1}{H}$ in order to satisfy Assumption 3.2. The second difference is that we consider the time-homogeneous model instead of the time-inhomogeneous one in theirs. By these changes, our lower bound for reward-free exploration provided in Theorem 5.6 removes the unnecessary polynomial dependency on episode length H introduced by the scale of total reward.

Corollary 5.8. Under the same conditions as Theorem 5.6 except replacing $\sum_{h=1}^{H} r_h(s_h,a_h)\le 1$ with $r_h\in[0,1]$, for any algorithm $\mathtt{ALG}_{Free}$ solving reward-free linear mixture MDP problems satisfying assumption 3.2, there exist a linear mixture MDP $\mathcal{M}$ such that $\mathtt{ALG}_{Free}$ needs to collect at least $\widetilde{\Omega}\left(H^2d^2\varepsilon^{-2}\right)$ episodes to output an ε-optimal policy with probability at least $1-\delta$. This means our upper bound is optimal.

6. Proof Sketch of Theorem 5.1

We provide the proof sketch of Theorem 5.1 along with several key lemmas in big-O notation. The detail of these lemmas is restated in Appendix A. The following lemmas are conditioned on some good events.

Firstly, Lemma 6.1 controls the suboptimality gap between optimal value functions and our estimated value function in the planning phase with the uncertainty along trajectories.

Lemma 6.1. For any reward function r in the planning phase, the suboptimality gap of the outputted policy $\widehat{\pi}_r$ can be bounded as

$$V_1^*(s_1;r)-V_1(s_1;\boldsymbol{\theta}^*,\widehat{\pi}_r,r)\le 4\widehat{V}_{K,1}(s_1). \tag{6.1}$$

Then, the next lemma shows that the uncertainty along trajectories decreases with respect to episodes. This lemma is

intuitively right since the uncertainty should decrease with more information collected.

Lemma 6.2. For uncertainty along trajectories, we have

$$\widehat{V}_{K,1}(s_1;\boldsymbol{\theta}_K,\pi_K,r_K) \leq \frac{1}{K}\bigg(\sum_{k=1}^{K}\widehat{V}_{k,1}(s_1;\boldsymbol{\theta}_k,\pi_k,r_k)\bigg).$$

The last lemma upper bounds the sum of the uncertainty along trajectories.

Lemma 6.3. For any $0<\delta<1$, with probability at least $1-4M\delta$, we have

$$\sum_{k=1}^{K}\widehat{V}_{k,1}(s_1;\boldsymbol{\theta}_k,\widehat{\pi}_k,r_k) = \widetilde{O}(d\sqrt{K}+d^2). \tag{6.2}$$

Equipped with the above lemmas, we are ready to prove Theorem 5.1.

Proof of Theorem 5.1. The suboptimality of the policy output in the planning phase can be bounded by the uncertainty along trajectories in the last episode of the exploration phase as the following equation according to Lemma 6.1.

$$V_1^*(s_1;r) - V_1(s_1;\boldsymbol{\theta}^*,\widehat{\pi}_r,r) \leq 4\widehat{V}_{K,1}(s_1) \tag{6.3}$$

Since Lemma 6.2 indicates that the uncertainty is decreasing with the episodes, the uncertainty of the last episode can be further upper bounded by the sum of uncertainty in each episode by substituting (6.1) in Lemma 6.1 into the above inequality:

$$V_1^*(s_1;r) - V_1(s_1;\boldsymbol{\theta}^*,\widehat{\pi}_r,r) \leq \frac{4}{K}\sum_{k=1}^{K}\widehat{V}_{k,1}(s_1). \tag{6.4}$$

At last, the sum of value functions can be upper bounded according to Lemma 6.3. Thus, plugging (6.2) into the (6.4) as follows concludes our proof.

$$V_1^*(s_1;r) - V_1(s_1;\boldsymbol{\theta}^*,\widehat{\pi}_r,r) = \widetilde{O}\left(\frac{d^2}{K}+\frac{d}{\sqrt{K}}\right).$$

□

7. Conclusion

We study model-based reward-free exploration for learning the linear mixture MDPs. Our algorithm is guaranteed to have horizon-free sample complexity in the exploration phase to find a near-optimal policy in the planning phase for any given reward function. To our knowledge, these are the first horizon-free result for reward-free RL with function approximation. We also give a sample complexity lower bound for reward-free exploration in linear mixture MDPs under our assumptions. The sample complexity upper bound of our algorithm matches the lower bound up to logarithmic factors.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. WZ, JZ and QG are supported in part by the National Science Foundation CAREER Award 1906169 and research fund from UCLA-Amazon Science Hub. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. PMLR, 2017.

Agarwal, A., Jin, Y., and Zhang, T. Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation. *arXiv preprint arXiv:2212.06069*, 2022.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

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

Chen, X., Hu, J., Yang, L., and Wang, L. Near-optimal reward-free exploration for linear mixture mdps with plug-in solver. In *International Conference on Learning Representations*, 2021.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. On oracle-efficient pac rl with rich observations. *Advances in neural information processing systems*, 31, 2018.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

He, J., Zhou, D., and Gu, Q. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 4171–4180. PMLR, 2021.

He, J., Zhao, H., Zhou, D., and Gu, Q. Nearly minimax optimal reinforcement learning for linear markov decision processes. *arXiv preprint arXiv:2212.06132*, 2022.

Huang, J., Chen, J., Zhao, L., Qin, T., Jiang, N., and Liu, T.-Y. Towards deployment-efficient reinforcement learning: Lower bound and optimality. *arXiv preprint arXiv:2202.06450*, 2022.

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pp. 666–686. PMLR, 2020.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020a.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143, 2020b.

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pp. 865–891. PMLR, 2021a.

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pp. 865–891. PMLR, 2021b.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Li, Y., Wang, R., and Yang, L. F. Settling the horizon-dependence of sample complexity in reinforcement learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 965–976. IEEE, 2022.

Mai, V., Mani, K., and Paull, L. Sample efficient deep reinforcement learning via uncertainty estimation. *arXiv preprint arXiv:2201.01666*, 2022.

Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 7599–7608. PMLR, 2021.

Miryoosefi, S., Brantley, K., Daume III, H., Dudik, M., and Schapire, R. E. Reinforcement learning with convex constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Modi, A., Jiang, N., Tewari, A., and Singh, S. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 2010–2020, 2020.

Ren, T., Li, J., Dai, B., Du, S. S., and Sanghavi, S. Nearly horizon-free offline reinforcement learning. *Advances in neural information processing systems*, 34:15621–15634, 2021.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pp. 2898–2933. PMLR, 2019.

Tarbouriech, J., Zhou, R., Du, S. S., Pirotta, M., Valko, M., and Lazaric, A. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in Neural Information Processing Systems*, 34: 6843–6855, 2021.

Tessler, C., Mankowitz, D. J., and Mannor, S. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, pp. 22430–22456. PMLR, 2022.

Wang, R., Du, S. S., Yang, L. F., and Kakade, S. M. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*, 2020a.

Wang, R., Du, S. S., Yang, L. F., and Salakhutdinov, R. On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 2020b.

Wang, R., Salakhutdinov, R. R., and Yang, L. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020c.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2019.

Weisz, G., Amortila, P., and Szepesvári, C. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pp. 1237–1264. PMLR, 2021.

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

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

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1964, 2020a.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020b.

Zanette, A., Lazaric, A., Kochenderfer, M. J., and Brunskill, E. Provably efficient reward-agnostic navigation with linear value iteration. *Advances in Neural Information Processing Systems*, 2020c.

Zhang, W., Zhou, D., and Gu, Q. Reward-free model-based reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34, 2021a.

Zhang, Z., Du, S. S., and Ji, X. Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*, 2020.

Zhang, Z., Ji, X., and Du, S. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pp. 4528–4531. PMLR, 2021b.

Zhang, Z., Yang, J., Ji, X., and Du, S. S. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems*, 34, 2021c.

Zhang, Z., Zhou, Y., and Ji, X. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*, pp. 12653–12662. PMLR, 2021d.

Zhang, Z., Ji, X., and Du, S. Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory*, pp. 3858–3904. PMLR, 2022.

Zhou, D. and Gu, Q. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *arXiv preprint arXiv:2205.11507*, 2022.

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

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021b.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*. PMLR, 2021c.

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

A. Omitted Proof in Section 6

In this section, we provide detailed proof for Theorem 5.1. For $k \in [K]$, $h \in [H]$, let $\mathcal{F}_{k,h}$ be the σ-algebra generated by the random variables representing the state-action pairs up to and including those that appear stage h of episode k. That is, $\mathcal{F}_{k,h}$ is generated by

$$\begin{array}{cccccc} s_1^1, a_1^1 & \cdots & s_h^1, a_h^1 & \cdots & s_H^1, a_H^1 \\ s_1^2, a_1^2 & \cdots & s_h^2, a_h^2 & \cdots & s_H^2, a_H^2 \\ \vdots & & \vdots & & \\ s_1^k, a_1^k & \cdots & s_h^k, a_h^k. & & \end{array}$$

A.1. Notations

To establish clarity and facilitate understanding, we provide the Table 2 that outlines the notations which will be utilized throughout the proof.

Notation	Meaning
s_h, a_h	States and actions introduced by a general policy π (not specified).
s_h^k, a_h^k	States and actions introduced in the episode k by the policy π_k.
$u_{k,h}(s, a; \boldsymbol{\theta}, \pi, r)$	The uncertainty of states and actions, defined in Equation (4.1).
$\boldsymbol{\theta}_k, \quad \pi_k,$ $r_k = \{r_{k,h}\}_{h\in[H]}$	The transition kernel parameter, the exploration policy, and the pseudo reward function obtained via the optimization oracle in Line 5 of Algorithm 1.
$V_h(s; \boldsymbol{\theta}, \pi, r)$	The value function of policy π in the linear mixture MDP with transition kernel parameter $\boldsymbol{\theta}$ and reward function r.
$\widehat{V}_{k,h}(s; \boldsymbol{\theta}, \pi, r)$	The uncertainty along the trajectory, defined in Equation (4.2)
$\widetilde{V}_{k,h}(s)$	The pseudo value function, equal to $V_h(s; \boldsymbol{\theta}_k, \pi_k, r_k)$
$\widehat{\boldsymbol{\theta}}_{k,m}$	The estimated transition kernel parameter obtained by value regression targeting $\widehat{V}_{k,h}^{2^m}$.
$\widetilde{\boldsymbol{\theta}}_{k,m}$	The estimated transition kernel parameter obtained by value regression targeting $\widetilde{V}_{k,h}^{2^m}$.
$\boldsymbol{\theta}^*$	The ground-truth transition kernel parameter.
$\mathcal{U}_k$	The confidence set containing $\boldsymbol{\theta}^*$ with high probability.
$\beta_k, (\beta = \beta_K)$	The radius of confidence set $\mathcal{U}_k$.
$\widetilde{\Sigma}_{k,h,m}, \widehat{\Sigma}_{k,h,m}$	The covariance matrix for $\widetilde{V}_{k,h}^{2^m}$ and $\widehat{V}_{k,h}^{2^m}$, respectively.
$\dot{\widetilde{\Sigma}}_{k,m}, \dot{\widehat{\Sigma}}_{k,m}$	Equal to $\widetilde{\Sigma}_{k-1,H+1,m}$ and $\widehat{\Sigma}_{k-1,H+1,m}$, respectively.
$\widetilde{\sigma}_{k,h,m}, \widehat{\sigma}_{k,h,m}$	The weights for regression problems targeting $\widetilde{V}_{k,h}^{2^m}$ and $\widehat{V}_{k,h}^{2^m}$ respectively, defined in Equation (4.4).
$\widetilde{\boldsymbol{\phi}}_{k,h,m}, \widehat{\boldsymbol{\phi}}_{k,h,m}$	Equal to $\boldsymbol{\phi}_{\widehat{V}_{k,h+1}^{2^m}}(s_h^k, a_h^k)$ and $\boldsymbol{\phi}_{\widetilde{V}_{k,h+1}^{2^m}}(s_h^k, a_h^k)$, respectively.
$\widehat{\pi}_r$	The policy obtained in the planning phase.

Table 2. Important Notations

A.2. Proof of Lemma 6.1

Lemma A.1. For all $0 < \delta < 1$, suppose β_k is set as in Theorem 5.1, the following event happens with probability at least $1 - 2M\delta$

$$\left\|\widehat{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^*\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}} \leq \beta_k \tag{A.1}$$

$$\left\|\widetilde{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^*\right\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,m}} \leq \beta_k \tag{A.2}$$

$$\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}} \leq 2\beta_k \tag{A.3}$$

$$\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\dot{\tilde{\boldsymbol{\Sigma}}}_{k,0}} \leq 2\beta_k. \tag{A.4}$$

In the following, we define the event that Lemma A.1 holds to be $\mathcal{E}_{A.1}$. And the following lemmas are conditioned on $\mathcal{E}_{A.1}$ by default. We define function W_h for certain sequence $\{R_h\}$ recursively as

$$W_h\left(\{R_h\}\right) = \min\left\{1, R_h + W_{h+1}\left(\{R_h\}\right)\right\}.$$

In addition we denote the trajectory of first h steps as $\texttt{traj}_h := (s_1, a_1, \cdots, s_{h-1}, a_{h-1}, s_h)$, and the trajectory sampled from $(\pi, \mathbb{P})$ conditioned on $\texttt{traj}_h$ as $\texttt{traj} \sim (\pi, \mathbb{P})|\texttt{traj}_h$.

Lemma A.2. For any policy π and reward function $r \in R$, we have

$$V_1(s_1; \boldsymbol{\theta}_K, \pi, r) - V_1(s_1; \boldsymbol{\theta}^*, \pi, r) = \mathbb{E}_{\texttt{traj}\sim(\pi,\mathbb{P})|\texttt{traj}_1} W_1(\{(\mathbb{P}_K - \mathbb{P})V_{h+1}(s_h; \boldsymbol{\theta}_K, \pi, r)\}) \tag{A.5}$$

Lemma A.3. For any policy π and reward function $r \in R$, we have

$$\mathbb{E}_{\texttt{traj}\sim(\pi,\mathbb{P})|\texttt{traj}_1} W_1\left(\{u_{k,h}(s_h, \pi(s_h); \boldsymbol{\theta}_K, \pi, r)\}\right) \leq \widehat{V}_{K,1}\left(s_1; \boldsymbol{\theta}_K, \pi_K, r_K\right).$$

Proof of Lemma 6.1. The proof follows the proof of Lemma 15 in Zhang et al. (2020). Firstly,

$$\begin{aligned}
&V_1^*(s_1; r) - V_1(s_1; \boldsymbol{\theta}^*, \widehat{\pi}_r, r) \\
&= (V_1^*(s_1; r) - V_1(s_1; \boldsymbol{\theta}_K, \widehat{\pi}_r, r)) + (V_1(s_1; \boldsymbol{\theta}_K, \widehat{\pi}_r, r) - V_1(s_1; \boldsymbol{\theta}^*, \widehat{\pi}_r, r)) \\
&\leq (V_1^*(s_1; r) - V_1(s_1; \boldsymbol{\theta}_K, \pi_r^*, r)) + (V_1(s_1; \boldsymbol{\theta}_K, \widehat{\pi}_r, r) - V_1(s_1; \boldsymbol{\theta}^*, \widehat{\pi}_r, r)),
\end{aligned} \tag{A.6}$$

where π_r^* is the optimal policy for $(\boldsymbol{\theta}, r)$, and $\widehat{\pi}_r$ is the optimal policy for $(\boldsymbol{\theta}_K, r)$. Then for any policy $\pi \in \Pi$,

$$\begin{aligned}
&|V_1(s_1; \boldsymbol{\theta}_K, \pi, r) - V_1(s_1; \boldsymbol{\theta}^*, \pi, r)| \\
&= \left|\mathbb{E}_{\texttt{traj}\sim(\pi,\mathbb{P})|\texttt{traj}_1} W_1\left(\{(\mathbb{P}_K - \mathbb{P})V_{h+1}(s_h, a_h; \boldsymbol{\theta}_K, \pi, r)\}\right)\right| \\
&= \left|\mathbb{E}_{\texttt{traj}\sim(\pi,\mathbb{P})|\texttt{traj}_1} W_1\left(\left\{(\boldsymbol{\theta}_K - \boldsymbol{\theta}^*)\boldsymbol{\phi}_{V_{h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h, a_h)\right\}\right)\right| \\
&\leq \mathbb{E}_{\texttt{traj}\sim(\pi,\mathbb{P})|\texttt{traj}_1} W_1\left(\left\{\|\boldsymbol{\theta}_K - \boldsymbol{\theta}^*\|_{\dot{\tilde{\boldsymbol{\Sigma}}}_{k,0}} \left\|\boldsymbol{\phi}_{V_{h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h, a_h)\right\|_{\dot{\tilde{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\right\}\right) \\
&\leq \mathbb{E}_{\texttt{traj}\sim(\pi,\mathbb{P})|\texttt{traj}_1} W_1\left(\left\{2\beta \left\|\boldsymbol{\phi}_{V_{h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h, a_h)\right\|_{\dot{\tilde{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\right\}\right) \\
&= 2\mathbb{E}_{\texttt{traj}\sim(\pi,\mathbb{P})|\texttt{traj}_1} W_1\left(\{u_h(s_h, a_h; \boldsymbol{\theta}_K, \pi, r)\}\right) \\
&\leq 2\widehat{V}_{K,1}(s_1; \boldsymbol{\theta}_K, \pi_K, r_K).
\end{aligned} \tag{A.7}$$

The first equality holds due to Lemma A.2, the second inequality holds due to Cauchy-Schwartz inequality, the third inequality holds due to Lemma A.1, and the last inequality holds due to Lemma A.3. Plugging (A.7) into (A.7), we obtain

$$\begin{aligned}
V_1^*(s_1; r) - V_1(s_1; \boldsymbol{\theta}^*, \widehat{\pi}_r, r) &\leq 2\widehat{V}_{K,1}(s_1; \boldsymbol{\theta}_K, \pi_K, r_K) + 2\widehat{V}_{K,1}(s_1; \boldsymbol{\theta}_K, \pi_K, r_K) \\
&= 4\widehat{V}_{K,1}(s_1; \boldsymbol{\theta}_K, \pi_K, r_K).
\end{aligned}$$

□

A.3. Proof of Lemma 6.2

Proof of Lemma 6.2. The proof follows the proof of Lemma 14 in Chen et al. (2021). Firstly, we prove that $\widehat{V}_{k,1}(s; \boldsymbol{\theta}, \pi, r)$ is non-increasing w.r.t. k for any fixed $\boldsymbol{\theta}, \pi, r$ by induction in h. Suppose for any $k_1 \leq k_2$, $\widehat{V}_{k_1,h+1}(s; \boldsymbol{\theta}, \pi, r) \geq \widehat{V}_{k_2,h+1}(s; \boldsymbol{\theta}, \pi, r)$ for any s. By definition,

$$\begin{aligned}
\widehat{V}_{k,h}(s; \boldsymbol{\theta}, \pi, r) = \min\Bigg\{1, u_{k,h}(s, a; \boldsymbol{\theta}, \pi, r) + 2\beta \Big\|\boldsymbol{\phi}_{\widehat{V}_{k,h+1}(\cdot;\boldsymbol{\theta},\pi,r)}(s, \pi(s))\Big\|_{\dot{\tilde{\boldsymbol{\Sigma}}}_{k,0}^{-1}} \\
+ \boldsymbol{\phi}_{\widehat{V}_{k,h+1}(\cdot;\boldsymbol{\theta},\pi,r)}^{\top}(s, \pi(s))\boldsymbol{\theta}\Bigg\}
\end{aligned}$$

$$u_{k,h}(s,a;\boldsymbol{\theta},\pi,r)=\beta\left\|\boldsymbol{\phi}_{V_h(\cdot;\boldsymbol{\theta},\pi,r)}(s,a)\right\|_{\dot{\tilde{\boldsymbol{\Sigma}}}_{k,0}^{-1}}$$

Since $\dot{\widehat{\boldsymbol{\Sigma}}}_{k_1,0}\preceq\dot{\widehat{\boldsymbol{\Sigma}}}_{k_2,0}$ and $\dot{\tilde{\boldsymbol{\Sigma}}}_{k_1,0}\preceq\dot{\tilde{\boldsymbol{\Sigma}}}_{k_2,0}$, we have

$$\begin{aligned}
u_{k_1,h}(s,a;\boldsymbol{\theta},\pi,r)&\geq u_{k_2,h}(s,a;\boldsymbol{\theta},\pi,r)\\
\left\|\boldsymbol{\phi}_{\widehat{V}_{k_1,h+1}(\cdot;\boldsymbol{\theta},\pi,r)}(s,\pi(s))\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}^{-1}}&\geq\left\|\boldsymbol{\phi}_{\widehat{V}_{k_2,h+1}(\cdot;\boldsymbol{\theta},\pi,r)}(s,\pi(s))\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\\
\boldsymbol{\phi}_{\widehat{V}_{k_1,h+1}(\cdot;\boldsymbol{\theta},\pi,r)}^{\top}(s,\pi(s))\boldsymbol{\theta}&\geq\boldsymbol{\phi}_{\widehat{V}_{k_2,h+1}(\cdot;\boldsymbol{\theta},\pi,r)}^{\top}(s,\pi(s))\boldsymbol{\theta}
\end{aligned}$$

Thus $\widehat{V}_{k_1,h}(s;\boldsymbol{\theta},\pi,r)\geq\widehat{V}_{k_2,h}(s;\boldsymbol{\theta},\pi,r)$ for any $k_1\leq k_2$. Furthermore, since $\mathcal{U}_{k_2}\subset\mathcal{U}_{k_1}$, and $\boldsymbol{\theta}_k,\pi_k,r_k$ are argmax over $\mathcal{U}_k$, we have

$$\widehat{V}_{k_1,1}(s_1;\boldsymbol{\theta}_{k_1},\pi_{k_1},r_{k_1})\geq\widehat{V}_{k_1,1}(s_1;\boldsymbol{\theta}_{k_2},\pi_{k_2},r_{k_2})\geq\widehat{V}_{k_2,1}(s_1;\boldsymbol{\theta}_{k_2},\pi_{k_2},r_{k_2})$$

It follows that $\widehat{V}_{k,1}(s_1^k;\boldsymbol{\theta}_k,\pi_k,r_k)$ is non-increasing w.r.t. k. Thus,

$$K\widehat{V}_{K,1}(s_1;\boldsymbol{\theta}_K,\pi_K,r_K)\leq\sum_{k=1}^{K}\widehat{V}_{k,1}(s_1;\boldsymbol{\theta}_k,\pi_k,r_k)$$

□

A.4. Proof of Lemma 6.3

Lemma A.4. Conditioned on the event $\mathcal{E}$, let $\widetilde{V}_{k,h}$, $\widehat{V}_{k,h}$, $\dot{\tilde{\boldsymbol{\Sigma}}}_{k,m}$, $\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}$, $\widetilde{\boldsymbol{\phi}}_{k,h,m}$, $\widehat{\boldsymbol{\phi}}_{k,h,m}$ be defined in Algorithm 1, for any $k\in[K]$, $h\in[H]$, $m\in\overline{[M]}$, we have

$$\widehat{V}_{k,h}(s_h^k)-u_{k,h}(s_h^k,a_h^k)-\mathbb{P}\widehat{V}_{k,h+1}(s_h^k,a_h^k)\leq 4\min\left\{1,\beta\left\|\widehat{\boldsymbol{\phi}}_{k,h,0}\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\right\}\tag{A.8}$$

$$\widetilde{V}_{k,h}(s_h^k)-r_{k,h}(s_h^k,a_h^k)-\mathbb{P}\widetilde{V}_{k,h+1}\leq 2\min\left\{1,\beta\left\|\widetilde{\boldsymbol{\phi}}_{k,h,0}\right\|_{\dot{\tilde{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\right\}\tag{A.9}$$

In order to prove Lemma 6.3, we introduce the following quantities used in Zhou & Gu (2022) as

$$\widehat{R}_m=\sum_{k=1}^{K}\sum_{h=1}^{H}I_h^j\min\left\{1,\beta\|\widehat{\boldsymbol{\phi}}_{k,h,m}\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}^{-1}}\right\},\forall m\in\overline{[M]}\tag{A.10}$$

$$\widetilde{R}_m=\sum_{k=1}^{K}\sum_{h=1}^{H}I_h^j\min\left\{1,\beta\|\widetilde{\boldsymbol{\phi}}_{k,h,m}\|_{\dot{\tilde{\boldsymbol{\Sigma}}}_{k,m}^{-1}}\right\},\forall m\in\overline{[M]}\tag{A.11}$$

$$\widehat{A}_m=\sum_{k=1}^{K}\sum_{h=1}^{H}I_h^k\left[\left[\mathbb{P}\widehat{V}_{k,h+1}^{2^m}\right](s_h^k,a_h^k)-\widehat{V}_{k,h+1}^{2^m}(s_{h+1}^k)\right],\forall m\in\overline{[M]}\tag{A.12}$$

$$\widetilde{A}_m=\sum_{k=1}^{K}\sum_{h=1}^{H}I_h^k\left[\left[\mathbb{P}\widetilde{V}_{k,h+1}^{2^m}\right](s_h^k,a_h^k)-\widetilde{V}_{k,h+1}^{2^m}(s_{h+1}^k)\right],\forall m\in\overline{[M]}\tag{A.13}$$

$$\widehat{S}_m=\sum_{k=1}^{K}\sum_{h=1}^{H}I_h^k\left[\mathbb{V}\widehat{V}_{k,h+1}^{2^m}\right](s_h^k,a_h^k),\forall m\in\overline{[M]}\tag{A.14}$$

$$\widetilde{S}_m=\sum_{k=1}^{K}\sum_{h=1}^{H}I_h^k\left[\mathbb{V}\widetilde{V}_{k,h+1}^{2^m}\right](s_h^k,a_h^k),\forall m\in\overline{[M]}\tag{A.15}$$

$$I_h^k=\mathbb{1}\left\{\forall m\in\overline{[M]},\det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}^{-1/2}\right)/\det\left(\widehat{\boldsymbol{\Sigma}}_{k,h,m}^{-1/2}\right)\leq 4\text{ and }\det\left(\dot{\tilde{\boldsymbol{\Sigma}}}_{k,m}^{-1/2}\right)/\det\left(\widetilde{\boldsymbol{\Sigma}}_{k,h,m}^{-1/2}\right)\leq 4\right\}\tag{A.16}$$

$$G=\sum_{k=1}^{K}\left(1-I_H^k\right),\tag{A.17}$$

Lemma A.5. Let γ, α, be defined in Algorithm 2, $\{\widehat{R}_m\}_{m\in\overline{[M]}}$, $\{\widetilde{R}_m\}_{m\in\overline{[M]}}$, $\{\widehat{S}_m\}_{m\in\overline{[M]}}$, $\{\widetilde{S}_m\}_{m\in\overline{[M]}}$ be defined in (A.10), (A.11), (A.14), (A.15). Then for $m \in \overline{[M-1]}$, we have

$$\widehat{R}_m \le \min\left\{KH, 4d\iota + 4\beta\gamma^2 d\iota + 2\beta\sqrt{d\iota}\sqrt{\widehat{S}_m + 4\widehat{R}_m + 2\widehat{R}_{m+1} + KH\alpha^2}\right\} \tag{A.18}$$

$$\widetilde{R}_m \le \min\left\{KH, 4d\iota + 4\beta\gamma^2 d\iota + 2\beta\sqrt{d\iota}\sqrt{\widetilde{S}_m + 4\widetilde{R}_m + 2\widetilde{R}_{m+1} + KH\alpha^2}\right\}, \tag{A.19}$$

where $\iota = \log(1 + KH/(d\lambda\alpha^2))$. For $\widehat{R}_{M-1}$ and $\widetilde{R}_{M-1}$, we have the trivial bound $\widehat{R}_{M-1} \le KH$ and $\widetilde{R}_{M-1} \le KH$.

Lemma A.6. Let $\{\widehat{R}_m\}_{m\in\overline{[M]}}$, $\{\widetilde{R}_m\}_{m\in\overline{[M]}}$, $\{\widehat{S}_m\}_{m\in\overline{[M]}}$, $\{\widetilde{S}_m\}_{m\in\overline{[M]}}$, $\{\widehat{A}_m\}_{m\in\overline{[M]}}$, $\{\widetilde{A}_m\}_{m\in\overline{[M]}}$, G be defined as (A.10), (A.11), (A.14), (A.15), (A.12), (A.13), (A.17). Then, conditioned on the event $\mathcal{E}$, for $m \in \overline{[M-1]}$, we have

$$\widehat{S}_m \le \left|\widehat{A}_{m+1}\right| + G + 2^{m+1}\left(\widetilde{R}_0 + 4\widehat{R}_0\right) \tag{A.20}$$

$$\widetilde{S}_m \le \left|\widetilde{A}_{m+1}\right| + G + 2^{m+1}\left(K + 2\widetilde{R}_0\right) \tag{A.21}$$

Lemma A.7. Let $\{\widehat{S}_m\}_{m\in\overline{[M]}}$, $\{\widetilde{S}_m\}_{m\in\overline{[M]}}$, $\{\widehat{A}_m\}_{m\in\overline{[M]}}$, $\{\widetilde{A}_m\}_{m\in\overline{[M]}}$ be defined as (A.14), (A.15), (A.12), (A.13). Then we have $\mathbb{P}(\mathcal{E}_{A.7}) > 1 - 2M\delta$, with $\mathcal{E}_{A.7}$ be defined as,

$$\mathcal{E}_{A.7} := \left\{\forall m \in \overline{[M]}, \left|\widehat{A}_m\right| \le \min\left\{\sqrt{2\zeta\widehat{S}_m} + \zeta, KH\right\} \text{ and } \left|\widetilde{A}_m\right| \le \min\left\{\sqrt{2\zeta\widetilde{S}_m} + \zeta, KH\right\}\right\}, \tag{A.22}$$

where $\zeta = 4\log(4\log(KH)/\delta)$.

Lemma A.8. Let G be defined in (A.17). Then we have

$$G \le Md\iota, \tag{A.23}$$

where $\iota = \log\left(1 + KH/\left(d\lambda\alpha^2\right)\right)$.

Lemma A.9. *(Restatement of Lemma 6.3)* For any $0 < \delta < 1$, with probability at least $1 - 4M\delta$, we have

$$\begin{aligned}
&\sum_{k=1}^{K} \widehat{V}_{k,1}(s_1^k; \boldsymbol{\theta}_k, \pi_k, r_k) \\
&\le 896\max\left\{64\beta^2 d\iota, 2\zeta\right\} + 24\zeta + 240d\iota + 240\beta\gamma^2 d\iota + 120\beta d\iota\sqrt{M} + 24\sqrt{\zeta M d\iota} + Md\iota \\
&\quad + \left(64\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\} + 120\beta\sqrt{d\iota H\alpha^2}\right)\sqrt{K}
\end{aligned}$$

where $\iota = \log(1 + KH/(d\lambda\alpha^2))$, $\zeta = 4\log(4\log(KH)/\delta)$.

Proof of Lemma A.9. All the following proofs are conditioned on $\mathcal{E}_{A.1} \cap \mathcal{E}_{A.7}$, which happens with probability at least $1 - 4M\delta$. Firstly, we have

$$\begin{aligned}
&\sum_{k=1}^{K} \widehat{V}_{k,1}(s_h^k) \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\left[I_h^k\left[\widehat{V}_{k,h}(s_h^k) - \widehat{V}_{k,h+1}(s_{h+1}^k)\right] + \left(1 - I_h^k\right)\left[\widehat{V}_{k,h}(s_h^k) - \widehat{V}_{k,h+1}(s_{h+1}^k)\right]\right] \\
&= \sum_{k=1}^{K}\left[\sum_{h=1}^{H} I_h^k u_{k,h}(s_h^k, a_h^k) + \sum_{h=1}^{H} I_h^k\left[\widehat{V}_{h,k}(s_h^k) - u_{k,h}(s_h^k, a_h^k) - \mathbb{P}\widehat{V}_{k,h+1}(s_h^k, a_h^k)\right]\right. \\
&\quad \left. + \sum_{h=1}^{H} I_h^k\left[\mathbb{P}\widehat{V}_{k,h+1}(s_h^k, a_h^k) - \widehat{V}_{k,h+1}(s_{h+1}^k)\right]\right] + \sum_{k=1}^{K}\sum_{h=1}^{H}(1 - I_h^k)\left[\widehat{V}_{k,h}(s_h^k) - \widehat{V}_{k,h+1}(s_{h+1}^k)\right]
\end{aligned}$$

$$\leq \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k u_{k,h}(s_h^k,a_h^k)}_{I_1} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\widehat{V}_{h,k}(s_h^k) - u_{k,h}(s_h^k,a_h^k) - \mathbb{P}\widehat{V}_{k,h+1}(s_h^k,a_h^k)\right]}_{I_2}$$
$$+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\mathbb{P}\widehat{V}_{k,h+1}(s_h^k,a_h^k) - \widehat{V}_{h+1,k}(s_{h+1}^k)\right]}_{I_3} + \underbrace{\sum_{k=1}^{K}\left(1-I_{h_k}^k\right)\widehat{V}_{k,h_k}(s_{h_k}^k)}_{I_4},$$

where h_k is the smallest index such that $I_{h_k}^k = 0$. Following the definition of $u_{k,h}$,

$$I_1 = \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \min\left\{1, \beta\left\|\widetilde{\boldsymbol{\phi}}_{k,h,0}\right\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\right\} = \widetilde{R}_0.$$

By Lemma A.4,

$$I_2 \leq 4\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \min\left\{1, \beta\left\|\widehat{\boldsymbol{\phi}}_{k,h,0}\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\right\} = 4\widehat{R}_0$$

By definitions,

$$I_3 = \widehat{A}_0,$$
$$I_4 \leq \sum_{k=1}^{K}\left(1-I_H^k\right) = G.$$

Thus,

$$\sum_{k=1}^{K}\widehat{V}_{k,1}(s_h^k) \leq \widetilde{R}_0 + 4\widehat{R}_0 + \widehat{A}_0 + G \tag{A.24}$$

Substituting (A.20) in Lemma A.6 into (A.18) in Lemma A.5, we have

$$\begin{aligned}\widehat{R}_m &\leq 4d\iota + 4\beta\gamma^2 d\iota + 2\beta\sqrt{d\iota}\sqrt{\left|\widehat{A}_{m+1}\right| + G + 2^{m+1}\left(\widetilde{R}_0 + 4\widehat{R}_0\right) + 4\widehat{R}_m + 2\widehat{R}_{m+1} + KH\alpha^2} \\ &\leq 2\beta\sqrt{d\iota}\sqrt{\left|\widehat{A}_{m+1}\right| + 2^{m+1}\left(\widetilde{R}_0 + 4\widehat{R}_0\right) + 4\widehat{R}_m + 2\widehat{R}_{m+1}} + \underbrace{4d\iota + 4\beta\gamma^2 d\iota + 2\beta\sqrt{d\iota}\sqrt{G + KH\alpha^2}}_{I_c},\end{aligned} \tag{A.25}$$

where the second inequality holds due to $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Substituting (A.20) in Lemma A.6 into (A.22) in Lemma A.7, we have

$$\begin{aligned}\left|\widehat{A}_m\right| &\leq \sqrt{2\zeta}\sqrt{\left|\widehat{A}_{m+1}\right| + G + 2^{m+1}\left(\widetilde{R}_0 + 4\widehat{R}_0\right)} + \zeta \\ &\leq \sqrt{2\zeta}\sqrt{\left|\widehat{A}_{m+1}\right| + 2^{m+1}\left(\widetilde{R}_0 + 4\widehat{R}_0\right)} + \sqrt{2\zeta G} + \zeta\end{aligned} \tag{A.26}$$

Substituting (A.21) in Lemma A.6 into (A.19) in Lemma A.5, we have

$$\begin{aligned}\widetilde{R}_m &\leq 4d\iota + 4\beta\gamma^2 d\iota + 2\beta\sqrt{d\iota}\sqrt{\left|\widetilde{A}_{m+1}\right| + G + 2^{m+1}\left(K + 2\widetilde{R}_0\right) + 4\widetilde{R}_m + 2\widetilde{R}_{m+1} + KH\alpha^2} \\ &\leq 2\beta\sqrt{d\iota}\sqrt{\left|\widetilde{A}_{m+1}\right| + 2^{m+1}\left(K + 2\widetilde{R}_0\right) + 4\widetilde{R}_m + 2\widetilde{R}_{m+1}} + \underbrace{4d\iota + 4\beta\gamma^2 d\iota + 2\beta\sqrt{d\iota}\sqrt{G + KH\alpha^2}}_{I_c}\end{aligned} \tag{A.27}$$

Substituting (A.21) in Lemma A.6 into (A.22) in Lemma A.7, we have

$$\begin{aligned}\left|\widetilde{A}_m\right| &\leq \sqrt{2\zeta}\sqrt{\left|\widetilde{A}_{m+1}\right| + G + 2^{m+1}\left(K + 2\widetilde{R}_0\right)} + \zeta \\ &\leq \sqrt{2\zeta}\sqrt{\left|\widetilde{A}_{m+1}\right| + 2^{m+1}\left(K + 2\widetilde{R}_0\right)} + \sqrt{2\zeta G} + \zeta \end{aligned} \tag{A.28}$$

Thus, calculating (A.27) + (A.28) + 4×(A.25) + (A.26) and using $\sqrt{a}+\sqrt{b}+\sqrt{c}+\sqrt{d} \leq 2\sqrt{a+b+c+d}$, we have

$$\begin{aligned}&\widetilde{R}_m + \left|\widetilde{A}_m\right| + 4\widehat{R}_m + \left|\widehat{A}_m\right| \\ &\leq 5I_c + 2\sqrt{2\zeta G} + 2\zeta + 2\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\}\sqrt{2\left|\widehat{A}_{m+1}\right| + 2\cdot 2^{m+1}\left(\widetilde{R}_0 + 4\widehat{R}_0\right)} \\ &\quad \overline{+4\widehat{R}_m + 2\widehat{R}_{m+1} + 2\left|\widetilde{A}_{m+1}\right| + 2\cdot 2^{m+1}\left(K + 2\widetilde{R}_0\right) + 4\widetilde{R}_m + 2\widetilde{R}_{m+1}} \\ &\leq 5I_c + +2\sqrt{2\zeta G} + 2\zeta + 4\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\}\sqrt{\left(\widetilde{R}_m + \left|\widetilde{A}_m\right| + 4\widehat{R}_m + \left|\widehat{A}_m\right|\right)} \\ &\quad \overline{+\left(\widetilde{R}_{m+1} + \left|\widetilde{A}_{m+1}\right| + 4\widehat{R}_{m+1} + \left|\widehat{A}_{m+1}\right|\right) + 2\cdot 2^{m+1}\left(K + \widetilde{R}_0 + \left|\widetilde{A}_0\right| + 4\widehat{R}_0 + \left|\widehat{A}_0\right|\right)}.\end{aligned}$$

Then by Lemma D.3 with $a_m = \widetilde{R}_m + \left|\widetilde{A}_m\right| + 4\widehat{R}_m + \left|\widehat{A}_m\right| \leq 7KH$ and $M = \log(7KH)/\log 2$, $\widetilde{R}_0 + \left|\widetilde{A}_0\right| + 4\widehat{R}_0 + \left|\widehat{A}_0\right|$ can be bounded as

$$\begin{aligned}&\widetilde{R}_0 + \left|\widetilde{A}_0\right| + 4\widehat{R}_0 + \left|\widehat{A}_0\right| \\ &\leq 22\cdot 16\max\{64\beta^2 d\iota, 2\zeta\} + 30I_c + 12\sqrt{\zeta G} + 12\zeta \\ &\quad + 32\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\}\sqrt{K + \widetilde{R}_0 + \left|\widetilde{A}_0\right| + 4\widehat{R}_0 + \left|\widehat{A}_0\right|} \\ &\leq 352\max\left\{64\beta^2 d\iota, 2\zeta\right\} + 30I_c + 12\sqrt{\zeta G} + 12\zeta + 32\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\}\sqrt{K} \\ &\quad + 32\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\}\sqrt{\widetilde{R}_0 + \left|\widetilde{A}_0\right| + 4\widehat{R}_0 + \left|\widehat{A}_0\right|}.\end{aligned} \tag{A.29}$$

By the fact that $x \leq a\sqrt{x} + b \Rightarrow x \leq 2a^2 + 2b$, (A.29) implies that

$$\begin{aligned}&\widetilde{R}_0 + \left|\widetilde{A}_0\right| + 4\widehat{R}_0 + \left|\widehat{A}_0\right| \\ &\leq 896\max\left\{64\beta^2 d\iota, 2\zeta\right\} + 60I_c + 24\sqrt{\zeta G} + 24\zeta + 64\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\}\sqrt{K}.\end{aligned} \tag{A.30}$$

Finally, plugging (A.30) into (A.24) and bounding G with Lemma A.8, we have

$$\begin{aligned}&\sum_{k=1}^{K}\widehat{V}_{k,1}(s_h^k) \\ &\leq \widetilde{R}_0 + \left|\widetilde{A}_0\right| + 4\widehat{R}_0 + \left|\widehat{A}_0\right| + G \\ &\leq 896\max\left\{64\beta^2 d\iota, 2\zeta\right\} + 24\zeta + 64\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\}\sqrt{K} \\ &\quad + 60\left(4d\iota + 4\beta\gamma^2 d\iota + 2\beta\sqrt{d\iota}\sqrt{Md\iota + KH\alpha^2}\right) + 24\sqrt{\zeta M d\iota} + Md\iota \\ &\leq 896\max\left\{64\beta^2 d\iota, 2\zeta\right\} + 24\zeta + 240d\iota + 240\beta\gamma^2 d\iota + 120\beta d\iota\sqrt{M} + 24\sqrt{\zeta M d\iota} + Md\iota \\ &\quad + \left(64\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\} + 120\beta\sqrt{d\iota H\alpha^2}\right)\sqrt{K}\end{aligned} \tag{A.31}$$

□

A.5. Proof of Main Results

Lemma A.10. *(Restatement of Theorem 5.1)* For Algorithm 1, set $M = \log(7KH)/\log(2)$, $\{\beta_k\}_{k\geq 1}$ as

$$\begin{aligned}\beta_k =& 12\sqrt{d\log(1+kH/(\alpha^2 d\lambda))\log(32(\log(\gamma^2/\alpha)+1)k^2H^2/\delta)} \\ &+ 30\log(32(\log(\gamma^2/\alpha)+1)k^2H^2/\delta)/\gamma^2 + \sqrt{\lambda}B,\end{aligned}$$

and denote $\beta = \beta_K$, then for any $0 < \delta' < 1$, we have with probability at least $1-\delta'$, where $\delta' = 4M\delta$, after collecting K trajectories, algorithm 1 returns a policy satisfying the following sub-optimality bound,

$$\begin{aligned}&V_1^*(s_1;r) - V_1(s_1;\boldsymbol{\theta}^*,\widehat{\pi}_r,r) \\ &\leq \frac{4}{K}\left(2752\max\left\{64\beta^2 d\iota, 2\zeta\right\} + 24\zeta + 240d\iota + 240\beta\gamma^2 d\iota + 120\beta d\iota\sqrt{M}\right) \\ &\quad + \frac{4}{\sqrt{K}}\left(64\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\} + 120\beta\sqrt{d\iota}\right),\end{aligned}$$

where $\iota = \log(1+KH/(d\lambda\alpha^2))$, $\zeta = 4\log(4\log(KH)/\delta)$. Moreover, setting $\alpha = H^{-1/2}$, $\gamma = d^{-1/4}$, and $\lambda = d/B^2$, we have the horizon-free suboptimality bound

$$V_1^*(s_1;r) - V_1(s_1;\boldsymbol{\theta}^*,\widehat{\pi}_r,r) = \widetilde{O}\left(\frac{d^2}{K} + \frac{d}{\sqrt{K}}\right). \tag{A.32}$$

Proof of Theorem A.10. The following proof is conditioned on $\mathcal{E}_{A.1}\cap\mathcal{E}_{A.7}$, which holds with probability at least $1-4M\delta = 1-\delta'$. We have

$$\begin{aligned}&V_1^*(s_1;r) - V_1(s_1;\boldsymbol{\theta}^*,\widehat{\pi}_r,r) \\ &\leq 4\widehat{V}_{K,1}\left(s_1;\boldsymbol{\theta}_K,\widehat{\pi}_K,r_K\right) \\ &\leq \frac{4}{K}\sum_{k=1}^{K} V_{k,1}(s_1;\boldsymbol{\theta}_k,\widehat{\pi}_k,r_k) \\ &\leq \frac{4}{K}\left(896\max\left\{64\beta^2 d\iota, 2\zeta\right\} + 24\zeta + 240d\iota + 240\beta\gamma^2 d\iota + 120\beta d\iota\sqrt{M} + 24\sqrt{\zeta M d\iota} + Md\iota\right) \\ &\quad + \frac{4}{\sqrt{K}}\left(64\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\} + 120\beta\sqrt{d\iota H\alpha^2}\right),\end{aligned}$$

where the first inequality holds due to Lemma 6.1, the second inequality holds due to Lemma 6.2, and the third equality holds due to Lemma A.9. □

Given the regret bound provided in Lemma A.10, we can prove the following sample complexity upper bound.

Lemma A.11. (Restatement of Corollary 5.2) Under the same conditions as in Theorem A.10, Algorithm 1 has sample complexity of

$$\begin{aligned}m(\varepsilon,\delta') =& \frac{16}{\varepsilon^2}\left(64\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\} + 120\beta\sqrt{d\iota H\alpha^2}\right)^2 \\ &+ \frac{8}{\varepsilon}\left(2752\max\left\{64\beta^2 d\iota, 2\zeta\right\} + 24\zeta + 240d\iota + 240\beta\gamma^2 d\iota + 120\beta d\iota\sqrt{M}\right)\end{aligned} \tag{A.33}$$

Moreover, setting $\alpha = H^{-1/2}$, $\gamma = d^{-1/4}$, and $\lambda = d/B^2$, we have the horizon-free sample complexity bound

$$m(\varepsilon,\delta') = \widetilde{O}\left(\frac{d^2}{\varepsilon^2}\right).$$

Proof of Lemma A.11. (A.33) is derived directly from Lemma A.10 by setting the suboptimality to ε and solving the K. □

Lemma A.12 (Restatement of Corollary 5.4)**.** When rescaling the assumption $\sum_{h=1}^{H} r_h(s_h, a_h) \le 1$ to $\sum_{h=1}^{H} r_h(s_h, a_h) \le H$, under the same conditions as Lemma A.10, Algorithm 1 has sample complexity of

$$\begin{aligned} m(\varepsilon, \delta') =& \frac{16H^2}{\varepsilon^2}\left(64\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\} + 120\beta\sqrt{d\iota H\alpha^2}\right)^2 \\ &+ \frac{8H}{\varepsilon}\left(2752\max\left\{64\beta^2 d\iota, 2\zeta\right\} + 24\zeta + 240d\iota + 240\beta\gamma^2 d\iota + 120\beta d\iota\sqrt{M}\right) \end{aligned} \tag{A.34}$$

Moreover, setting $\alpha = H^{-1/2}$, $\gamma = d^{-1/4}$, and $\lambda = d/B^2$, we have the horizon-free sample complexity bound

$$m(\varepsilon, \delta') = \widetilde{O}\left(\frac{H^2 d^2}{\varepsilon^2}\right).$$

Proof of Lemma A.12. (A.34) is a direct result of Lemma A.11. Let $r'_h(s_h, a_h) = r_h(s_h, a_h)/H$, then $\sum_{h=1}^{H} r'_h(s_h, a_h) \le 1$. Thus the sample complexity of achieving an ε/H-optimal policy for reward r'_h with probability $1-\delta'$ is

$$\begin{aligned} m(\varepsilon, \delta') =& \frac{16H^2}{\varepsilon^2}\left(64\max\left\{8\beta\sqrt{d\iota}, \sqrt{2\zeta}\right\} + 120\beta\sqrt{d\iota H\alpha^2}\right)^2 \\ &+ \frac{8H}{\varepsilon}\left(2752\max\left\{64\beta^2 d\iota, 2\zeta\right\} + 24\zeta + 240d\iota + 240\beta\gamma^2 d\iota + 120\beta d\iota\sqrt{M}\right). \end{aligned} \tag{A.35}$$

Since $r_h(s_h, a_h) = Hr'_h(s_h, a_h)$, for the same policy, the suboptimality for rewards r_h is H times the suboptimality for rewards r'_h. Thus, the ε/H-optimal policy for r'_h is a ε-optimal policy for r_h. Therefore, the sample complexity of achieving an ε-optimal policy for reward r_h with probability $1-\delta'$ is $m(\varepsilon, \delta')$. □

B. Proof of Lemmas in Appendix A

B.1. Proof of Lemma A.1

Lemma B.1. *(Theorem 4.3 in Zhou & Gu (2022))* Let $\{\mathcal{G}_k\}_{k=1}^{\infty}$ be a filtration, and $\{\mathrm{x}_k, \eta_k\}_{k\ge 1}$ be a stochastic process such that $\mathrm{x}_k \in \mathbb{R}^d$ is $\mathcal{G}_k$-measurable and $\eta_k \in \mathbb{R}$ is $\mathcal{G}_{k+1}$-measurable. Let $L, \sigma, \lambda, \varepsilon > 0, \boldsymbol{\mu}^* \in \mathbb{R}^d$. For $k \ge 1$, let $y_k = \langle \boldsymbol{\mu}^*, \mathbf{x}_k\rangle + \eta_k$ and suppose that $\eta_k, \mathbf{x}_k$ also satisfy

$$\mathbb{E}\left[\eta_k \mid \mathcal{G}_k\right] = 0, \mathbb{E}\left[\eta_k^2 \mid \mathcal{G}_k\right] \le \sigma^2, |\eta_k| \le R, \|\mathrm{x}_k\|_2 \le L \tag{B.1}$$

For $k \ge 1$, let $\mathbf{Z}_k = \lambda\mathbf{I} + \sum_{i=1}^{k} \mathbf{x}_i\mathbf{x}_i^\top$, $\mathbf{b}_k = \sum_{i=1}^{k} y_i\mathbf{x}_i$, $\boldsymbol{\mu}_k = \mathbf{Z}_k^{-1}\mathbf{b}_k$, and

$$\begin{aligned} \beta_k =& 12\sqrt{\sigma^2 d\log\left(1 + kL^2/(d\lambda)\right)\log\left(32(\log(R/\varepsilon)+1)k^2/\delta\right)} \\ &+ 24\log\left(32(\log(R/\varepsilon)+1)k^2/\delta\right)\max_{1\le i\le k}\left\{|\eta_i|\min\left\{1, \|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}}\right\}\right\} + 6\log\left(32(\log(R/\varepsilon)+1)k^2/\delta\right)\varepsilon. \end{aligned} \tag{B.2}$$

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

$$\forall k \ge 1, \left\|\sum_{i=1}^{k}\mathbf{x}_i\eta_i\right\|_{\mathbf{Z}_k^{-1}} \le \beta_k, \quad \|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|_{\mathbf{Z}_k} \le \beta_k + \sqrt{\lambda}\,\|\boldsymbol{\mu}^*\|_2$$

Lemma B.2. Let $\widetilde{V}_{k,h}$, $\widehat{V}_{k,h}$, $\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,m}$, $\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}$, $\widetilde{\boldsymbol{\theta}}_{k,m}$, $\widehat{\boldsymbol{\theta}}_{k,m}$, $\widetilde{\boldsymbol{\phi}}_{k,h,m}$, $\widehat{\boldsymbol{\phi}}_{k,h,m}$ be defined in Algorithm 1, for any $k \in [K]$, $h \in [H]$, $m \in \overline{[M]}$. We have

$$\begin{aligned} &\left|\mathbb{V}\widehat{V}_{k,h+1}^{2^m}\left(s_h^k, a_h^k\right) - \widehat{\mathbb{V}}\widehat{V}_{k,h+1}^{2^m}\left(s_h^k, a_h^k\right)\right| \\ &\le \min\left\{1, \left\|\widehat{\boldsymbol{\phi}}_{k,h,m+1}\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m+1}^{-1}}\left\|\widehat{\boldsymbol{\theta}}_{k,m+1} - \boldsymbol{\theta}^*\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m+1}}\right\} \end{aligned}$$

$$+ \min\left\{1, 2\left\|\widehat{\boldsymbol{\phi}}_{k,h,m}\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}^{-1}}\left\|\widehat{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^*\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}}\right\}, \tag{B.3}$$

and

$$\begin{aligned}
&\left|\mathbb{V}\widetilde{V}_{k,h+1}^{2^m}\left(s_h^k, a_h^k\right) - \widetilde{\mathbb{V}}\widetilde{V}_{k,h+1}^{2^m}\left(s_h^k, a_h^k\right)\right| \\
&\leq \min\left\{1, \left\|\widetilde{\boldsymbol{\phi}}_{k,h,m+1}\right\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,m+1}^{-1}}\left\|\widetilde{\boldsymbol{\theta}}_{k,m+1} - \boldsymbol{\theta}^*\right\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,m+1}}\right\} \\
&\quad + \min\left\{1, 2\left\|\widetilde{\boldsymbol{\phi}}_{k,h,m}\right\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,m}^{-1}}\left\|\widetilde{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^*\right\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,m}}\right\}.
\end{aligned} \tag{B.4}$$

Proof of Lemma B.2. The proof follows the proof of Lemma C.1 in (Zhou et al., 2021b). We first prove (B.3), and the proof of (B.4) is similar. We have

$$\begin{aligned}
&|[\widehat{\mathbb{V}}_{k,h}\widehat{V}_{k,h+1}^{2^m}](s_h^k, a_h^k) - [\mathbb{V}_{k,h}\widehat{V}_{k,h+1}](s_h^k, a_h^k)| \\
&= |[\langle\widehat{\boldsymbol{\phi}}_{k,h,m+1}, \widehat{\boldsymbol{\theta}}_{k,m+1}\rangle]_{[0,1]} - \langle\widehat{\boldsymbol{\phi}}_{k,h,m+1}, \boldsymbol{\theta}^*\rangle \\
&\qquad + (\langle\widehat{\boldsymbol{\phi}}_{k,h,m}, \boldsymbol{\theta}^*\rangle)^2 - [\langle\widehat{\boldsymbol{\phi}}_{k,h,m}, \widehat{\boldsymbol{\theta}}_{k,m}\rangle]_{[0,1]}^2| \\
&\leq \underbrace{|[\langle\widehat{\boldsymbol{\phi}}_{k,h,m+1}, \widehat{\boldsymbol{\theta}}_{k,m+1}\rangle]_{[0,1]} - \langle\widehat{\boldsymbol{\phi}}_{k,h,m+1}, \boldsymbol{\theta}^*\rangle|}_{I_1} \\
&\qquad + \underbrace{|(\langle\widehat{\boldsymbol{\phi}}_{k,h,m}, \boldsymbol{\theta}^*\rangle)^2 - [\langle\widehat{\boldsymbol{\phi}}_{k,h,m}, \widehat{\boldsymbol{\theta}}_{k,m}\rangle]_{[0,1]}^2|}_{I_2}
\end{aligned} \tag{B.5}$$

where the inequality holds due to triangle inequality. We have $I_1 \leq 1$ since both terms in I_1 lie in the interval $[0, 1]$. Furthermore,

$$\begin{aligned}
I_1 &\leq |[\langle\widehat{\boldsymbol{\phi}}_{k,h,m+1}, \widehat{\boldsymbol{\theta}}_{k,m+1}\rangle] - \langle\widehat{\boldsymbol{\phi}}_{k,h,m+1}, \boldsymbol{\theta}^*\rangle| \\
&= |[\langle\widehat{\boldsymbol{\phi}}_{k,h,m+1}, \widehat{\boldsymbol{\theta}}_{k,m+1} - \boldsymbol{\theta}^*\rangle| \\
&\leq \|\boldsymbol{\phi}_{k,h,m+1}\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m+1}^{-1}}\|\widehat{\boldsymbol{\theta}}_{k,m+1} - \boldsymbol{\theta}^*\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m+1}},
\end{aligned}$$

where the first inequality holds due to $\langle\widehat{\boldsymbol{\phi}}_{k,h,m+1}(s_h^k, a_h^k), \boldsymbol{\theta}^*\rangle \in [0, 1]$, the second inequality holds due to Cauchy-Schwarz inequality. Thus, we obtain

$$I_1 \leq \min\{1, \|\boldsymbol{\phi}_{k,h,m+1}\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m+1}^{-1}}\|\widehat{\boldsymbol{\theta}}_{k,m+1} - \boldsymbol{\theta}^*\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m+1}}\} \tag{B.6}$$

For I_2, we have

$$\begin{aligned}
I_2 &= \left|(\langle\widehat{\boldsymbol{\phi}}_{k,h,m}(s_h^k, a_h^k), \boldsymbol{\theta}^*\rangle) - [\langle\widehat{\boldsymbol{\phi}}_{k,h,m}, \widehat{\boldsymbol{\theta}}_{k,m}\rangle]_{[0,1]}\right| \cdot \left|(\langle\widehat{\boldsymbol{\phi}}_{k,h,m}(s_h^k, a_h^k), \boldsymbol{\theta}^*\rangle) + [\langle\widehat{\boldsymbol{\phi}}_{k,h,m}, \widehat{\boldsymbol{\theta}}_{k,m}\rangle]_{[0,1]}\right| \\
&\leq 2\left|(\langle\widehat{\boldsymbol{\phi}}_{k,h,m}(s_h^k, a_h^k), \boldsymbol{\theta}^*\rangle) - \langle\widehat{\boldsymbol{\phi}}_{k,h,m}, \widehat{\boldsymbol{\theta}}_{k,m}, \rangle\right| \\
&\leq 2\|\boldsymbol{\phi}_{k,h,m}(s_h^k, a_h^k)\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}^{-1}}\|\widehat{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^*\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}}
\end{aligned}$$

where the first inequality holds due to that both $\langle\widehat{\boldsymbol{\phi}}_{k,h,m}(s_h^k, a_h^k), \boldsymbol{\theta}^*\rangle$ and $[\langle\widehat{\boldsymbol{\phi}}_{k,h,m}, \widehat{\boldsymbol{\theta}}_{k,m}\rangle]_{[0,1]}$ lie in the interval $[0, 1]$, and the second inequality holds due to Cauchy-Schwarz inequality. Since I_2 belongs to the interval $[0, 1]$, we have

$$I_2 \leq \min\{1, 2\|\boldsymbol{\phi}_{k,h,m}(s_h^k, a_h^k)\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}^{-1}}\|\widehat{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^*\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}}\} \tag{B.7}$$

Substituting (B.6) and (B.7) into (B.5), we obtain (B.3). The proof of B.4 is nearly identical to the proof of (B.5). The only difference is to replace $\widehat{\boldsymbol{\phi}}$ with $\widetilde{\boldsymbol{\phi}}$, $\widehat{\boldsymbol{\theta}}$ with $\widetilde{\boldsymbol{\theta}}$, $\dot{\widehat{\boldsymbol{\Sigma}}}$ with $\dot{\widetilde{\boldsymbol{\Sigma}}}$. □

Proof of Lemma A.1. The proof follows Lemma C.2 in Zhou & Gu (2022). Symbols we used here may have small intuitively understandable modification compared to Algorithm 2 since we have to distinguish between Algorithm 2 applied to $\widetilde{V}_{k,h}$ and $\widehat{V}_{k,h}$. We first prove that Equation (A.1) holds with high probability. By definitions,

$$\widehat{\sigma}^2_{k,h,m} = \max\left\{\gamma^2 \left\|\widehat{\boldsymbol{\phi}}_{k,h,m}\right\|_{\widehat{\boldsymbol{\Sigma}}^{-1}_{k,h,m}}, \left[\widehat{\mathbb{V}}_{k,m}\widehat{V}^{2^m}_{k,h+1}\right](s^k_h, a^k_h) + \widehat{E}_{k,h,m}, \alpha^2\right\}$$
$$\widehat{\sigma}^2_{k,h,m} = \max\left\{\gamma^2 \left\|\widehat{\boldsymbol{\phi}}_{k,h,M-1}\right\|_{\widehat{\boldsymbol{\Sigma}}^{-1}_{k,h,M-1}}, 1, \alpha^2\right\}.$$

We define $\mathcal{C}_{k,m}$ as

$$\widehat{\mathcal{C}}_{k,m} := \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{k,m}\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}} \le \beta_k\}.$$

For each m, let

$$\begin{aligned}
\mathbf{x}_{k,h,m} &= \widehat{\sigma}^{-1}_{k,h,m}\widehat{\boldsymbol{\phi}}_{k,h,m},\\
\eta_{k,h,m} &= \widehat{\sigma}^{-1}_{k,h,m}\,\mathbb{1}\{\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{k,m} \cap \widehat{\mathcal{C}}_{k,m+1}\}[\widehat{V}^{2^m}_{k,h+1}(s^k_{h+1}) - \langle\widehat{\boldsymbol{\phi}}_{k,h,m}, \boldsymbol{\theta}^*\rangle],\\
\eta_{k,h,M-1} &= \widehat{\sigma}^{-1}_{k,h,M-1}[\widehat{V}^{2^{M-1}}_{k,h+1} - \langle\widehat{\boldsymbol{\phi}}_{k,h,M-1}, \boldsymbol{\theta}^*\rangle],\\
\mathcal{G}_{k,h} &= \mathcal{F}_{k,h},\\
\boldsymbol{\mu}^* &= \boldsymbol{\theta}^*.
\end{aligned}$$

We have

$$\mathbb{E}[\eta_{k,h,m}|\mathcal{G}_{k,h}] = 0, \quad \|\mathbf{x}_{k,h,m}\|_2 \le \widehat{\sigma}^{-1}_{k,h,m} \le 1/\alpha, \quad |\eta_{k,h,m}| \le 1/\alpha$$

Since $\mathbb{1}\{\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{k,m} \cap \widehat{\mathcal{C}}_{k,m+1}\}$ is $\mathcal{G}_{k,h}$-measurable, then we can bound the variance for $m \in \overline{[M]}$ as follows:

$$\begin{aligned}
\mathbb{E}[\eta^2_{k,h,m}|\mathcal{G}_{k,h}] &= \widehat{\sigma}^{-2}_{k,h,m}\,\mathbb{1}\{\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{k,m} \cap \widehat{\mathcal{C}}_{k,m+1}\}[\mathbb{V}\widehat{V}^{2^m}_{k,h+1}](s^k_h, a^k_h)\\
&\le \widehat{\sigma}^{-2}_{k,h,m}\,\mathbb{1}\{\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{k,m} \cap \widehat{\mathcal{C}}_{k,m+1}\}\Bigg[\widehat{\mathbb{V}}\widehat{V}^{2^m}_{k,h+1}\left(s^k_h, a^k_h\right)\\
&\quad + \min\left\{1, \left\|\widehat{\boldsymbol{\phi}}_{k,h,m+1}\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}^{-1}_{k,m+1}} \left\|\widehat{\boldsymbol{\theta}}_{k,m+1} - \boldsymbol{\theta}^*\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m+1}}\right\}\\
&\quad + \min\left\{1, 2\left\|\widehat{\boldsymbol{\phi}}_{k,h,m}\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}^{-1}_{k,m}} \left\|\widehat{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^*\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}}\right\}\Bigg]\\
&\le \widehat{\sigma}^{-2}_{k,h,m}\Bigg[\widehat{\mathbb{V}}\widehat{V}^{2^m}_{k,h+1}\left(s^k_h, a^k_h\right) + \min\left\{1, \beta_k\left\|\widehat{\boldsymbol{\phi}}_{k,h,m+1}\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}^{-1}_{k,m+1}}\right\}\\
&\quad + \min\left\{1, 2\beta_k\left\|\widehat{\boldsymbol{\phi}}_{k,h,m}\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}^{-1}_{k,m}}\right\}\Bigg]\\
&\le 1,
\end{aligned}$$

where the first inequality holds due to Lemma B.2, the second inequality holds due to the definition of the indicator function, and the third inequality holds due to the definition of $\widehat{\sigma}^{-2}_{k,h,m}$. For $m = M-1$, we have $\mathbb{E}[\eta^2_{k,h,m}|\mathcal{G}_{k,h}] \le 1$ directly by the definition of $\widehat{\sigma}^2_{k,h,m}$. For any $m \in \overline{[M]}$, we have

$$|\eta_{k,h,m}|\max\{1, \|\mathbf{x}_{k,h,m}\|_{\widehat{\boldsymbol{\Sigma}}^{-1}_{k,h-1,m}}\} \le \widehat{\sigma}^{-2}_{k,h,m}\|\widehat{\boldsymbol{\phi}}_{k,h,m}\|_{\widehat{\boldsymbol{\Sigma}}^{-1}_{k,h-1,m}} \le 1/\gamma^2,$$

where the first inequality follows from the definition of $\eta_{k,h,m}$ and $\mathbf{x}_{k,h,m}$, and the second inequality follows from the definition of $\widehat{\sigma}_{k,h,m}$. Let

$$y_{k,h,m} = \langle\boldsymbol{\mu}^*, \mathbf{x}_{x,h,m}\rangle + \eta_{k,h,m},$$

$$\begin{aligned}
\mathbf{Z}_{k,m} &= \lambda\mathbf{I} + \sum_{i=1}^{k}\sum_{h=1}^{H}\mathbf{x}_{i,h,m}\mathbf{x}_{i,h,m}^\top = \dot{\widehat{\mathbf{\Sigma}}}_{k+1,m},\\
\boldsymbol{b}_{k,m} &= \sum_{i=1}^{k}\sum_{h=1}^{H}\mathbf{x}_{i,h,m}y_{i,h,m},\\
\boldsymbol{\mu}_{k,m} &= \mathbf{Z}_{k,m}^{-1}\boldsymbol{b}_{k,m},\\
\varepsilon &= 1/\gamma^2.
\end{aligned}$$

Then, by Lemma B.1, for each $m \in \overline{[M]}$, with probability at least $1-\delta$, $\forall k \in [K+1]$,

$$\begin{aligned}
\|\boldsymbol{\mu}_{k-1,m} - \boldsymbol{\theta}^*\|_{\dot{\widehat{\mathbf{\Sigma}}}_{k,m}} \leq& 12\sqrt{d\log(1+kH/(\alpha^2 d\lambda))\log(32(\log(\gamma^2/\alpha)+1)k^2H^2/\delta)}\\
&+ 30\log(32(\log(\gamma^2/\alpha)+1)k^2H^2/\delta)/\gamma^2 + \sqrt{\lambda}B\\
&= \beta_k
\end{aligned} \tag{B.8}$$

Define the event that (B.8) happens for all k and m as $\widehat{\mathcal{E}}$. Conditioned on $\widehat{\mathcal{E}}$, the following properties hold:

- For $k=1$, $m \in \overline{[M]}$, by the definition of $\widehat{\boldsymbol{\theta}}_{1,m}$ and $\dot{\widehat{\mathbf{\Sigma}}}_{1,m}$, we have $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{1,m}\|_{\dot{\widehat{\mathbf{\Sigma}}}_{1,m}} = \|\boldsymbol{\theta}^*\|_{\lambda\mathbf{I}} \leq \sqrt{\lambda}B = \beta_1$, which implies

$$\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{1,m} \tag{B.9}$$

- For $k \in [K]$ and $m = M-1$, we directly have $\boldsymbol{\mu}_{k,M-1} = \widehat{\boldsymbol{\theta}}_{k+1,M-1}$, which implies

$$\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{k+1,M-1}. \tag{B.10}$$

- For $k \in [K]$ and $m \in \overline{[M-1]}$, we have

$$\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{k,m} \cap \widehat{\mathcal{C}}_{k,m+1} \Rightarrow y_{k,h,m} = \widehat{\sigma}^{-1}\widehat{V}_{k,h+1}^{2^m}(s_{h+1}^k) \Rightarrow \boldsymbol{\mu}_{k,m} = \widehat{\boldsymbol{\theta}}_{k+1,m} \Rightarrow \boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{k+1,m}. \tag{B.11}$$

  Therefore, by induction based on initial conditions (B.9) and (B.10), induction rule (B.11), we have for $k \in [K]$ and $m \in [M]$, $\boldsymbol{\theta}^* \in \widehat{\mathcal{C}}_{k,m}$. Taking the union bound gives that (A.1) happens with probability at least $1 - M\delta$. We can use the nearly identical argument to prove that (A.2) holds with probability at least $1 - M\delta$. The only difference is to replace $\widehat{\sigma}$ with $\widetilde{\sigma}$, $\widehat{\boldsymbol{\phi}}$ with $\widetilde{\boldsymbol{\phi}}$, $\widehat{\mathbb{V}}$ with $\widetilde{\mathbb{V}}$, $\widehat{V}$ with $\widetilde{\mathbb{V}}$, $\widehat{\mathbf{\Sigma}}$ with $\widetilde{\mathbf{\Sigma}}$, $\dot{\widehat{\mathbf{\Sigma}}}$ with $\dot{\widetilde{\mathbf{\Sigma}}}$, $\widehat{\boldsymbol{\theta}}$ with $\widetilde{\boldsymbol{\theta}}$. By taking the union bound, we obtain that with probability at least $1-2M\delta$, Equations (A.1) (A.2) both hold.

For (A.3) and (A.4), we have

$$\begin{aligned}
\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\dot{\widehat{\mathbf{\Sigma}}}_{k,0}} &\leq \left\|\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_{k,m}\right\|_{\dot{\widehat{\mathbf{\Sigma}}}_{k,0}} + \left\|\widehat{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^*\right\|_{\dot{\widehat{\mathbf{\Sigma}}}_{k,0}} \leq 2\beta_k,\\
\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\dot{\widetilde{\mathbf{\Sigma}}}_{k,0}} &\leq \left\|\boldsymbol{\theta}_k - \widetilde{\boldsymbol{\theta}}_{k,m}\right\|_{\dot{\widetilde{\mathbf{\Sigma}}}_{k,0}} + \left\|\widetilde{\boldsymbol{\theta}}_{k,m} - \boldsymbol{\theta}^*\right\|_{\dot{\widetilde{\mathbf{\Sigma}}}_{k,0}} \leq 2\beta_k
\end{aligned}$$

□

B.2. Proof of Lemma A.2

Proof of Lemma A.2. We prove this inequality by induction. Suppose

$$\begin{aligned}
V_{h+1}(s_{h+1};\boldsymbol{\theta}_K,\pi,r) - V_{h+1}(s_{h+1};\boldsymbol{\theta}^*,\pi,r)&\\
= \mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_{h+1}} W_{h+1}(\{(\mathbb{P}_K - \mathbb{P})V_{h+1}(s_h,a_h;\boldsymbol{\theta}_K,\pi,r)\}),&
\end{aligned} \tag{B.12}$$

which is true for $h = H$. Then, we have

$$V_h(s_h;\boldsymbol{\theta}_K,\pi,r) - V_h(s_h;\boldsymbol{\theta}^*,\pi,r)$$

$$
\begin{aligned}
&= \min\left\{1, r_h(s_h,a_h) + \mathbb{P}_K V_{h+1}(s_h,a_h;\boldsymbol{\theta}_K,\pi,r) - \big(r_h(s_h,a_h) + \mathbb{P}V_{h+1}(s_h,a_h;\boldsymbol{\theta}^*,\pi,r)\big)\right\}\\
&= \min\left\{1, \mathbb{P}_K V_{h+1}(s_h,a_h;\boldsymbol{\theta}_K,\pi,r) - \mathbb{P}V_{h+1}(s_h,a_h;\boldsymbol{\theta}^*,\pi,r)\right\}\\
&= \min\{1, (\mathbb{P}_K-\mathbb{P})V_{h+1}(s_h,a_h;\boldsymbol{\theta}_K,\pi,r) + \mathbb{P}(V_{h+1}(s_h,a_h;\boldsymbol{\theta}_K,\pi,r) - V_{h+1}(s_h,a_h;\boldsymbol{\theta}^*,\pi,r))\}\\
&= \min\{1, (\mathbb{P}_K-\mathbb{P})V_{h+1}(s_h,a_h;\boldsymbol{\theta}_K,\pi,r)\\
&\quad + \mathbb{E}_{s_{h+1}\sim\mathbb{P}(\cdot|s_h,a_h)}\mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_{h+1}} W_{h+1}(\{(\mathbb{P}_K-\mathbb{P})V_{h+1}(s_h;\boldsymbol{\theta}_K,\pi,r)\})\}\\
&= \mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_h} \min\{1, (\mathbb{P}_K-\mathbb{P})V_{h+1}(s_h,a_h;\boldsymbol{\theta}_K,\pi,r) + W_{h+1}(\{(\mathbb{P}_k-\mathbb{P})V_{h+1}(s_h,a_h;\boldsymbol{\theta}_K,\pi,r)\})\}\\
&= \mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_h} W_h(\{(\mathbb{P}_k-\mathbb{P})V_{h+1}(s_h,a_h;\boldsymbol{\theta}_K,\pi,r)\}).
\end{aligned}
$$

The first equality holds due to that $V_h(s_h;\boldsymbol{\theta}_K,\pi,r)$ and $V_h(s_h;\boldsymbol{\theta}^*,\pi,r)$ both belong to $[0,1]$, the third equality holds due to (B.12), and the forth equality holds due to that $\mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_h} = \mathbb{E}_{s_{h+1}\sim\mathbb{P}(\cdot|s_h,a_h)}\mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_{h+1}}$. Thus, by induction, we obtain the desired result (A.5). □

B.3. Proof of Lemma A.3

Proof of Lemma A.3. We first prove (B.13) by induction.

$$
\mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_1} W_1\left(\{u_{K,h}(s_h,\pi(s_h);\boldsymbol{\theta}_K,\pi,r)\}\right) \le \widehat{V}_{K,1}\left(s_1;\boldsymbol{\theta}_K,\pi,r\right). \tag{B.13}
$$

Suppose

$$
\mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_{h+1}} W_{h+1}(\{u_{K,h}(s_h,\pi(s_h);\boldsymbol{\theta}_K,\pi,r)\}) \le \widehat{V}_{K,h+1}(s_1;\boldsymbol{\theta}_K,\pi,r), \tag{B.14}
$$

which is true for $h=H$. Then,

$$
\begin{aligned}
&\widehat{V}_{K,h}(s_h;\boldsymbol{\theta}_K,\pi,r) - \mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_h} W_h(\{u_{K,h}(s_h,\pi(a_h);\boldsymbol{\theta}_K,\pi,r)\})\\
&\ge \min\Bigg\{0, u_{K,h}(s_h,\pi(s_h);\boldsymbol{\theta}_K,\pi,r) + 2\beta\left\|\boldsymbol{\phi}_{\widehat{V}_{K,h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h,\pi(s_h))\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{K,0}}\\
&\qquad +\boldsymbol{\phi}^\top_{\widehat{V}_{K,h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h,\pi(s_h))\,\boldsymbol{\theta}_K - \mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_h} W_h(\{u_{K,h}(s_h,\pi(s_h);\boldsymbol{\theta}_K,\pi,r))\}\Bigg\}\\
&\ge \min\Bigg\{0, u_{K,h}(s_h,\pi(s_h);\boldsymbol{\theta}_K,\pi,r) + 2\beta\left\|\boldsymbol{\phi}_{\widehat{V}_{K,h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h,\pi(s_h))\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{K,0}}\\
&\qquad + \boldsymbol{\phi}^\top_{\widehat{V}_{K,h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h,\pi(s_h))\boldsymbol{\theta}_K - u_{K,h}(s_h,\pi(s_h);\boldsymbol{\theta}_K,\pi,r)\\
&\qquad - \mathbb{E}_{s_{h+1}\sim\mathbb{P}(\cdot|s_h,\pi(s_h))}\mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_{h+1}} W_{h+1}(\{u_{K,h}(s_h,\pi(s_h);\boldsymbol{\theta}_K,\pi,r))\Bigg\}\\
&\ge \min\Bigg\{0, 2\beta\left\|\boldsymbol{\phi}_{\widehat{V}_{K,h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h,\pi(s_h))\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{K,0}} + \boldsymbol{\phi}^\top_{\widehat{V}_{K,h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h,\pi(s_h))\boldsymbol{\theta}_K\\
&\qquad -\mathbb{E}_{s_{h+1}\sim\mathbb{P}(\cdot|s_h,\pi(s_h))}\widehat{V}_{K,h+1}(s_{h+1};\boldsymbol{\theta}_K,\pi,r)\Bigg\}\\
&\ge \min\Bigg\{0, 2\beta\left\|\boldsymbol{\phi}_{\widehat{V}_{K,h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h,\pi(s_h))\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{K,0}} + \boldsymbol{\phi}^\top_{\widehat{V}_{K,h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h,\pi(s_h))(\boldsymbol{\theta}_K-\boldsymbol{\theta}^*)\Bigg\}\\
&\ge \min\Bigg\{0, 2\beta\left\|\boldsymbol{\phi}_{\widehat{V}_{K,h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h,\pi(s_h))\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{K,0}} - 2\beta\left\|\boldsymbol{\phi}_{\widehat{V}_{K,h+1}(\cdot;\boldsymbol{\theta}_K,\pi,r)}(s_h,\pi(s_h))\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{K,0}}\Bigg\}\\
&\ge 0,
\end{aligned}
$$

where the first inequality holds due to the definition of $\widehat{V}_{K,h}$, the second inequality holds due to the definition of $W_h(\cdot)$ and $\mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_h} = \mathbb{E}_{s_{h+1}\sim\mathbb{P}(\cdot|s_h,\pi(s_h))}\mathbb{E}_{\mathtt{traj}\sim(\pi,\mathbb{P})|\mathtt{traj}_{h+1}}$, the third inequality holds due to B.14, the fifth inequality holds due to Lemma A.1. Thus, by induction, B.13 holds. Thanks to the optimism of $\widehat{V}_{K,1}(s_1;\boldsymbol{\theta}_K,\pi_K,r_K)$, we have

$$
\widehat{V}_{K,1}(s_1;\boldsymbol{\theta}_K,\pi,r) \le \widehat{V}_{K,1}(s_1;\boldsymbol{\theta}_K,\pi_K,r_K),
$$

which concludes the proof. □

B.4. Proof of Lemma A.4

Proof of Lemma A.4. For the equation (A.8), we have

$$\begin{aligned}
&\widehat{V}_{k,h}(s_h^k) - u_{k,h}(s_h^k, a_h^k) - \mathbb{P}\widehat{V}_{k,h+1}(s_h^k, a_h^k) \\
&\leq \min\left\{1, 2\beta\left\|\widehat{\boldsymbol{\phi}}_{k,h,0}(s_h^k, a_h^k)\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}^{-1}} + \widehat{\boldsymbol{\phi}}_{k,h,0}^\top(s_h^k, a_h^k)\boldsymbol{\theta}_k - \widehat{\boldsymbol{\phi}}_{k,h,0}^\top(s_h^k, a_h^k)\boldsymbol{\theta}\right\} \\
&= \min\left\{1, 2\beta\left\|\widehat{\boldsymbol{\phi}}_{k,h,0}(s_h^k, a_h^k)\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}^{-1}} + \widehat{\boldsymbol{\phi}}_{k,h,0}^\top(s_h^k, a_h^k)(\boldsymbol{\theta}_k - \boldsymbol{\theta})\right\} \\
&\leq \min\left\{1, 2\beta\left\|\widehat{\boldsymbol{\phi}}_{k,h,0}(s_h^k, a_h^k)\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}^{-1}} + \left\|\widehat{\boldsymbol{\phi}}_{k,h,0}^\top(s_h^k, a_h^k)\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}^{-1}} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}}\right\} \\
&\leq \min\left\{1, 4\beta\left\|\widehat{\boldsymbol{\phi}}_{k,h,0}(s_h^k, a_h^k)\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\right\} \\
&\leq 4\min\left\{1, \beta\left\|\widehat{\boldsymbol{\phi}}_{k,h,0}(s_h^k, a_h^k)\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\right\}
\end{aligned}$$

where the first inequality holds due to that each term lies in the interval $[0, 1]$, the second inequality holds due to Cauchy-Schwartz inequality, and the third inequality holds due to lemma A.1. For the equation (A.9), we have

$$\begin{aligned}
&\widetilde{V}_{k,h}(s_h^k) - r_{k,h}(s_h^k, a_h^k) - \mathbb{P}\widetilde{V}_{k,h+1}(s_h^k, a_h^k) \\
&\leq \min\left\{1, \widetilde{\boldsymbol{\phi}}_{k,h,0}^\top(s_h^k, a_h^k)\boldsymbol{\theta}_k - \widetilde{\boldsymbol{\phi}}_{k,h,0}^\top(s_h^k, a_h^k)\boldsymbol{\theta}\right\} \\
&= \min\left\{1, \widetilde{\boldsymbol{\phi}}_{k,h,0}^\top(s_h^k, a_h^k)(\boldsymbol{\theta}_k - \boldsymbol{\theta})\right\} \\
&\leq \min\left\{1, \left\|\widetilde{\boldsymbol{\phi}}_{k,h,0}^\top(s_h^k, a_h^k)\right\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,0}^{-1}} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,0}}\right\} \\
&\leq \min\left\{1, 2\beta\left\|\widetilde{\boldsymbol{\phi}}_{k,h,0}^\top(s_h^k, a_h^k)\right\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\right\} \\
&\leq 2\min\left\{1, \beta\left\|\widetilde{\boldsymbol{\phi}}_{k,h,0}^\top(s_h^k, a_h^k)\right\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\right\},
\end{aligned}$$

where the first inequality holds due to that each term lies in the interval $[0, 1]$, the second inequality holds due to the Cauchy-Schwartz inequality, and the third inequality holds due to Lemma A.1. □

B.5. Proof of Lemma A.5

Lemma B.3 (Lemma B.1, Zhou & Gu (2022))**.** Let $\{\sigma_k, \beta_k\}_{k\geq 1}$ be a sequence of non-negative numbers, $\alpha, \gamma > 0, \{\mathbf{x}_k\}_{k\geq 1} \subset \mathbb{R}^d$ and $\|\mathbf{x}_k\|_2 \leq L$. Let $\{\mathbf{Z}_k\}_{k\geq 1}$ and $\{\bar{\sigma}_k\}_{k\geq 1}$ be recursively defined as follows: $\mathbf{Z}_1 = \lambda\mathbf{I}$

$$\forall k \geq 1, \bar{\sigma}_k = \max\left\{\sigma_k, \alpha, \gamma\|\mathbf{x}_k\|_{\mathbf{z}_k^{-1}}^{1/2}\right\}, \mathbf{Z}_{k+1} = \mathbf{Z}_k + \mathbf{x}_k\mathbf{x}_k^\top/\bar{\sigma}_k^2.$$

Let $\iota = \log\left(1 + KL^2/\left(d\lambda\alpha^2\right)\right)$. Then we have

$$\sum_{k=1}^K \min\left\{1, \beta_k\|\mathbf{x}_k\|_{\mathbf{Z}_k^{-1}}\right\} \leq 2d\iota + 2\max_{k\in[K]}\beta_k\gamma^2 d\iota + 2\sqrt{d\iota}\sqrt{\sum_{k=1}^K \beta_k^2(\sigma_k^2 + \alpha^2)}.$$

Proof of Lemma A.5. The proof is nearly identical to the proof of Lemma C.5 in Zhou & Gu (2022). The only difference is to replace $\widehat{\boldsymbol{\Sigma}}_{k,m}$ with $\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}$ (or $\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,m}$), $\widetilde{\boldsymbol{\Sigma}}_{k,h,m}$ with $\widehat{\boldsymbol{\Sigma}}_{k,h,m}$ (or still $\widetilde{\boldsymbol{\Sigma}}_{k,h,m}$), $\boldsymbol{\phi}_{k,h,m}$ with $\widehat{\boldsymbol{\phi}}_{k,h,m}$ (or $\widetilde{\boldsymbol{\phi}}_{k,h,m}$). □

B.6. Proof of Lemma A.6

Proof of Lemma A.6. The proof follows the proof of Lemma 25 in Zhang et al. (2021c) and Lemma C.6 in Zhou & Gu (2022). We have,

$$\begin{aligned}
\widehat{S}_m =& \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\left[\mathbb{P}\widehat{V}_{k,h+1}^{2^{m+1}}\right]\left(s_h^k,a_h^k\right) - \left(\left[\mathbb{P}\widehat{V}_{k,h+1}^{2^m}\right]\left(s_h^k,a_h^k\right)\right)^2\right] \\
=& \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\left[\mathbb{P}\widehat{V}_{k,h+1}^{2^{m+1}}\right]\left(s_h^k,a_h^k\right) - \widehat{V}_{k,h+1}^{2^{m+1}}\left(s_{h+1}^k\right)\right] + \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \Bigg[\widehat{V}_{k,h}^{2^{m+1}}\left(s_h^k\right) \\
& - \left(\left[\mathbb{P}\widehat{V}_{k,h+1}^{2^m}\right]\left(s_h^k,a_h^k\right)\right)^2\Bigg] + \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left(\widehat{V}_{k,h+1}^{2^{m+1}}\left(s_{h+1}^k\right) - \widehat{V}_{k,h}^{2^{m+1}}\left(s_h^k\right)\right) \\
\leq& \widehat{A}_{m+1} + \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\widehat{V}_{k,h}^{2^{m+1}}\left(s_h^k\right) - \left(\left[\mathbb{P}\widehat{V}_{k,h+1}^{2^m}\right]\left(s_h^k,a_h^k\right)\right)^2\right] + \sum_{k=1}^{K} I_{h_k}^k \widehat{V}_{k,h_k+1}^{2^{m+1}}\left(s_{h_k+1}^k\right),
\end{aligned} \tag{B.15}$$

where h_k is the largest index satisfying $I_h^k = 1$. For the second term, we have

$$\begin{aligned}
& \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\widehat{V}_{k,h}^{2^{m+1}}\left(s_h^k\right) - \left(\left[\mathbb{P}\widehat{V}_{k,h+1}^{2^m}\right]\left(s_h^k,a_h^k\right)\right)^2\right] \\
\leq& \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\widehat{V}_{k,h}^{2^{m+1}}\left(s_h^k\right) - \left(\left[\mathbb{P}\widehat{V}_{k,h+1}\right]\left(s_h^k,a_h^k\right)\right)^{2^{m+1}}\right] \\
=& \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left(\widehat{V}_{k,h}\left(s_h^k\right) - \left[\mathbb{P}\widehat{V}_{k,h+1}\right]\left(s_h^k,a_h^k\right)\right)\prod_{i=0}^{m}\left(\widehat{V}_{k,h}^{2^i}\left(s_h^k\right) + \left[\mathbb{P}\widehat{V}_{k,h+1}\right]\left(s_h^k,a_h^k\right)^{2^i}\right) \\
\leq& 2^{m+1}\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \max\left\{\widehat{V}_{k,h}\left(s_h^k\right) - \left[\mathbb{P}\widehat{V}_{k,h+1}\right]\left(s_h^k,a_h^k\right),0\right\} \\
\leq& 2^{m+1}\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[u_{k,h}\left(s_h^k,a_h^k\right) + 4\min\left\{1,\beta\left\|\widehat{\boldsymbol{\phi}}_{k,h,0}\right\|_{\dot{\widehat{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\right\}\right] \\
\leq& 2^{m+1}\left(\widetilde{R}_0 + 4\widehat{R}_0\right),
\end{aligned} \tag{B.16}$$

where the first inequality holds due to using $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$ recursively, the first equality holds due to the fact $x^{2^{m+1}} - y^{2^{m+1}} = (x-y)\prod_{i=0}^{m}(x^{2^m} - y^{2^m})$, the second inequality holds due to $\widehat{V}_{k,h}$ belongs to the interval $[0,1]$, the third inequality holds due to Lemma A.4, and the last inequality holds due to $u_{k,h}(s_h^k,a_h^k) = \beta\|\boldsymbol{\phi}_{V_{h+1}(\cdot;\boldsymbol{\theta}_k,\pi_k,r_k)}(s_h^k,a_h^k)\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,0}} = \beta\|\widetilde{\boldsymbol{\phi}}_{k,h,0}\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,0}}$. If $h_K \leq H$, we have $I_{h_k}^k \widehat{V}_{k,h_k+1}^{2^{m+1}}\left(s_{h_k+1}^k\right) \leq 1 = 1 - I_H^k$, and if $h_K = H$, $I_{h_k}^k \widehat{V}_{k,h_k+1}^{2^{m+1}}\left(s_{h_k+1}^k\right) = 0 = 1 - I_H^k$, which both give

$$\sum_{k=1}^{K} I_{h_k}^k \widehat{V}_{k,h_k+1}^{2^{m+1}}\left(s_{h_k+1}^k\right) \leq \sum_{k=1}^{K}(1 - I_H^k) = G \tag{B.17}$$

Substituting Equations (B.15), (B.16),(B.17) into (B.15), we can get (A.20). For Equation (A.15), similarly, we have

$$\widetilde{S}_m \leq \widetilde{A}_{m+1} + \sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\widetilde{V}_{k,h}^{2^{m+1}}\left(s_h^k\right) - \left(\left[\mathbb{P}\widetilde{V}_{k,h+1}^{2^m}\right]\left(s_h^k,a_h^k\right)\right)^2\right] + \sum_{k=1}^{K} I_{h_k}^k \widetilde{V}_{k,h_k+1}^{2^{m+1}}\left(s_{h_k+1}^k\right), \tag{B.18}$$

$$\sum_{k=1}^{K} I_{h_k}^k \widetilde{V}_{k,h_k+1}^{2^{m+1}}\left(s_{h_k+1}^k\right) \leq \sum_{k=1}^{K}\left(1 - I_H^k\right) = G. \tag{B.19}$$

And we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[\widehat{V}_{k,h}^{2^{m+1}}\left(s_h^k\right) - \left(\left[\mathbb{P}\widehat{V}_{k,h+1}^{2^m}\right]\left(s_h^k, a_h^k\right)\right)^2\right] \\
&\leq 2^{m+1}\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \max\left\{\widetilde{V}_{k,h}\left(s_h^k\right) - \left[\mathbb{P}\widetilde{V}_{k,h+1}\right]\left(s_h^k, a_h^k\right), 0\right\} \\
&\leq 2^{m+1}\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k \left[r_{k,h}(s_h^k, s_h^k) + \min\left\{1, 2\beta\left\|\widetilde{\boldsymbol{\phi}}_{k,h,0}\right\|_{\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,0}^{-1}}\right\}\right] \\
&\leq 2^{m+1}\left(K + 2\widetilde{R}_0\right)
\end{aligned} \tag{B.20}$$

where the first inequality holds similar to the derivation of (B.16), second inequality follows Lemma A.4, and the third inequality holds due to $\sum_{h=1}^{H} r_{k,h}(s_h^k, a_h^k) \leq 1$. Plugging Equations (B.19) (B.20) into B.18, we obtain A.21 □

B.7. Proof of Lemma A.7

Proof of Lemma A.7. The proof follows the proof of Lemma 25 in Zhang et al. (2021c) and Lemma C.7 in Zhou & Gu (2022). We use Lemma D.2 for $\widehat{A}_m$ and $\widetilde{A}_m$ for each m. To avoid confusion, we write ϵ, δ in Lemma D.2 as ϵ', δ'.

Let $x_{k,h} = I_h^k\left[\left[\mathbb{P}\widehat{V}_{k,h+1}^{2^m}\right]\left(s_h^k, a_h^k\right) - \widehat{V}_{k,h+1}^{2^m}(s_{h+1}^k)\right]$, $n = KH$, $\epsilon' = \sqrt{\log(1/\delta')}$, and $\delta' = \delta/(4\log(KH))$. Thus, $\mathbb{E}\left[\widehat{x}_{k,h}|\mathcal{F}_{k,h}\right] = 0$ and $\mathbb{E}\left[\widehat{x}_{k,h}^2|\mathcal{F}_{k,h}\right] = I_h^k\left[\mathbb{V}V_{k,h+1}^{2^m}\right]\left(s_h^k, a_h^k\right)$. Therefore, for each $m \in \overline{[M]}$, with probability at least $1-\delta$, we have

$$|\widehat{A}_m| = \left|\sum_{k=1}^{K}\sum_{h=1}^{H} x_{k,h}\right| \leq \sqrt{2\zeta\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k\left[\mathbb{V}\widehat{V}_{k,h+1}^{2^m}\right]\left(s_h^k, a_h^k\right)} + \zeta.$$

Similarly, let $x_{k,h} = I_h^k\left[\left[\mathbb{P}\widetilde{V}_{k,h+1}^{2^m}\right]\left(s_h^k, a_h^k\right) - \widetilde{V}_{k,h+1}^{2^m}(s_{h+1}^k)\right]$, $n = KH$, $\epsilon' = \sqrt{\log(1/\delta')}$, and $\delta' = \delta/(4\log(KH))$. With probability at least $1-\delta$, we have

$$|\widetilde{A}_m| = \left|\sum_{k=1}^{K}\sum_{h=1}^{H} x_{k,h}\right| \leq \sqrt{2\zeta\sum_{k=1}^{K}\sum_{h=1}^{H} I_h^k\left[\mathbb{V}\widetilde{V}_{k,h+1}^{2^m}\right]\left(s_h^k, a_h^k\right)} + \zeta.$$

Taking union bound over $m \in \overline{[M]}$ completes the proof. □

B.8. Proof of Lemma A.8

Proof of Lemma A.8. By the fact that $\det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k+1,m}^{-1/2}\right) < \det\left(\widehat{\boldsymbol{\Sigma}}_{k,H,m}^{-1/2}\right)$ and $\det\left(\dot{\widetilde{\boldsymbol{\Sigma}}}_{k+1,m}^{-1/2}\right)$ $< \det\left(\widetilde{\boldsymbol{\Sigma}}_{k,H,m}^{-1/2}\right)$, we have

$$\begin{aligned}
(1 - I_H^k) = 1 \Leftrightarrow\ & \exists m \in \overline{[M]}, \det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}^{-1/2}\right) / \det\left(\widehat{\boldsymbol{\Sigma}}_{k,H,m}^{-1/2}\right) > 4 \text{ or } \det\left(\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,m}^{-1/2}\right) / \det\left(\widetilde{\boldsymbol{\Sigma}}_{k,H,m}^{-1/2}\right) > 4 \\
\Rightarrow\ & \exists m \in \overline{[M]}, \det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}^{-1/2}\right) / \det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k+1,m}^{-1/2}\right) > 4 \text{ or } \det\left(\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,m}^{-1/2}\right) / \det\left(\dot{\widetilde{\boldsymbol{\Sigma}}}_{k+1,m}^{-1/2}\right) > 4
\end{aligned} \tag{B.21}$$

Let $\widehat{\mathcal{D}}_m$ and $\widetilde{\mathcal{D}}_m$ denote the indices k such that

$$\begin{aligned}
\widehat{\mathcal{D}}_m &:= \left\{k \in [K] :\ \det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k+1,m}\right) / \det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}\right) > 16\right\} \\
\widetilde{\mathcal{D}}_m &:= \left\{k \in [K] :\ \det\left(\dot{\widetilde{\boldsymbol{\Sigma}}}_{k+1,m}\right) / \det\left(\dot{\widetilde{\boldsymbol{\Sigma}}}_{k,m}\right) > 16\right\}
\end{aligned}$$

Then we have

$$G \leq \left| \bigcup_{m=0}^{M-1} \widehat{\mathcal{D}}_m \cup \bigcup_{m=0}^{M-1} \widetilde{\mathcal{D}}_m \right| \leq \sum_{m=0}^{M-1} \left| \widehat{\mathcal{D}}_m \right| + \sum_{m=0}^{M-1} |\widetilde{\mathcal{D}}_m|$$

For each m, we have

$$2\left|\widehat{\mathcal{D}}_m\right| < \sum_{k \in \widehat{\mathcal{D}}_m} \log 16 < \sum_{k \in \widehat{\mathcal{D}}_m} \log\left(\det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k+1,m}\right) / \det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}\right)\right) \leq \sum_{k=1}^{K} \log\left(\det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k+1,m}\right) / \det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}\right)\right)$$

Furthermore, since $\det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{K+1,m}\right) \leq \left(\operatorname{tr}\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{K+1,m}\right)/d\right)^d$ and $\operatorname{tr}\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{K+1,m}\right) \leq \operatorname{tr}(\lambda I) + \sum_{k,h} \left\|\widehat{\boldsymbol{\phi}}_{k,h.m}\right\|_2^2 / \widehat{\sigma}_{k,h,m}^2 \leq d\lambda + KH/\alpha^2$

$$\begin{aligned}\sum_{k=1}^{K} \log\left(\det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k+1,m}\right) / \det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{k,m}\right)\right) &= \log\left(\det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{K+1,m}\right) / \det\left(\dot{\widehat{\boldsymbol{\Sigma}}}_{1,m}\right)\right) \\ &\leq d\left(\log\left(\lambda + KH/(d\alpha^2)\right) - \log(\lambda)\right)\end{aligned}$$

Therefore $|\widehat{\mathcal{D}}_m|$ is upper bounded by

$$|\widehat{\mathcal{D}}_m| < d/2 \log(1 + KH/(d\lambda\alpha^2)).$$

And same for $|\widetilde{\mathcal{D}}_m|$. Taking summation gives the upper bound of G. □

C. Proof of Lower Bound

Reward-free exploration is more difficult than non-reward-free MDP by definitions since we can easily solve non-reward-free MDP by ignoring its reward and executing reward-free exploration. Thus, we will start with acquiring lower bounds under non-reward-free MDP settings and then obtain sample complexity lower bounds of reward-free exploration. The proof follows ideas of Zhou & Gu (2022) and Chen et al. (2021).

As noted in Section 5.2, we will consider the *hard-to-learn linear mixture MDPs* constructed in Zhou & Gu (2022). The state space $\mathcal{S} = \{x_1, x_2, x_3\}$ and the action space $\mathcal{A} = \{\mathbf{a}\} = \{-1, 1\}^{d-1}$. The reward function satisfies $r(x_1, \cdot) = r(x_2, \cdot) = 0$, and $r(x_3, \cdot) = \frac{1}{H}$. The transition probability satisfies $\mathbb{P}(x_2 \mid x_1, \boldsymbol{a}) = 1 - (\delta + \langle \boldsymbol{\mu}, \boldsymbol{a} \rangle)$ and $\mathbb{P}(x_3 \mid x_1, \boldsymbol{a}) = \delta + \langle \boldsymbol{\mu}, \boldsymbol{a} \rangle$, where $\delta = 1/6$ and $\boldsymbol{\mu} \in \{-\Delta, \Delta\}^{d-1}$ with $\Delta = \sqrt{\delta/K'}/(4\sqrt{2})$. The transition kernel is formulated as

$$\begin{aligned}\boldsymbol{\phi}\left(s' \mid s, \boldsymbol{a}\right) &= \begin{cases} \left(\alpha(1-\delta), -\beta \boldsymbol{a}^\top\right)^\top, & s = x_1, s' = x_2; \\ \left(\alpha\delta, \beta \boldsymbol{a}^\top\right)^\top, & s = x_1, s' = x_3; \\ \left(\alpha, \mathbf{0}^\top\right)^\top, & s \in \{x_2, x_3\}, s' = s; \\ \mathbf{0}, & \text{otherwise.} \end{cases} \\ \boldsymbol{\theta} &= \left(1/\alpha, \boldsymbol{\mu}^\top/\beta\right)^\top.\end{aligned}$$

The following lemma from Zhou & Gu (2022) lower bounds the regret for linear mixture MDP.

Lemma C.1. (Theorem 5.4 in Zhou & Gu (2022)) Let $B > 1$. Then for any algorithm, when $K' \geq \max\left\{3d^2, (d-1)/(192(B-1))\right\}$, there exists a B-bounded linear mixture MDP satisfying Assumptions 3.2 such that its expected regret $\mathbb{E}[\operatorname{Regret}(K')]$ is lower bounded by $\Omega\left(d\sqrt{K'}/(16\sqrt{3})\right)$.

Given Lemma C.1, we will use the regret lower bound of non-reward-free linear mixture MDPs to derive the sample coomplexity lower bound.

Lemma C.2. Suppose $B > 1$. Then for any algorithm $\texttt{ALG}_{NonFree}$ solving non-reward-free linear mixture MDP problems satisfying assumption 3.2, there exist a linear mixture $\mathcal{M}$ such that $\texttt{ALG}_{NonFree}$ needs to collect at least $\frac{Cd^2}{\varepsilon^2}$ episodes to output an ε-policy with probability at least $1-\delta$. Here C is an absolute constant.

Proof of Lemma C.2. For any algorithm $\texttt{ALG}_{NonFree}$, we construct an algorithm $\texttt{ALG}'_{NonFree}$ executing totally $K_1 = cK$ episodes, where c is a constant integer larger than 1. The first K episodes of $\texttt{ALG}'_{NonFree}$ are the same as $\texttt{ALG}_{NonFree}$, and the rest episodes keep executing the policy at the end of episode K. By Lemma C.1, we have

$$\sum_{k=1}^{K_1} \mathbb{E}\left[V(s_1;\boldsymbol{\theta}^*,\pi^*,r) - V(s_1;\boldsymbol{\theta}^*,\pi_k,r)\right] \geq \frac{c'd\sqrt{K_1}}{16\sqrt{3}}, \tag{C.1}$$

for some constant c'. In addition, based on the construction of *the hard-to-learn MDPs*, where $K^{'} = K_1$, the per-episode regret is upper bounded by

$$\mathbb{E}\left[V(s_1;\boldsymbol{\theta}^*,\pi^*,r) - V(s_1;\boldsymbol{\theta}^*,\pi_k,r)\right] \leq \frac{d}{4\sqrt{3K_1}}. \tag{C.2}$$

Thus, calculating (C.1) - $(K_1 - K)\times$ (C.2), and choosing$c = \max\left\{5/c', 2\right\}$, we have

$$\sum_{k=K+1}^{K_1} \mathbb{E}\left[V(s_1;\boldsymbol{\theta}^*,\pi^*,r) - V(s_1;\boldsymbol{\theta}^*,\pi_k,r)\right] \geq \frac{d\sqrt{K}}{16\sqrt{3c}}.$$

Since the policies in episode $K+1$ to episode K_1 are same to π_K, we have

$$\mathbb{E}\left[V(s_1;\boldsymbol{\theta}^*,\pi^*,r) - V(s_1;\boldsymbol{\theta}^*,\pi_K,r)\right] \geq \frac{d}{16\sqrt{3cK}c}.$$

Suppose the $\texttt{ALG}_{NonFree}$ return return a ε-optimal policy with probability $1-\delta$. Then,

$$(1-\delta)\varepsilon + \delta\frac{d}{4\sqrt{3cK}} \geq \frac{d}{16\sqrt{3cK}c}.$$

Setting $\delta < \min\{1, 1/(4c)\}$, by solving the inequality, we have $K \geq \frac{Cd^2}{\varepsilon^2}$ for some constant C. □

Since reward-free MDP is more difficult than non-reward-free MDP, Lemma C.2 directly indicates Theorem 5.6.

Proof of Theorem 5.6. We will prove the theorem by contradiction. Assume all reward-free linear mixture MDPs can be solved with sample complexity of $o(\frac{d^2}{\varepsilon^2})$. Then, for any non-reward-free MDP $\mathcal{M}$, there exists an algorithm $\texttt{ALG}'\ (\varepsilon,\delta)$ learning its reward-free counterpart $\mathcal{M}'$ with sample complexity of $o(\frac{d^2}{\varepsilon^2})$. We define $\texttt{ALG}$ solving $\mathcal{M}$ as follows: it collects K episodes of data and outputs the policy in the same way as $\texttt{ALG}'$ by ignoring the rewards. Then $\texttt{ALG}$ can also (ε,δ) learning $\mathcal{M}$ with sample complexity of $o(\frac{d^2}{\varepsilon^2})$, which contradicts Theorem C.2. □

Corollary 5.8 can be viewed as an direct result of Theorem 5.6.

Proof of Corollary 5.8. The hard-to-learn case we consider here is basically same as we consider in Theorem 5.6, except replacing reward function with $r(x_1,\cdot) = r(x_2,\cdot) = 0$, and $r(x_3,\cdot) = 1$. Since the reward here is H times the reward in Theorem 5.6, the suboptimality is also H times. Therefore, ε/H-optimal policy in Theorem 5.6 is a ε-optimal policy here. According to Theorem 5.6, the sample complexity required to achieve such a policy with probability at least $1-\delta$ is $\Omega(\frac{H^2d^2}{\varepsilon^2})$.

□

D. Auxiliary Lemmas

Lemma D.1. Suppose $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ are two positive definite matrices satisfying $\boldsymbol{A} \succeq \mathbf{B}$, then for any $\boldsymbol{x} \in \mathbb{R}^d$, $\|x\|_{\mathbf{A}} \leq \|\mathbf{x}\|_{\mathbf{B}} \sqrt{\det(\mathbf{A})/\det(\mathbf{B})}$

Lemma D.2 (Lemma 11, Zhang et al. 2021d)**.** Let $M > 0$ be a constant. Let $\{x_i\}_{i=1}^n$ be a stochastic process, $\mathcal{G}_i = \sigma(x_1, \ldots, x_i)$ be the σ-algebra of $x_1, \ldots, x_i$. Suppose $\mathbb{E}[x_i \mid \mathcal{G}_{i-1}] = 0, |x_i| \leq M$ and $\mathbb{E}\left[x_i^2 \mid \mathcal{G}_{i-1}\right] < \infty$ almost surely. Then, for any $\delta, \varepsilon > 0$, we have

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{i=1}^n x_i\right| \leq 2\sqrt{2\log(1/\delta)\sum_{i=1}^n \mathbb{E}[x_i^2 \mid \mathcal{G}_{i-1}]} + 2\sqrt{\log(1/\delta)}\varepsilon + 2M\log(1/\delta)\right)\\ &> 1 - 2\left(\log\left(M^2 n/\varepsilon^2\right) + 1\right)\delta \end{aligned} \tag{D.1}$$

Lemma D.3. (Lemma 12 in Zhang et al. (2021c)) Let $\lambda_1, \lambda_2, \lambda_4 > 0, \lambda_3 \geq 1$ and $\kappa = \max\{\log_2 \lambda_1, 1\}$. Let $a_1, \ldots, a_\kappa$ be non-negative real numbers such that $a_i \leq \min\left\{\lambda_1, \lambda_2\sqrt{a_i + a_{i+1} + 2^{i+1}\lambda_3} + \lambda_4\right\}$ for any $1 \leq i \leq \kappa$. Let $a_{\kappa+1} = \lambda_1$. Then we have $a_1 \leq 22\lambda_2^2 + 6\lambda_4 + 4\lambda_2\sqrt{2\lambda_3}$.