Dynamic Regret of Adversarial Linear Mixture MDPs

Long-Fei Li, Peng Zhao, Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology, Nanjing University, China
School of Artificial Intelligence, Nanjing University, China
{lilf, zhaop, zhouzh}@lamda.nju.edu.cn

Abstract
We study reinforcement learning in episodic inhomogeneous MDPs with adversarial full-information rewards and the unknown transition kernel. We consider the linear mixture MDPs whose transition kernel is a linear mixture model and choose the dynamic regret as the performance measure. Denote by $d$ the dimension of the feature mapping, $H$ the length of each episode, $K$ the number of episodes, $P_T$ the non-stationary measure, we propose a novel algorithm that enjoys an $O(\sqrt{d^2 H^3 K} + H^4 (1 + P_T)(1 + P_T))$ dynamic regret under the condition that $P_T$ is known, which improves previously best-known dynamic regret for adversarial linear mixture MDP and adversarial tabular MDPs. We also establish an $\Omega(\sqrt{d^2 H^3 K} + H^4 (1 + P_T)(1 + P_T))$ lower bound, indicating our algorithm is optimal in $K$ and $P_T$. Furthermore, when the non-stationary measure $P_T$ is unknown, we design an online ensemble algorithm with a meta-base structure, which is proved to achieve an $O(\sqrt{d^2 H^3 K} + H^4 (1 + P_T)(1 + P_T) + H^2 S_T^2)$ dynamic regret and here $S_T$ is the expected switching number of the best base-learner. The result can be optimal under certain regimes.

1 Introduction
Reinforcement Learning (RL) aims to learn a policy that maximizes the cumulative reward through interacting with the environment, which has achieved tremendous successes in various fields, including games [1, 2], robotic control [3, 4], and dialogue generation [5, 6]. In reinforcement learning, the Markov Decision Process (MDP) [7] is the most widely used model to describe the environment.

Traditional MDPs assume that the reward functions are stochastic and the number of actions and states is small. However, in many real-world applications, the reward functions may be adversarially changing and the state and action spaces are large or even infinite. Previous work studies these two problems separately. To deal with the adversarial reward functions, Even-Dar et al. [8] first consider learning adversarial MDPs with known transition and full-information reward feedback. They propose the MDP-E algorithm that enjoys $O(\sqrt{n^2 T})$ regret, where $n$ is the mixing time and $T$ is the number of steps. There is a line of subsequent work studying adversarial MDPs [9, 10, 11, 12, 13, 14, 15, 16], which studies various settings depending on whether the transition kernel is known, and whether the feedback is full-information or bandit. To overcome the large state and action space issue, a widely used method is linear function approximation, which reparameterizes the value function as a linear function over some feature mapping that maps the state and action to a low-dimensional space. Amongst these work, linear MDPs [17, 18, 19, 20, 21, 22, 23] and linear mixture MDPs [24, 25, 26, 27, 28, 29, 30, 31] are two of the most popular MDP models with linear function approximation. In particular, He et al. [23] and Zhou et al. [31] attain the minimax optimal $O(\sqrt{d^2 H^3 K})$ regret for linear MDPs and linear mixture MDPs with stochastic rewards respectively.

Recent studies try to combine two lines of work to establish the theoretical foundation of adversarial MDPs with large state and action space. In particular, Cai et al. [32] study adversarial linear mixture
We discuss the technical difficulty of removing with the dynamic regret guarantee by the meta-base two-layer structure. Our algorithm enjoys with linear function approximation. Both algorithms of Fei et al. [35] and Zhong et al. [36] require

where \( \pi \) \( V \)

This work investigates the dynamic regret of adversarial linear mixture MDPs, with a focus on

• We first show that the dynamic regret depends on the inverse of the minimal probability over

regime, the degree of environmental non-stationarity to some extent. Moreover, under specific regimes, the

switching number of the best base-learner. Though

an

e

Furthermore, we study the case when

\( P \)\( V \)

is the number of episodes, \( K \)

Dynamic regret is a more appropriate measure in non-stationary environments, but it is more chal-

\( P \)

measure \( d \)

as the algorithmic input ahead of time. For the unknown transition setting, Fei et al. [35]

study adversarial tabular MDPs and propose an algorithm with dynamic regret guarantee. Zhong et al. [36] further extend the algorithm of Fei et al. [35] to accommodate non-stationary transition kernels with linear function approximation. Both algorithms of Fei et al. [35] and Zhong et al. [36] require the quantity of \( P_T \) as the input. Moreover, their dynamic regret bounds are suboptimal in terms of \( K \) and \( P_T \) as demonstrated by the lower bound established by our work (see Theorem 4).

This work investigates the dynamic regret of adversarial linear mixture MDPs, with a focus on

the full-information feedback and the unknown transition. We first propose

\( \text{POWERS-FixShare} \)

algorithm when \( P_T \) is known, an algorithm combining optimistic policy optimization with a Bernstein bonus and fixed-share mechanism. We show it enjoys an \( \tilde{O}(d^4H^4K + \sqrt{H^4(K + P_T)(1 + P_T)}) \) dynamic regret, where \( d \) is the dimension of the feature mapping, \( H \) is the length of each episode, \( K \) is the number of episodes, \( P_T \) is the non-stationary measure. We also establish a dynamic regret lower bound of \( \Omega(d^4H^4K + \sqrt{H}K(H + P_T)) \). We stress four remarks regarding our results:

1. Our result can recover the \( \tilde{O}(d^4H^4K) \) minimax optimal static regret in He et al. [33].
2. Our result improves upon the previously best-known \( \tilde{O}(dH^{7/4}K^{3/4} + H^2K^{2/3}P_T^{1/3}) \) dynamic regret for the fixed transition of Zhong et al. [36, Theorem 4.6] in terms of \( H, K \) and \( P_T \).
3. Our result can imply an \( \tilde{O}(\sqrt{H^4S^2AK + H^2(K + P_T)(1 + P_T)}) \) dynamic regret for ad-

versarial tabular MDPs, strictly improving upon the previously best-known \( \tilde{O}(\sqrt{H^4S^2AK + H^2K^{2/3}P_T^{1/3}}) \) dynamic regret of Fei et al. [35, Theorem 1], in terms of both \( K \) and \( P_T \).
4. As the lower bound suggests, our result is the first optimal regarding the dependence on \( d \), \( K \) and \( P_T \) and can be optimal in terms of \( H \) under certain regimes (\( H \leq d \) and \( P_T \leq d^2/H \)).

Furthermore, we study the case when \( P_T \) is unknown and design a novel algorithm equipped

with the dynamic regret guarantee by the meta-base two-layer structure. Our algorithm enjoys an
\( \tilde{O}(d^4H^4K + \sqrt{H^4(K + P_T)(1 + P_T) + H^2S_T^2}) \) dynamic regret, where \( S_T \) is the expected

switching number of the best base-learner. Though \( S_T \) is a data-dependent quantity, it also reflects the degree of environmental non-stationarity to some extent. Moreover, under specific regimes, the magnitude of \( S_T \) may be relatively negligible, resulting in our results still being optimal. Indeed, given that \( S_T \) is a data-dependent quantity, its inclusion in the regret bound is not ideal. Deriving bounds that rely exclusively on problem-dependent quantities, like \( P_T \), remains an open challenge. We discuss the technical difficulty of removing \( S_T \) in Section 5 and take this issue for future work.

Finally, we also highlight the main technical challenges and our solutions as follows.

• We first show that the dynamic regret depends on the inverse of the minimal probability over

the action space of our policies, which can be arbitrarily small. To this end, we propose a novel
algorithm with the fixed-share mechanism [37]. While this mechanism is proved to enjoy favorable dynamic regret in online convex optimization [38], it suffers an additional term that can be regarded as the weighted sum of the difference between the occupancy measure of the compared policies in online MDPs. To overcome the difficulty, we exploit the multiplicative stability to bound this term, eliminating the need for a restart strategy to handle the environmental non-stationarity as in previous studies [35, 36] and allows us to attain the dynamic regret optimal in terms of $K$ and $T$.

• We show the dynamic regret of online MDPs can be written as the weighted average of “multi-armed bandits” problems over all states, where the weight for each state is the unknown and changing probability of being visited by $π_1^c, . . . , π_K^c$. For the unknown $P_T$ case, we first show the standard two-layer structure used in non-stationary online learning studies [39, 40, 41] fails to achieve a favorable dynamic regret guarantee, which characterizes the unique difficulty of online MDPs. Then, we present an initial attempt to address this issue by a specifically designed two-layer structure. We prove our algorithm enjoys nice dynamic regret guarantees under certain regimes.

2 Related Work

RL with adversarial rewards. There are many studies on learning adversarial MDPs where the reward functions are adversarially chosen, yielding fruitful results that can be categorized into three lines [8, 9, 10, 11, 12, 13, 14, 15, 16]. In particular, the first line of work considers the infinite-horizon MDPs with uniform mixing assumption. In the known transition and full-information setting, the seminal work of Even-Dar et al. [8] proposes MDP-E algorithm that achieves the $\tilde{O}(\sqrt{\tau^3T})$ regret, where $\tau$ is the mixing time and $T$ is the number of steps. Another concurrent work of Yu et al. [9] achieves $\tilde{O}(T^{2/3})$ in the same setting. In the known transition and bandit-feedback setting, Neu et al. [10] propose MDP-EXP3 algorithm that attains $\tilde{O}(T^{2/3})$ regret. The second line of work considers the episodic loop-free MDPs. Neu et al. [11] first study this problem under the known transition setting and propose algorithms that achieve $\tilde{O}(\sqrt{T})$ and $\tilde{O}(\sqrt{T/\alpha})$ for full-information and bandit-feedback respectively, where $\alpha$ is the lower bound of the probability of all states under any policy. Zimin and Neu [12] propose O-REPS algorithm that enjoys $\tilde{O}(\sqrt{T})$ regret in both full-information and bandit-feedback setting without any additional assumption. Rosenberg and Mansour [13] and Jin et al. [14] further consider the harder unknown transition and bandit-feedback setting. The last line of work studies the episodic Stochastic Shortest Path (SSP) setting [15, 16]. In this paper, we focus on episodic MDPs with the unknown transition and full-information setting.

RL with linear function approximation. To design RL algorithms in large state and action space scenarios, recent works focus on solving MDPs with linear function approximation. In general, these works can be divided into three lines based on the specific assumption of the underlying MDP. The first line of work considers the low Bellman-rank assumption [42, 43, 44, 45], which assumes the Bellman error matrix has a low-rank factorization. The second line of work is based on the linear MDP assumption [17, 18, 19, 20, 21, 22, 23], where both the transition kernel and reward functions can be parameterized as linear functions of given state-action feature mappings $φ : S × A → \mathbb{R}^d$. The last line of work studies the linear mixture MDP [24, 25, 26, 28, 29, 30, 31], where the transition kernel can be parameterized as a linear function of a feature mapping $φ : S × A × S → \mathbb{R}^d$ but without the linear reward functions assumption. Amongst these works, He et al. [23] and Zhou et al. [31] attain the minimax optimal $\tilde{O}(\sqrt{TH^3K})$ regret for both episodic linear MDPs and linear mixture MDPs respectively. However, all the above studies consider the stochastic reward setting. In this paper, we study the episodic linear mixture MDP setting but with adversarial reward functions.

Non-stationary RL. Another related line of research is online non-stationary MDPs. In contrast to adversarial MDPs where the reward functions are generated in an adversarial manner, online non-stationary MDPs consider the setting where the reward functions are generated stochastically.
We focus on episodic inhomogeneous MDPs with full-information reward functions and the unknown transition kernel. The learner interacts with the MDP through a linear function of a feature mapping. In this setup, the transition kernel can be parameterized as a linear function of a feature mapping. The formal definition of linear mixture MDPs is as follows.

The learner interacts with the MDP through $K$ episodes without knowledge of transition kernel $\{P_h\}_{h \in [H]}$. In each episode $k$, the environment chooses the reward function $r_{k,h}$ and the learner observes the reward function $P_h$. The learner observes current state $s_{k,1}$, where the reward function may be chosen in an adversarial manner and depend on the history of the past $(k-1)$ episodes. Simultaneously, the learner decides a policy $\pi_k = \{\pi_{k,h}\}_{h \in [H]}$ where each $\pi_{k,h} : S \to \Delta(A)$ is a function that maps a state $s$ to a distribution over action space $A$. In the $h$ stage in episode $k$, the learner observes current state $s_{k,h}$, chooses an action $a_{k,h} \sim \pi_{k,h}(\cdot | s_{k,h})$, and transits to the next state $s_{k,h+1} \sim P_h(\cdot | s_{k,h}, a_{k,h})$. Then the learner observes the reward $r_{k,h}(s_{k,h}, a_{k,h})$ and observes the reward function $r_{k,h}$ as we consider the full-information setting. At stage $H+1$, the learner observes the final state $s_{k,H+1}$ but does not take any action, and the episode $k$ terminates. Denote by $T = KH$ the total steps throughout $K$ episodes.

Linear Mixture MDPs. In this work, we focus on a special class of MDPs called linear mixture MDPs, a setting initiated by Ayoub et al. [24] and further studied in the subsequent work [32, 31, 33]. In this setup, the transition kernel can be parameterized as a linear function of a feature mapping $\phi : S \times A \times S \to \mathbb{R}^d$. The formal definition of linear mixture MDPs is as follows.

**Definition 1** (Linear Mixture MDPs). An MDP $M = \{S, A, H, \{r_{k,h}\}_{k \in [K], h \in [H], \{P_h\}_{h \in [H]}\}$ is called an inhomogeneous, episode $B$-bounded linear mixture MDP, if there exist a known feature mapping $\phi(s' \mid s, a) : S \times A \times S \to \mathbb{R}^d$ and an unknown vector $\theta_h^* \in \mathbb{R}^d$ with $\|\theta_h^*\|_2 \leq B$, $\forall h \in [H]$ such that (i) $P_h(s' \mid s, a) = \phi(s' \mid s, a)^\top \theta_h^*$ for all $(s, a, s') \in S \times A \times S$ and $h \in [H]$, (ii) $\|\phi_V(s, a)\|_2 \equiv \|\sum_{s' \in S} \phi(s' \mid s, a)V(s')\|_2 \leq 1$ for any $(s, a) \in S \times A$ and function $V : S \to [0, 1]$.

Dynamic Regret. Dynamic regret of RL with adversarial rewards is only recently studied in the literature [34, 35, 36]. Zhao et al. [34] investigate the dynamic regret of adversarial tabular MDPs with the known transition kernel and present an algorithm with optimal dynamic regret. Importantly, their algorithm does not require the non-stationarity measure $P_T$ as the algorithmic input ahead of time. For the unknown transition setting, Fei et al. [35] study adversarial tabular MDPs and propose an algorithm with dynamic regret guarantees. Zhong et al. [36] further extend the algorithm of Fei et al. [35] to accommodate non-stationary transition kernels with linear function approximation. Both algorithms [35, 36] require the quantity of $P_T$ as the input. Moreover, their dynamic regret bounds are suboptimal in $K$ and $P_T$ as shown by the lower bound established in our work. In this work, we first design an optimal algorithm in terms of $K$ and $P_T$ when $P_T$ is known. Further, we develop the first algorithm to handle the unknown $P_T$ issue in adversarial MDPs with unknown transition.

3 Problem Setup

We focus on episodic inhomogeneous MDPs with full-information reward functions and the unknown transition kernel. Denote by $M = \{S, A, H, \{r_{k,h}\}_{k \in [K], h \in [H], \{P_h\}_{h \in [H]}\}$ an episodic inhomogeneous MDP, where $S$ is the state space, $A$ is the action space, $K$ is the number of episodes, $H$ is the horizon, $r_{k,h} : S \times A \to [0, 1]$ is the reward function, $P_h(\cdot | \cdot) : S \times A \times S \to [0, 1]$ is the transition kernel. We assume the rewards are deterministic without loss of generality and extending our results to stochastic rewards is straightforward. Let $S = |S|$ and $A = |A|$.

The learner interacts with the MDP through $K$ episodes without knowledge of transition kernel $\{P_h\}_{h \in [H]}$. In each episode $k$, the environment chooses the reward function $r_{k,h}$ and decides the initial state $s_{k,1}$, where the reward function may be chosen in an adversarial manner and depend on the history of the past $(k-1)$ episodes. Simultaneously, the learner decides a policy $\pi_k = \{\pi_{k,h}\}_{h \in [H]}$ where each $\pi_{k,h} : S \to \Delta(A)$ is a function that maps a state $s$ to a distribution over action space $A$. In the $h$ stage in episode $k$, the learner observes current state $s_{k,h}$, chooses an action $a_{k,h} \sim \pi_{k,h}(\cdot | s_{k,h})$, and transits to the next state $s_{k,h+1} \sim P_h(\cdot | s_{k,h}, a_{k,h})$. Then the learner observes the reward $r_{k,h}(s_{k,h}, a_{k,h})$ and observes the reward function $r_{k,h}$ as we consider the full-information setting. At stage $H+1$, the learner observes the final state $s_{k,H+1}$ but does not take any action, and the episode $k$ terminates. Denote by $T = KH$ the total steps throughout $K$ episodes.

Linear Mixture MDPs. In this work, we focus on a special class of MDPs called linear mixture MDPs, a setting initiated by Ayoub et al. [24] and further studied in the subsequent work [32, 31, 33]. In this setup, the transition kernel can be parameterized as a linear function of a feature mapping $\phi : S \times A \times S \to \mathbb{R}^d$. The formal definition of linear mixture MDPs is as follows.

**Definition 1** (Linear Mixture MDPs). An MDP $M = \{S, A, H, \{r_{k,h}\}_{k \in [K], h \in [H], \{P_h\}_{h \in [H]}\}$ is called an inhomogeneous, episode $B$-bounded linear mixture MDP, if there exist a known feature mapping $\phi(s' \mid s, a) : S \times A \times S \to \mathbb{R}^d$ and an unknown vector $\theta_h^* \in \mathbb{R}^d$ with $\|\theta_h^*\|_2 \leq B$, $\forall h \in [H]$ such that (i) $P_h(s' \mid s, a) = \phi(s' \mid s, a)^\top \theta_h^*$ for all $(s, a, s') \in S \times A \times S$ and $h \in [H]$, (ii) $\|\phi_V(s, a)\|_2 \equiv \|\sum_{s' \in S} \phi(s' \mid s, a)V(s')\|_2 \leq 1$ for any $(s, a) \in S \times A$ and function $V : S \to [0, 1]$.
We present our proposed algorithm in Algorithm 1. Similar to previous studies, the algorithm consists of

**Algorithm 1**

\[
\begin{align*}
\text{Input:} & \quad \text{step size } \eta, \text{ exploration parameter } \gamma \text{ and regularization parameter } \lambda. \\
1: & \quad \text{Initialize } \{\pi_{0,h}(\cdot | s)\}^H_{h=1}, s \in \mathcal{S} \text{ as uniform distribution, and set } \{Q_{0,h}(\cdot, \cdot)\}^H_{h=1} \text{ as zero function.} \\
2: & \quad \text{for } h = 1, 2, \ldots, K \text{ do} \\
3: & \quad \text{Receive the initial state } s_{k,1}. \\
4: & \quad \text{for } h = 1, 2, \ldots, H \text{ do} \\
5: & \quad \text{for all } h \in [H], s \in \mathcal{S}, \text{ updates policy by} \\
6: & \quad \pi_{k,h}(\cdot | s) \propto \pi_{k-1,h}(\cdot | s) \exp (\eta Q_{k-1,h}(s, \cdot)), \pi_{k,h}(\cdot | s) = (1 - \gamma) \pi_{k-1,h}(\cdot | s) + \gamma \pi^u(\cdot | s). \\
7: & \quad \text{Take the action following } a_{k,h} \sim \pi_{k,h}(\cdot | s_{k,h}) \text{ and transit to the next state } s_{k,h+1}. \\
8: & \quad \text{Obtain reward } r_{k,h}(s_{k,h}, a_{k,h}) \text{ and observe the reward function } r_{k,h}(\cdot, \cdot). \\
9: & \quad \text{end for} \\
10: & \quad \text{Initialize } V_{k,H+1}(\cdot) \text{ as a zero function.} \\
11: & \quad \text{for } h = H, H - 1, \ldots, 1 \text{ do} \\
12: & \quad \text{Set } Q_{k,h}(\cdot, \cdot) \leftarrow \left[ r_{k,h}(\cdot, \cdot) + (\theta_{k,h} \phi_{V,k,h+1}(\cdot, \cdot)) + \frac{\beta_k}{H} \sum_{k,h}^2 \phi_{V,k,h+1}(\cdot, \cdot) \right]_{[0,H-h+1]}. \\
13: & \quad \text{Set the estimated variance } \hat{V}_{k,h+1}(s_{k,h}, a_{k,h}) \text{ as in (11), bonus } E_{k,h} \text{ as in (21}). \\
14: & \quad \hat{\sigma}_{k,h} \leftarrow \sqrt{\max \left\{ \frac{H^2}{d}, \frac{\hat{V}_{k,h+1}(s_{k,h}, a_{k,h}) + E_{k,h}}{H} \right\}}. \\
15: & \quad \hat{\Sigma}_{k+1,h} \leftarrow \hat{\Sigma}_{k,h} + \hat{\sigma}_{k,h}^2 \phi_{V,k,h+1}(s_{k,h}, a_{k,h}) \phi_{V,k,h+1}(s_{k,h}, a_{k,h})^\top. \\
16: & \quad \hat{b}_{k+1,h} \leftarrow \hat{b}_k + \hat{\sigma}_{k,h}^2 \phi_{V,k,h+1}(s_{k,h}, a_{k,h}) V_{k,h+1}(s_{k,h+1}). \\
17: & \quad \hat{e}_{k+1,h} \leftarrow \hat{e}_k + \phi_{V,k,h+1}(s_{k,h}, a_{k,h}) V_{k,h+1}^2(s_{k,h+1}). \\
18: & \quad \hat{\theta}_{k+1,h} \leftarrow \hat{\Sigma}_{k+1,h}^{-1} \hat{b}_{k+1,h}, \hat{\theta}_{k+1,h} \leftarrow \hat{\Sigma}_{k+1,h}^{-1} \hat{b}_{k+1,h}. \\
19: & \quad \text{end for} \\
20: & \quad \text{end for} \\
\end{align*}
\]

**Dynamic Regret.** For any policy \( \pi = \{\pi_h\}_{h \in [H]} \) and any \( (k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A} \), we define the action-value function \( Q^\pi_{k,h} \) and value function \( V^\pi_{k,h} \) as

\[
Q^\pi_{k,h}(s, a) = \mathbb{E}_s \left[ \sum_{h'=h}^{H} r_{k,h'}(s_{h'}, a_{h'}) \bigg| s_h = s, a_h = a \right], V^\pi_{k,h}(s) = \mathbb{E}_s \left[ \sum_{h'=h}^{H} r_{k,h'}(s_{h'}, a_{h'}) \bigg| s_h = s \right].
\]

The Bellman equation is given by \( Q^\pi_{k,h} = r_{k,h} + \mathbb{E}_h V^\pi_{k,h+1} \), and \( V^\pi_{k,h}(s) = \mathbb{E}_{s \sim \pi_h(\cdot | s)} [Q^\pi_{k,h}(s, a)] \) with \( V^\pi_{k,H+1} = 0 \). For simplicity, for any function \( V : \mathcal{S} \rightarrow \mathbb{R} \), we define the operator

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

As stated in Section 1, dynamic regret is a more appropriate measure compared with static regret for the adversarial environments, which is defined in (2) and we rewrite it below for clarity:

\[
\text{D-Regret}(K) = \sum_{k=1}^{K} V^\pi_{k,1}(s_{k,1}) - \sum_{k=1}^{K} V^\pi_{k,1}(s_{k,1}),
\]

where \( \pi^1, \ldots, \pi^K \) is any sequence of compared policies. We define \( \pi^0 = \pi^1 \) to simplify the notation. The non-stationarity measure is defined as \( P_T = \sum_{k=1}^{K} \sum_{h=1}^{H} \| \pi^\pi_{k,h} - \pi^\pi_{k-1,h} \|_{1,\infty} \).

### 4 Optimal Dynamic Regret with Known \( P_T \)

We present our proposed algorithm in Algorithm 1. Similar to previous studies, the algorithm consists of (i) policy improvement phase, and (ii) policy evaluation phase. We introduce the details below. In Sections 4.1, we first consider the case when the transition is known to highlight the challenges even under the ideal setting. Then, we extend the results to the unknown transition setting in Section 4.2.
4.1 Policy Improvement Phase

In the policy improvement phase, the algorithm updates $\pi_k$ based on $\pi_{k-1}$ using the proximal policy optimization (PPO) method [50]. Specifically, at episode $k$, we define the following linear function:

$$L_{k-1}(\pi) = V_{k-1}^\pi(s_{k,1}) + \mathbb{E}_{\pi_{k-1}} \left[ \sum_{h=1}^{H} \mathbb{E}_{\pi_{k-1}} \left[ Q_{k-1,h}^\pi(s_k) \right] | s_1 = s_{k,1} \right],$$

which is the first-order Taylor approximation of $V_{k-1}^\pi(s_{k,1})$ around $\pi_{k-1}$. Then, we update $\pi_k$ by

$$\pi_k = \arg \max_{\pi \in \Delta(A | S,H)} L_{k-1}(\pi) - \frac{1}{\eta} \mathbb{E}_{\pi_{k-1}} \left[ \sum_{h=1}^{H} D_{KL}(\pi_h(\cdot | s_h) | \pi_{k-1,h}(\cdot | s_h)) \right],$$

where $\eta > 0$ is the stepsize and the KL-divergence encourages $\pi_k$ to be close to $\pi_{k-1}$ so that $L_{k-1}(\pi)$ is a good approximation of $V_{k-1}^\pi(s_{k,1})$. The update rule in (5) takes the following closed form,

$$\pi_{k,h}(\cdot | s) \propto \pi_{k-1,h}(\cdot | s) \cdot \exp \left( \eta \cdot Q_{k-1,h}^\pi(s, \cdot) \right), \forall h \in [H], s \in S. \quad (6)$$

We show the update rule in (6) ensures the following guarantee and the proof is in Appendix C.1.

**Lemma 1.** The update rule in (6) ensures the following dynamic regret guarantee:

$$\text{D-Regret}(K) \leq \frac{\eta KH^3}{2} + \frac{1}{\eta} \sum_{k=1}^{K} \mathbb{E}_{\pi_k} \left[ \widetilde{D}_{\text{KL}}(\pi_k^\pi || \pi_{k+1}) - \widetilde{D}_{\text{KL}}(\pi_k^\pi || \pi_k) \right]. \quad (7)$$

Note that the expectation in the last term in (7) is taken over $\pi_k^\pi$ which may change over episode $k$. For static regret, i.e., $\pi_1 = \ldots = \pi_K = \pi^*$, we can control this term by a standard telescoping argument, which is not viable for dynamic regret analysis. Fei et al. [35] propose a restart strategy to handle this term. Specifically, they restart the algorithm every certain number of steps and decompose the above expectation into $\mathbb{E}_{\pi_k^\pi} = \mathbb{E}_{\pi_{k_0}} + \mathbb{E}_{\pi_{k_0} - \pi_{k_0}^\pi}$ where $k_0 < k$ is the episode in which restart takes place most recently before episode $k$. For the first expectation $\mathbb{E}_{\pi_{k_0}}$, they apply a customized telescoping argument to each period as the expectation is taken over the fixed policy. The second expectation $\mathbb{E}_{\pi_{k_0} - \pi_{k_0}^\pi}$ involves the difference $\pi_{k_0}^\pi - \pi_{k_0}$ and can be bounded by $P_T$. However, as we will show in Theorem 4, their regret bound is suboptimal in terms of $K$ and $P_T$.

We introduce our approach below. Let us first consider taking expectations over any fixed policy $\pi$. Denote by $\delta$ the minimal probability over any action at any state for policies $\pi_1, \ldots, \pi_K$, i.e., $\delta = \min_{a \in A, s \in S} \pi_k(a | s), \forall a \in A, s \in S$, the last term in (7) can be upper bounded by $\sum_{k=1}^{K} \mathbb{E}_{\pi_k} [\widetilde{D}_{\text{KL}}(\pi_k^\pi || \pi_k) - \widetilde{D}_{\text{KL}}(\pi_k^\pi || \pi_{k+1})] \leq H \log A + P_T \log \frac{1}{\delta}$, showing that we need to control the minimal value of $\delta$ to obtain a favorable dynamic regret bound. To this end, we slightly modify the update rule in (6) and add a uniform distribution $\pi^u(\cdot | s)$, $\forall s \in S$. That is, the policy $\pi^u$ chooses each action with equal probability at any state. Thus, the update rule in (6) is modified as:

$$\pi_{k,h}^u(\cdot | s) \propto \pi_{k-1,h}(\cdot | s) \cdot \exp(\eta \cdot Q_{k-1,h}^{\pi_k}(s, \cdot)), \pi_{k,h}(\cdot | s) = (1 - \gamma) \pi_{k,h}^u(\cdot | s) + \gamma \pi_{k,h}(\cdot | s) \quad (8)$$

for any $s \in S, h \in [H]$, where $\gamma \geq 0$ is the exploration parameter. This update is called the fixed-share mechanism in online learning literature [37]. While the fixed-share mechanism is standard to obtain dynamic regret in modern online learning [38], several important new challenges arise in online MDPs due to taking expectations over the policy sequence of changing policies $\pi_1, \ldots, \pi_K$. In particular, we prove that performing (8) ensures the following dynamic regret.

**Lemma 2.** Set $\pi_u$ as uniform distribution on $A$ for any state $s \in S$. The update rule in (8) ensures

$$\text{D-Regret}(K) \leq \frac{\eta KH^3}{2} + \frac{1}{\eta} \left( P_T \log A + KH \log \frac{1}{\gamma} \right)$$

$$+ \frac{1}{\eta} \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k} \left[ \sum_{a \in A} \left( \pi_{k-1,h}(a | s_h) \log \frac{1}{\pi_{k,h}(a | s_h)} - \pi_{k,h}(a | s_h) \log \frac{1}{\pi_{k+1,h}(a | s_h)} \right) \right]. \quad (9)$$

The proof can be found in Appendix C.2. In the dynamic regret analysis in online learning, the last term in (9) is usually canceled out through telescoping since we do not need to take expectations [38].
However, this is not the case in online MDPs. Since the expectation is taken over the policy sequence of *changing* policies \( \pi_1^*, \ldots, \pi_K^* \), this term cannot be canceled out, which requires a more refined analysis. To address this issue, we decompose one step of the expectation in (9) as follows.

\[
\left( E_{\pi_{k-1}^c} \left[ \sum_{a \in A} \pi_{k-1}^c \log \frac{1}{\pi_{k,h}^c} \right] - E_{\pi_{k-1}^c} \left[ \sum_{a \in A} \pi_{k}^c \log \frac{1}{\pi_{k+1,h}^c} \right] \right) + \left( E_{\pi_{k-1}^c} - E_{\pi_{k-1}^c} \right) \sum_{a \in A} \pi_{k-1}^c \log \frac{1}{\pi_{k,h}^c}.
\]

With this decomposition, the first term can be canceled out through telescoping, yet it remains to control the second term — the weighted difference between the state-action occupancy measures of policy \( \pi_{k}^c \) and \( \pi_{k-1}^c \) with weight \(-\pi_{k-1}^c(a \mid s_h) \log \pi_{k}^c(a \mid s_h)\) for state-action \((s_h,a)\). To control it, we need to (i) ensure the weight is upper bounded by some universal constant, and (ii) bound the unweighted difference between the state-action occupancy measures, which are new challenges that arose in online MDPs compared with standard online learning.

For the first challenge, note that the weight \(-\pi_{k-1}^c(a \mid s_h) \log \pi_{k}^c(a \mid s_h)\) can be large or even infinite since \( \pi_{k-1}^c \) is the policy before uniform exploration and \( \pi_{k}^c \) can be arbitrarily small. Fortunately, \( \pi_{k}^c \) is obtained by one-step descent from \( \pi_{k-1}^c \), which is the policy after uniform exploration and can be lower bounded. We provide the following *multiplicative stability lemma* for the one-step update, which shows \( \pi_{k}^c \) is not far from \( \pi_{k-1}^c \) and thus is also lower bounded.

**Lemma 3** (Multiplicative Stability). For any distributions \( p \in \Delta(A) \) with \( p(a) > 0 \), for all \( a \in A \), and any function \( Q : S \times A \rightarrow [0, H] \), it holds for \( p' \in \Delta(A) \) with \( p'(a) \propto p(a) \cdot \exp(\eta \cdot Q(s,a)) \) and \( \eta \leq 1/H \) that \( p'(a) \in [p(a)/4, 4p(a)] \), for all \( a \in A \).

For the second challenge, we show the unweighted difference between the state-action occupancy measures can be bounded by the path length of policies. In particular, we have the following lemma.

**Lemma 4.** For any policy sequence \( \pi_1^*, \ldots, \pi_K^* \), it holds that

\[
\sum_{k=1}^{K} \left( E_{\pi_{k}^c} - E_{\pi_{k-1}^c} \right) \left[ \sum_{h=1}^{H} \sum_{i=1}^{h} \| \pi_{k,i}^c - \pi_{k-1,i}^c \|_1 \right] \leq H P_T.
\]

**Remark 1.** We note that a similar argument is also used in Fei et al. [35, Appendix B.2.2]. However, they prove this lemma by imposing an additional smooth visitation measures assumption [35, Assumption 1], which is not required in our analysis.

The proofs for Lemmas 3 and 4 can be found in Appendices C.3 and C.4 respectively. Combining Lemmas 2, 3 and 4, we can prove the guarantee for update rule (8). The proof is in Appendix C.5.

**Theorem 1.** Set \( \pi_1 \) as uniform distribution on \( A \) for any state \( s \in S \). The update rule in (8) ensures

\[
\text{D-Regret}(K) \leq \frac{\eta K H^3}{2} + \frac{1}{\eta} \left( H \log A + (1 + H) \log \frac{4A}{\gamma} \right) + K H \log \frac{1}{1 - \gamma}.
\]

**Remark 2.** Considering the static regret where \( \pi_1^c = \ldots = \pi_K^c = \pi^* \), we can recover the \( O(\sqrt{H^3 K \log A}) \) static regret in Cai et al. [32] under the stationary scenario by setting \( \gamma = 0 \), that is, without uniform exploration. However, when \( \gamma = 0 \), the dynamic regret is not bounded as there lacks an upper bound for \(-\log \gamma\), showing the necessity of the fixed-share mechanism.

### 4.2 Policy Evaluation Phase

Sections 4.1 focus on the simplified scenario where the transition is known. In this subsection, we further consider the unknown transition setting such that it is necessary to evaluate the policy \( \pi_k \) based on the \((k-1)\) historical trajectories. To see how the model estimation error enters the dynamic regret, we decompose the dynamic regret in the following lemma.

**Lemma 5** (Fei et al. [35, Lemma 1]). Define the model prediction error as \( \epsilon_{k,h} = r_{k,h} + \mathbb{E}_{a \in A} V_{k+1,h}(s_h) - Q_{k,h} \), the dynamic regret \( \text{D-Regret}(K) = \sum_{k=1}^{K} V_{k,1}(s_k) - \sum_{k=1}^{K} V_{k,1}(s_k, 1) \) can be written as

\[
\sum_{k,h} E_{\pi_{k}^c} \left[ (Q_{k,h}(s_h), \pi_{k}^c(- \mid s_h) - \pi_{k,h}^c(- \mid s_h)) \right] + M_{K,H} + \sum_{k,h} \left( E_{\pi_{k}^c} [L_{k,h}(s_h, a_h)] - \epsilon_{k,h}(s_h, a_h) \right),
\]

where \( M_{K,H} = \sum_{k=1}^{K} \sum_{h=1}^{H} M_{k,h} \) is a martingale that satisfies \( M_{k,h} \leq 4H, \forall k \in [K], h \in [H] \).

\(^{\dagger}\)With a slight abuse of notations, we omit \((- \mid s_h)\) for simplicity.
When applied to linear mixture MDPs, the key idea is learning the unknown parameter $\theta^*_s$ of the linear mixture MDP and using the learned parameter $\hat{\theta}_{k,h}$ to build an optimistic estimator $Q_{k,h}(\cdot, \cdot)$ such that the model prediction error is non-positive, which is more or less standard. From the definition of linear mixture MDP, for the learned value function $V_{k,h}(\cdot)$, we have $[\mathbb{P}_h V_{k,h+1}(s, a) = (\sum_{s'} \phi(s' | s, a)) V_{k,h+1}(s', a), \theta^*_s]$. Inspired by recent advances in policy evaluation for linear mixture MDPs [31], we adopt the weighted ridge regression to estimate the parameter $\theta^*_s$, that is, we construct the estimator $\hat{\theta}_{k,h}$ by solving the following weighted ridge regression problem:

$$\hat{\theta}_{k,h} = \arg \min_{\theta \in \mathbb{R}^d} \sum_{j=1}^{k-1} \left[ \langle \phi_{V_{j,h+1}}(s_j, a_j, \cdot), \theta \rangle - V_{j,h+1}(s_j, a_j) \right]^2 / \sigma_{j,h}^2 + \lambda \|\theta\|_2^2.$$

Here, $\sigma_{j,h}$ is the upper confidence bound of the variance $[\mathbb{V}_h V_{j,h+1}(s_j, a_j, \cdot)]$, and we set it as $\sigma_{j,h} = \sqrt{\max[H^2/d, \mathbb{V}_h V_{j,h+1}(s_j, a_j, \cdot) + E_{k,h}]}$, where $\mathbb{V}_h V_{k,h+1}(s_k, a_k, \cdot)$ is a scalar-valued empirical estimate for the variance of the value function $V_{k,h+1}$ under the transition probability $P_h(\cdot | s_k, a_k)$, and $E_{k,h}$ is the bonus term to guarantee that the true variance $[\mathbb{V}_h V_{k,h+1}(s_k, a_k, \cdot)]$ is upper bounded by $[\mathbb{V}_h V_{k,h+1}(s_k, a_k, \cdot) + E_{k,h}]$ with high probability. Then, the confidence set $\tilde{C}_{k,h}$ is constructed as follows:

$$\tilde{C}_{k,h} = \{ \theta | \|\mathbb{S}_{k,h}^{1/2}(\theta - \hat{\theta}_{k,h})\|_2 \leq \beta_k \}.$$  \hspace{1cm} (10)

where $\mathbb{S}_{k,h}$ is a covariance matrix based on the observed data, and $\beta_k$ is a radius of the confidence set. Given $\tilde{C}_{k,h}$, we estimate the Q-function following the principle of “optimism in the face of uncertainty” [51] and set it as $Q_{k,h}(\cdot, \cdot) = [\mathbb{E}_{h} \tilde{C}_{k,h}(\cdot, \cdot) + \max_{\theta \in \tilde{C}_{k,h}} \phi_{V_{k,h+1}}(\cdot, \cdot)]_{0:H-1}$.

It remains to estimate the variance $[\mathbb{V}_h V_{k,h+1}(s_k, a_k, \cdot)]$. By the definition of linear mixture MDPs, we have $[\mathbb{V}_h V_{k,h+1}(s_k, a_k, \cdot)] = \langle \phi_{V_{k,h+1}}(s_k, a_k, \cdot), \theta^*_s \rangle - \langle \phi_{V_{k,h+1}}(s_k, a_k, \cdot), \hat{\theta}_{k,h} \rangle^2$. Therefore, we estimate $[\mathbb{V}_h V_{k,h+1}(s_k, a_k, \cdot)]$ by the expression below

$$\left[ \langle \phi_{V_{k,h+1}}(s_k, a_k, \cdot), \hat{\theta}_{k,h} \rangle \right]_{0:H-1} - \left[ \langle \phi_{V_{k,h+1}}(s_k, a_k, \cdot), \hat{\theta}_{k,h} \rangle \right]_{0:H-1}^2,$$

where $\hat{\theta}_{k,h} = \arg \min_{\theta \in \mathbb{R}^d} \sum_{j=1}^{k-1} \left[ \langle \phi_{V_{j,h+1}}(s_j, a_j, \cdot), \theta \rangle - V_{j,h+1}(s_j, a_j) \right]^2 + \lambda \|\theta\|_2^2$. The details are summarized in Lines 10-20 of Algorithm 1 and we provide the following guarantee.

**Theorem 2.** Set the parameters as in Lemma 8, with probability at least $1 - \delta$, we have

$$\sum_{k=1}^{K} \left( V_{k,1}(s_0^k) - V_{k,1}(s_0^k) \right) = M_{K,H} - \sum_{k=1}^{K} \sum_{h=1}^{H} \pi_{k,h}(s_k^h, a_k) \leq \tilde{O} \left( \sqrt{dH^4K + d^2H^3} \right).$$

The proof is given in Appendix C.6. Theorem 2 shows the model estimation error can be bounded. Combining Theorems 1, 2 and Lemma 5, we present the dynamic regret bound in the next section.

### 4.3 Regret Guarantee: Upper and Lower Bounds

In this section, we provide the regret bound for our algorithm and present a lower bound of the dynamic regret for any algorithm for adversarial linear mixture MDPs with the unknown transition.

**Theorem 3.** Set $\eta = \min\{\sqrt{(P_T + \log A)/K, 1}/H, \gamma = 1/(KH)\}$ and $\beta_k$ as in Lemma 8, then with probability at least $1 - \delta$, it holds

$$\text{D-Regret}(K) \leq \tilde{O} \left( \sqrt{dH^4K + d^2H^3} + \sqrt{H^4(K + P_T)(1 + P_T)} \right),$$

where $P_T = \sum_{k=1}^{K} \sum_{h=1}^{H} \left| \pi_{k,h}^e - \pi_{k-1,h}^e \right|_{1,\infty}$ is the path length of the compared policies.
When we finally establish the lower bound of this problem. The proof can be found in Appendix C.8.

Theorem 4. Suppose $B \geq 2, d \geq 4, H \geq 3, K \geq \lceil (d-1)^2H/2 \rceil$, for any algorithm and any constant $\Gamma \in [0, 2KH]$, there exists an episodic B-bounded adversarial linear mixture MDP and compared policies $\pi_1, \ldots, \pi_K$ such that $P_T \leq \Gamma$, and $D$-Regret($K$) $\geq \Omega(\sqrt{dH^3K} + \sqrt{HK(H+\Gamma)})$.

When $H \leq d$, the upper bound is $\tilde{O}(\sqrt{dH^3K} + \sqrt{H^2(K+P_T)(1+P_T)})$. Combining it with Theorem 4, we discuss the optimality of our result. We consider the following three regimes.

- Small $P_T$: when $0 \leq P_T \leq d^2/H$, the upper bound (12) can be simplified as $\tilde{O}(\sqrt{d^2H^3K})$, and the lower bound is $\tilde{O}(\sqrt{d^2H^3K})$, hence our result is optimal in terms of $d, H$, and $K$.
- Moderate $P_T$: when $d^2/H \leq P_T \leq K$, the upper bound (12) can be simplified as $\tilde{O}(\sqrt{d^2H^3K} + \sqrt{H^4K(1+P_T)})$, and it is minimax optimal in $d, K$ and $P_T$ but loses a factor of $H^{3/2}$.
- Large $P_T$: when $P_T \geq K$, any algorithm suffers at most $O(HK^2)$ dynamic regret, while the lower bound is $\Omega(K\sqrt{H})$. So our result is minimax optimal in $K$ but loses a factor of $\sqrt{H}$.

5 Towards Optimal Dynamic Regret with Unknown $P_T$

This section further considers the case when the non-stationarity measure $P_T$ is unknown. By Theorem 1, we need to tune the step size $\eta$ optimally to balance the number of episodes $K$ and $P_T$ to achieve a favorable dynamic regret. To address the difficulty of not knowing $P_T$ ahead of time, we develop an online ensemble method to handle this uncertainty, in which a two-layer meta-base structure is maintained. While the methodology can be standard in recent non-stationary online learning [52, 39, 40, 41], new challenges arise in online MDPs. We introduce the details below.

By the performance difference lemma in Cai et al. [32, Lemma 3.2] (as restate in Lemma 13), we can rewrite the dynamic regret as

$$\sum_{k=1}^{K} \left[ V^{\pi_k}_{k, h}(s^k) - V^\pi_k(s^k) \right] = \sum_{k=1}^{K} \mathbb{E}_{\pi_k} \left[ \sum_{h=1}^{H} \left\langle Q^\pi_{k,h}(s_h, \cdot), \pi_k^h(\cdot | s_h) - \pi_{k,h}(\cdot | s_h) \right\rangle \right],$$

where the expectation is taken over the randomness of the state trajectory sampled according to $\pi_k^h$. The dynamic regret of online MDPs can be written as the weighted average of some “multi-armed bandits” problems over all states, where the weight for each state is the unknown and changing probability of being visited by $\pi_1^h, \ldots, \pi_K^h$. As the optimal step size depends on the unknown non-stationarity measure $P_T$ as shown in Section 4, a natural idea is to use the two-layer structure to learn the optimal step size as in recent online convex optimization literature [52, 39, 40, 41].

The general idea is constructing a step size pool $H = \{\eta_1, \ldots, \eta_N\}$ to discretize the value range of the optimal step size; and then maintaining multiple base-learners $B_1, \ldots, B_N$, each of which works with a specific step size $\eta_i$. Finally, a meta-algorithm is used to track the best base-learner and yield the final policy. Then, the dynamic regret can be decomposed as follows (omit $\cdot | s_h$) for simplicity):

$$\sum_{k=1}^{K} \mathbb{E}_{\pi_k} \left[ \sum_{h=1}^{H} \left\langle Q^\pi_{k,h}, \pi_k^h - \pi_{k,h} \right\rangle \right] + \sum_{i=1}^{N} \mathbb{E}_{\pi_i} \left[ \sum_{h=1}^{H} \left\langle Q^\pi_{i,h}, \pi_i^h - \pi_{i,h} \right\rangle \right].$$

Since the above decomposition holds for any index $i \in [N]$, we can always choose the base-learner with optimal step size to analyze and the first term is easy to control. The challenge is to control the
second term, which is the regret of the meta-algorithm. Different from the standard “Prediction with Expert Advice” problem, it involves an additional expectation over the randomness of states sampled according to \(\pi^*_i\). This poses a grand challenge compared to conventional online convex optimization where the expectation is not required. Although we can bound this term by \(P_T\), again, optimal tuning of the meta-algorithm is hindered as \(P_T\) is unknown. Consequently, we opt to upper bound it by the worst-case dynamic regret [53], that is, benchmarking the performance with the best choice of each round, which in turn introduces the dependence on the switching number of the best base-learner.

We introduce our approach as follows. We maintain multiple base-learners, each of which works with a specific step size \(\eta_i\). All base-learners update their policies according to the same action-value function \(Q_{k-1,h}(s_h, \cdot)\) of the combined policy \(\pi_{k-1}\), that is, the base-learner \(\mathcal{B}_i\) updates policy by

\[
\pi^i_{k,h'}(\cdot | s) \propto \pi^i_{k-1,h'}(\cdot | s) \exp(\eta_i Q_{k-1,h}(s_h, \cdot)),
\]

where \(\pi^i_{k-1,h'}(\cdot | s) = (1 - \gamma)\pi^i_{k-1,h'}(\cdot | s) + \gamma\pi^i(\cdot | s)\), (13)

Then, the meta-algorithm chooses the base-learner by measuring the quality of each base-learner. In our approach, we choose the best base-learner at the last episode, that is,

\[
\pi^*_{k,h}(\cdot | s) = \pi^i^k_{k-1,h}(\cdot | s) \text{ with } i^*_{k-1,h}(s_h) = \arg\max_{i \in [N]}(Q_{k-1,h}(s_h, \cdot), \pi^i_{k-1,h}(\cdot | s)).
\]

(14)

The details are summarized in Algorithm 2 of Appendix A and the guarantee is as follows.

**Theorem 5.** Set \(\gamma = 1/(KH)\), step size pool \(\mathcal{H} = \{\eta_i = (2i/H)\sqrt{(\log A)/K} \mid i \in [N]\}\) with \(N = \lfloor 1/2 \log \left(\frac{K}{2}\right)\). Algorithm 2 ensures

\[
\text{D-Regret}(K) \leq \tilde{O}\left(\sqrt{dH^4K + d^2H^3K} + \sqrt{H^4(K + P_T)(1 + P_T) + H^3S_T^2}\right),
\]

where \(P_T = \sum_{k=1}^{K} \sum_{h=1}^{H} \|\pi^i_{k,h} - \pi^i_{k-1,h}\|_1\), \(S_T = \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi^i_k} \mathbb{1}[i^*_{k,h}(s_h) \neq i^*_{k-1,h}(s_h)]\) is the expected switching number of best base-learner.

Combining it with Theorem 3, we discuss the optimality of our result. We consider two regimes.

- **Small \(S_T\):** when \(S_T \leq \max\{d\sqrt{HK}, H\sqrt{(K + P_T)(1 + P_T)}\}\), the term \(S_T\) can be subsumed by other terms. In this case, the upper bound in Theorem 5 is entirely the \(\gamma\) same as that in Theorem 3. This implies we maintain the same guarantees without \(P_T\) as algorithmic input.

- **Large \(S_T\):** when \(S_T > \max\{d\sqrt{HK}, H\sqrt{(K + P_T)(1 + P_T)}\}\), our result loses a factor of \(H S_T\) compared with the result in Theorem 3 for the known \(P_T\) setting.

By the above discussion, our result can be optimal in terms of \(K\) and \(P_T\) under certain regimes when \(P_T\) is unknown. In comparison, the regret bounds achieved via the restart mechanism [35, 36] remain sub-optimal across all regimes even \(P_T\) is known. Note that we introduce the notation \(S_T\) in the regret analysis, which also reflects the degree of environmental non-stationarity to some extent. Consider the following two examples: (i) in the stationary environment, \(S_T\) could be relatively small as the best base-learner would seldom change, and (ii) in the piecewise-stationary environment, \(S_T\) would align with the frequency of environmental changes. Indeed, given that \(S_T\) is a data-dependent quantity, its inclusion in the regret analysis is not ideal. Deriving bounds that rely exclusively on problem-dependent quantities, like \(P_T\), remains a significant open challenge.

### 6 Conclusion and Future Work

In this work, we study the dynamic regret of adversarial linear mixture MDPs with the unknown transition. For the case when \(P_T\) is known, we propose a novel policy optimization algorithm that incorporates a fixed-share mechanism without the need for restarts. We show it enjoys a dynamic regret of \(\tilde{O}\left(\sqrt{d^2H^3K} + \sqrt{H^4(K + P_T)(1 + P_T)}\right)\), strictly improving the previously best-known result of Zhong et al. [36] for the same setting and Fei et al. [35] when specialized to tabular MDPs. We also establish an \(\Omega(\sqrt{d^2H^3K} + \sqrt{HK(H + P_T)})\) lower bound, indicating that our algorithm is optimal regarding \(d, K\) and \(P_T\) and can be optimal in terms of \(H\) under certain regimes. Moreover, we explore the more complex scenario where \(P_T\) is unknown. We show this setting presents unique challenges that distinguish online MDPs from conventional online convex optimization. We introduce a novel two-layer algorithm and show its dynamic regret guarantee is attractive under certain regimes.

There are several important future works to investigate. First, how to remove the dependence on the switching number \(S_T\) is an important open question. Moreover, we focus on the full-information feedback in this work, it remains an open problem to extend the results to the bandit feedback.
Acknowledgements

This research was supported by National Key R&D Program of China (2022ZD0114800) and NSFC (62206125, 61921006). Peng Zhao was supported in part by the Xiaomi Foundation.

References


11


Algorithm 2 POWERS-FixShare-OnE

Input: step size pool \( H \), exploration parameter \( \gamma \) and regularization parameter \( \lambda \).
1: Initialize \( \{\pi_{0,h}(\cdot | s)\}_h=1, \forall i \in [N], \forall s \in S \) as uniform distribution on \( A \). \( \{p_{0,h}(s)\}_h=1, \forall s \in S \) as uniform distribution on \( N \) and set \( \{Q_{0,h}(\cdot, \cdot)\}_h=1 \) as zero function.
2: for \( k = 1, 2, \ldots, K \) do
3:     Receive the initial state \( s_{k,1} \).
4:     for \( h = 1, 2, \ldots, H \) do
5:         For all \( h \in [H] \), \( s \in S \), each base-learner \( B_i, \forall i \in [N] \) updates policy by
6:             \[ \pi_{k+1,h}^i(\cdot | s) \propto \pi_{k,h}^i(\cdot | s) \exp \left( \eta_i Q_{k-1,h}(s, \cdot) \right), \] \[ \pi_{k+1,h}^i(\cdot | s) = (1 - \gamma) \pi_{k,h}^i(\cdot | s) + \gamma \pi^u(\cdot | s). \]
7:         Set the estimated variance \( \sigma_{k,h}^2 \) as in \( (11) \), bonus \( E_{k,h} \) as in \( (21) \).
8:     end for
9:     Initialize \( V_{k,H+1}(\cdot) \) as a zero function.
10:    Set \( Q_{k,h}(\cdot, \cdot) \leftarrow \left[ V_{k,h}(\cdot, \cdot) + \langle \tilde{\theta}_{k,h}, \phi_{V_{k,h+1}(\cdot, \cdot)} \rangle + \beta_k \| \Sigma_{k,h} \|^{1/2} \| \phi_{V_{k,h+1}(\cdot, \cdot)} \| \right]_{[0,H-h+1]} \).
11:    Set \( V_{k,h}(\cdot) \leftarrow E_{k,h} \cdot \pi_{k,h}(\cdot | s_{k,h}) \).
12:    Set the estimated variance \( \sigma_{k,h}^2 \) as in \( (11) \), bonus \( E_{k,h} \) as in \( (21) \).
13:    \[ \tilde{\theta}_{k,h} + \Sigma_{k,h}^{-1} \phi_{V_{k,h+1}}(s_{k,h}, a_{k,h}) \]
14:    \[ V_{k,h+1}(\cdot) \leftarrow V_{k,h}(\cdot) + \sigma_{k,h}^{-2} \phi_{V_{k,h+1}}(s_{k,h}, a_{k,h}) \]
15:    \[ \Sigma_{k+1,h} \leftarrow \Sigma_{k,h} + \sigma_{k,h}^{-2} V_{k,h}^2(s_{k,h}, a_{k,h}) \]
16:    \[ \tilde{\theta}_{k+1,h} + \Sigma_{k+1,h} \phi_{V_{k,h+1}}(s_{k,h}, a_{k,h}) \]
17:    \[ V_{k,h+1}(\cdot) \leftarrow V_{k,h}(\cdot) + \lambda \| \tilde{\theta}_{k,h} \|_2. \]
18:    \[ \tilde{\theta}_{k+1,h} + \Sigma_{k+1,h} \phi_{V_{k,h+1}}(s_{k,h}, a_{k,h}) \]
19:    \[ V_{k,h+1}(\cdot) \leftarrow V_{k,h}(\cdot) + \lambda \| \tilde{\theta}_{k,h} \|_2. \]
20: end for
21: end for

B Recovering Tabular Case

In this part, we show the result of our algorithm when specialized to the tabular case. Note in the linear case, we adopt the weighted ridge regression to estimate the parameter \( \theta^*_{k,h} \) that is, we construct the estimator \( \tilde{\theta}_{k,h} \) by solving the following weighted ridge regression problem:

\[
\tilde{\theta}_{k,h} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{j=1}^{k-1} \left[ \langle \phi_{V_{j,h+1}}(s_{j,h}, a_{j,h}), \theta \rangle - V_{j,h+1}(s_{j,h+1}) \right]^2 / \sigma_{j,h}^2 + \lambda \| \theta \|_2^2.
\]

In the tabular case, we simply set \( \sigma_{j,h} = 1 \), and compute \( \phi_{V_{k,h+1}} \) by taking the sample mean of \( \{V_{k,h+1}(s_{j,h+1})\}_{j \in [k-1]} \). That is, we set it as

\[
\phi_{V_{j,h+1}}(s_{j,h}, a_{j,h}) = \sum_{s' \in S} \frac{N_k(s_{j,h}, a_{j,h}, s')}{N_k(s_{j,h}, a_{j,h}) + \lambda} \cdot V_{j,h+1}(s'),
\]

\[
\tilde{\theta}_{k,h} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{j=1}^{k-1} \left[ \langle \phi_{V_{j,h+1}}(s_{j,h}, a_{j,h}), \theta \rangle - V_{j,h+1}(s_{j,h+1}) \right]^2 / \sigma_{j,h}^2 + \lambda \| \theta \|_2^2.
\]
of each $r(s, a)$, where $N_k$ counts the number of times each tuple $(s, a, s')$ has been visited up to episode $k$. Then, we construct the estimator $\hat{\theta}_{k,h}$ by solving the linear regression problem:

$$\hat{\theta}_{k,h} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{j=1}^{k-1} \left[ \left( \phi V_{j,h+1}(s_{j,h}, a_{j,h}) \right) - V_{j,h+1}(s_{j,h+1}) \right]^2 + \lambda \|\theta\|_2^2.$$ 

Then, the confidence set $\hat{C}_{k,h}$ is constructed as follows:

$$\hat{C}_{k,h} = \{ \theta \mid \|\hat{S}_{k,h}^{1/2}(\theta - \hat{\theta}_{k,h})\|_2 \leq \beta \}.$$ 

The remaining part of the algorithm is the same as the linear case. The following theorem shows the result of our algorithm in the tabular case.

**Theorem 6.** Set $\gamma = 1/(KH)$, step size pool $\mathcal{H} = \{\eta_i = (2^i/H)\sqrt{(\log A)/K} \mid i \in [N]\}$ with $N = \lfloor 1/\log(\frac{K}{\log A}) \rfloor$ and $\beta = H\sqrt{S\log(dKH/\delta)}$, with probability at least $1 - \delta$, it holds

$$\text{D-Regret}(K) \leq \tilde{O}\left(\sqrt{H^4S^2AK} + \sqrt{H^4(K + P_T)(1 + P_T)}\right),$$

where $P_T = \sum_{k=1}^{K} \sum_{h=1}^{H} \|\pi_{k,h}^c - \pi_{k-1,h}^c\|_{1,\infty}$ is the path length of the compared policy sequence.

**Proof.** By Lemma 5, we decompose the dynamic regret as follows.

$$\text{D-Regret}(K) \leq \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k^c} \left[ \langle Q_{k,h}(s_h, \cdot), \pi_{k,h}(\cdot \mid s_h) \rangle - \pi_{k,h}(\cdot \mid s_h) \right] + M_{K,H}$$

$$+ \sum_{k=1}^{K} \sum_{h=1}^{H} \left( \mathbb{E}_{\pi_k^c} [t_{k,h}(s_h, a_h)] - t_{k,h}(s_h, a_h) \right),$$

Note that the policy evaluation phase is the same as the one in Fei et al. [35, Algorithm 3], by the estimator bound in Fei et al. [35, Lemmas 7 and 9], we have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k^c} [t_{k,h}(s_h, a_h)] \leq 0, \quad \sum_{k=1}^{K} \sum_{h=1}^{H} -t_{k,h}(s_h, a_h) \leq \mathcal{O}\left(\sqrt{H^4S^2AT} \log^2(SAKH/\delta)\right).$$

(15)

The remaining proof is the same as the proof of Theorem 3 in Appendix C.7.

**C Proofs**

In this section, we provide the proof of the results in this paper.

**C.1 Proof of Lemma 1**

**Proof.** By the definition of dynamic regret, we have

$$\text{D-Regret}(K)$$

$$= \sum_{k=1}^{K} V_{k,1}^{\pi_k}(s_{k,1}) - \sum_{k=1}^{K} V_{k,1}^{\pi_k}(s_{k,1})$$

$$= \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k^c} \left[ \langle Q_{k,h}(s_h, \cdot), \pi_{k,h}(\cdot \mid s_h) \rangle - \pi_{k,h}(\cdot \mid s_h) \right] \mid s_1 = s_{k,1}$$

$$\leq \frac{\eta K H^3}{2} + \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k^c} \left[ D_{KL}(\pi_{k,h}(\cdot \mid s_h) \| \pi_{k+1,h}(\cdot \mid s_h)) - D_{KL}(\pi_{k,h}(\cdot \mid s_h) \| \pi_{k+1,h}(\cdot \mid s_h)) \right],$$

where the equality holds by the performance difference lemma in Lemma 13 and the inequality holds by the one-step descent guarantee in Lemma 14. This finishes the proof. 

---

16
C.2 Proof of Lemma 2

Proof. Note the update rule is

\[ \pi_{k,h}(\cdot | s) \propto \pi_{k-1,h}(\cdot | s) \exp(\eta \cdot Q_{k-1,h}^{\pi_{k-1}}(s, \cdot)), \pi_{k,h}(\cdot | s) = (1 - \gamma)\pi_{k,h}(\cdot | s) + \gamma \pi^u(\cdot | s) \]

By Lemma 14, for any \( s_h \in S, h \in [H], k \in [K], \) we have

\[ \langle Q_{k,h}^\pi(s_h, \cdot), \pi_{k,h}^c(\cdot | s) - \pi_{k,h}(\cdot | s_h) \rangle \leq \frac{\eta \gamma H^2}{2} + \frac{1}{\eta} \sum_{a \in A} \left( \pi_{k,h}^c(a | s_h) \log \frac{1}{\pi_{k,h}(a | s_h)} - \pi_{k,h}^c(a | s_h) \log \frac{1}{\pi_{k+1,h}(a | s_h)} \right) \]

where the equality holds by the definition of KL divergence. We decompose the last term as follows:

\[ \pi_{k,h}^c(a | s_h) \log \frac{1}{\pi_{k,h}(a | s_h)} - \pi_{k-1,h}^c(a | s_h) \log \frac{1}{\pi_{k-1,h}(a | s_h)} + \pi_{k-1,h}^c(a | s_h) \log \frac{1}{\pi_{k,h}(a | s_h)} \]

For the first term, we have

\[ \sum_{a \in A} \left( \pi_{k,h}^c(a | s_h) \log \frac{1}{\pi_{k,h}(a | s_h)} - \pi_{k-1,h}^c(a | s_h) \log \frac{1}{\pi_{k-1,h}(a | s_h)} \right) \]

By the update rule, we have \( 1 \leq 1/\pi_{k,h}(a | s_h) \leq A/\gamma \) and \( \pi_{k,h}^c(a | s_h)/\pi_{k,h}(a | s_h) \leq 1/(1 - \gamma) \). Therefore, we have

\[ \sum_{a \in A} \left( \pi_{k,h}^c(a | s_h) - \pi_{k-1,h}^c(a | s_h) \right) \log \frac{1}{\pi_{k,h}(a | s_h)} + \pi_{k-1,h}^c(a | s_h) \log \frac{1}{\pi_{k,h}(a | s_h)} \]

Then, the dynamic regret is bounded as follows:

\[ \text{D-Regret}(K) \]

\[ = \sum_{k=1}^{K} V_{k,1}^{\pi_k}(s_{k,1}) - \sum_{k=1}^{K} V_{k,1}^{\pi_k}(s_{k,1}) \]

\[ = \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k^c} \left[ \langle Q_{k,h}^\pi(s_h, \cdot), \pi_{k,h}^c(\cdot | s) - \pi_{k,h}(\cdot | s_h) \rangle | s_1 = s_{k,1} \right] \]

\[ \leq \frac{\eta \gamma H^3}{2} + \frac{1}{\eta} \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k^c} \left[ \pi_{k,h}^c(a | s_h) \log \frac{1}{\pi_{k,h}(a | s_h)} - \pi_{k,h}^c(a | s_h) \log \frac{1}{\pi_{k+1,h}(a | s_h)} \right] \]

\[ \leq \frac{\eta \gamma H^3}{2} + \frac{1}{\eta} \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k^c} \left[ \| \pi_{k,h}^c(\cdot | s_h) - \pi_{k-1,h}^c(\cdot | s_h) \|_1 \log \frac{A}{\gamma} + \log \frac{1}{1 - \gamma} \right] \]

\[ + \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k^c} \left[ \pi_{k-1,h}^c(a | s_h) \log \frac{1}{\pi_{k-1,h}(a | s_h)} - \pi_{k-1,h}^c(a | s_h) \log \frac{1}{\pi_{k+1,h}(a | s_h)} \right] \]
we prove that for all $a$ where the first inequality holds by (16), the second inequality is due to (17) and (18), and the last inequality holds by the definition of $P_T = \sum_{k=1}^{K} \sum_{h=1}^{H} \|\pi_k,h - \pi_{k-1,h}\|_{1,\infty}$. It ends the proof. ■

C.3 Proof of Lemma 3

Proof. Lemma 3 is a simplified version of Chen et al. [54, Lemma 17] and we prove it in our notations below for self-containedness. It is easy to verify that the update rule $p'(a) \propto p(a) \cdot \exp(\eta \cdot Q(s, a))$ is equivalent to the update

$$p' = \arg \max_{p' \in \Delta(A)} \eta(p', Q(s, \cdot)) + D_{KL}(p' || p).$$

Thus, by the KKT condition, we have for some $\lambda$ and $\mu(a) \geq 0$, such that

$$Q(s, a) - \frac{1}{\eta} \log \frac{p'(a)}{p(a)} + \lambda + \mu(a) = 0, \text{ and } \mu(a)p'(a) = 0, \forall a \in A.$$ 

The above equations give the closed-form solution $p'(a) = p(a) \exp(\eta(Q(s, a) + \lambda + \mu(a)))$. First, we prove that for all $a \in A$ we have $\mu(a) = 0$. Indeed, when $\mu(a) \neq 0$, by $\mu(a)p'(a) = 0$, we have $p'(a) = p(a) \exp(\eta(Q(s, a) + \lambda + \mu(a))) = 0$, which contradicts with $p(a) > 0$. Then, we now separately discuss two cases.

**Case 1:** $\min_{a \in A} Q(s, a) \neq \max_{a \in A} Q(s, a)$. In this case, we first show that $\min_{a \in A} -Q(s, a) \leq \lambda \leq \max_{a \in A} -Q(s, a)$. We prove it by contradiction: If $\lambda \geq \max_{a \in A} -Q(s, a)$, then we have

$$\sum_{a \in A} p'(a) = \sum_{a \in A} p(a) \exp(\eta(Q(s, a) + \lambda + \mu(a))) \geq \sum_{a \in A} p(a) \geq 1,$$

contradicting with $p' \in \Delta(A)$. A similar argument holds for the case $\lambda \leq \min_{a \in A}(-Q(s, i))$. Thus, we have $\lambda \in [\min_{a \in A}(-Q(s, a)), \max_{a \in A}(-Q(s, a))]$. Then, we have $|Q(s, a) + \lambda + \mu(a)| \leq \max_{a \in A} Q(s, a) - \min_{a \in A} Q(s, a) \leq H$. By the condition on $\eta H \leq 1$, we have $p'(a) \in [\exp(-1)p(a), \exp(1)p(a)] \in [1/(4p(a)), 4p(a)]$.

**Case 2:** $\min_{a \in A} Q(s, a) = \max_{a \in A} Q(s, a)$. In this case, it is clear that $\lambda = -Q(s, a)$ must hold for all $a \in A$ to make $p$ and $p'$ both discussions. Thus $p'(a) = p(a)$ for all $a \in A$.

Combining the above two cases finishes the proof. ■

C.4 Proof of Lemma 4

To prove Lemma 4, we first introduce the following two lemmas. Denote by $\mathbb{P}_{\pi}^{k_h}(s' \mid s) = \sum_{a \in A} \mathbb{P}_{h}(s' \mid s, a) \pi_h(a \mid s)$ the transition kernel of the MDP in step $h$ under policy $\pi_h$ and recall that $\|\pi - \pi'\|_{1,\infty} = \max_{s \in S} \|\pi(\cdot \mid s) - \pi'(\cdot \mid s)\|$. The first lemma shows the difference between the state distribution of two policies can be bounded by the path length of the policies.

**Lemma 6** (Zhao et al. [34, Lemma 7]). For any state distribution $\mathbb{P}$, policy pair $\pi$ and $\pi'$, and transition kernel $\mathbb{P}_{\pi}$, we have

$$\left\| d\mathbb{P}_{\pi}^{k_h}(\cdot, s) - d\mathbb{P}_{\pi'}^{k_h}(\cdot, s) \right\|_1 \leq \|\pi_h - \pi'_h\|_{1,\infty}, \forall h \in [H].$$

**Proof.** Consider the case when $d$ is a delta function on $s$. The difference in the next distributions is

$$\left\| \mathbb{P}_{\pi}^{k_h}(\cdot, s) - \mathbb{P}_{\pi'}^{k_h}(\cdot, s) \right\|_1 = \sum_{s' \in \mathcal{S}} \sum_{a \in A} |\mathbb{P}(s' \mid s, a) (\pi_h(a \mid s) - \pi_h(a \mid s))|$$

$$\leq \sum_{s' \in \mathcal{S}} \sum_{a \in A} \mathbb{P}(s' \mid s, a) \|\pi_h(a \mid s) - \pi'_h(a \mid s)\|_1$$

where the first inequality holds by (16), the second inequality is due to (17) and (18), and the last inequality holds by the definition of $P_T = \sum_{k=1}^{K} \sum_{h=1}^{H} \|\pi_k,h - \pi_{k-1,h}\|_{1,\infty}$. It ends the proof. ■
where the second inequality holds by Lemmas 6 and 7 and the last inequality holds by a recursive
The second lemma shows the difference between the state distribution of the policy starting from the
different initial distributions can be bounded by the difference between the initial distributions.

**Lemma 7** (Zhao et al. [34, Lemma 8]). For any two initial distributions $d$ and $d'$, transition kernel $P$
and policy $\pi$, we have
\[ \|dP^\pi_h - d'P^\pi_h\|_1 \leq \|d - d'\|_1, \forall h \in [H]. \]

**Proof.** Note the relationship that $d(s') = \sum_{s \in S} d(s)P^\pi_{s,s'}$, we have
\[
\|dP^\pi_h - d'P^\pi_h\|_1 = \sum_{s' \in S} \left| \sum_{s \in S} d(s)P^\pi_{s,s'}(s' | s) - d'(s)P^\pi_{s,s'}(s' | s) \right|
\]
\[
\leq \sum_{s' \in S} \left| \sum_{s \in S} d(s)P^\pi_{s,s'}(s' | s) - d'(s)P^\pi_{s,s'}(s' | s) \right|
\]
\[
= \sum_{s' \in S} \left| d(s) - d'(s) \right| P^\pi_{s,s'}(s' | s)
\]
\[
= \sum_{s} \left| d(s) - d'(s) \right| \sum_{s'} P^\pi_{s,s'}(s' | s)
\]
\[
= \sum_{s} \left| d(s) - d'(s) \right| = \|d - d'\|_1.
\]

This finishes the proof. ■

Now, we are ready to prove Lemma 4.

**Proof of Lemma 4.** For any policy $\pi$ and $\pi'$ and any initial state $s_1$, denote by $d_h$ the distribution of
the MDP in step $h$ under policy $\pi_h$ and $d'_h$ the distribution of the MDP in step $h$ under policy $\pi'_h$, that
is $d_h(s_h) = \mathbb{E}_\pi[I(s_h) | s_1]$ and $d'_h(s_h) = \mathbb{E}_{\pi'}[I(s_h) | s_1]$. We have
\[
\|d_h - d'_h\|_1 = \|d_{h-1}P^\pi_h - d'_{h-1}P^\pi_h\|_1
\]
\[
\leq \|d_{h-1}P^\pi_h - d'_{h-1}P^\pi_h\|_1 + \|d'_{h-1}P^\pi_h - d'_{h-1}P^\pi'_h\|_1
\]
\[
\leq \|d_{h-1} - d'_{h-1}\|_1 + \|\pi_h - \pi'_h\|_{1,\infty}
\]
\[
\leq \sum_{i \in [h]} \|\pi_i - \pi'_i\|_{1,\infty}.
\]

where the second inequality holds by Lemmas 6 and 7 and the last inequality holds by a recursive
calculation. Thus, we have
\[
\sum_{k=1}^K \left( \mathbb{E}_{\pi_k^c} - \mathbb{E}_{\pi_{k-1}^c} \right) \left[ \sum_{h=1}^H I(s_h) | s_1 = s_{k,1} \right] \leq \sum_{k=1}^K \sum_{h=1}^H \sum_{i=1}^h \|\pi_{k,i}^c - \pi_{k-1,i}^c\|_{1,\infty} = HP_T.
\]

This finishes the proof. ■

**C.5 Proof of Theorem 1**

**Proof.** By Lemma 2, we have
\[
\text{D-Regret}(K) \leq \frac{\eta KH^3}{2} + \frac{1}{\eta} \left( P_T \log \frac{4A}{\gamma} + KH \log \frac{1}{1 - \gamma} \right) + \frac{1}{\eta} \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi_k^c} \left[ \sum_{a \in A} \pi_{k-1,h}(a | s_h) \log \frac{1}{\pi_{k,h}(a | s_h)} - \pi_{k,h}(a | s_h) \log \frac{1}{\pi_{k+1,h}(a | s_h)} \right]
\]

(19)
Then, the last term can be bounded as follows:
\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_h} \left( \sum_{a \in A} \left( \pi_{k-1,h}(a \mid s_h) \log \frac{1}{\pi_{k,h}(a \mid s_h)} - \pi_{k,h}(a \mid s_h) \log \frac{1}{\pi_{k+1,h}(a \mid s_h)} \right) \right)
\]
\[
\leq \sum_{k=1}^{K} \sum_{h=1}^{H} \left( \mathbb{E}_{\pi_{k-1}} \left( \sum_{a \in A} \pi_{k-1,h}(a \mid s_h) \log \frac{1}{\pi_{k,h}(a \mid s_h)} \right) - \mathbb{E}_{\pi_{k}} \left( \sum_{a \in A} \pi_{k,h}(a \mid s_h) \log \frac{1}{\pi_{k+1,h}(a \mid s_h)} \right) \right)
\]
\[
+ \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_{k-1}^{\pi_{k-1} - \pi_{k-1}}} \left[ \sum_{a \in A} \pi_{k-1,h}(a \mid s_h) \log \frac{1}{\pi_{k,h}(a \mid s_h)} \right]
\]
\[
\leq H \log A + \log \frac{4A}{\gamma} H P_T,
\]
where the second inequality holds by the telescoping sum and the third inequality holds because \( \pi_{k,h} = \pi_{k,h} \) is the uniform distribution and \( \pi_{k,h}(a \mid s) \geq \frac{2}{4A} \) by the multiplicative stability lemma in Lemma 3 and the last inequality holds by Lemma 4.

Combining (19) and (20), we have
\[
\text{D-Regret}(K) \leq \frac{\eta K H^3}{2} + \frac{1}{\eta} \left( H \log A + (1 + H) \log \frac{4A}{\gamma} P_T + KH \log \frac{1}{1 - \gamma} \right),
\]
which finishes the proof.

C.6 Proof of Theorem 2

To prove Theorem 2, we first introduce the following key lemma omitted in the main paper due to the space limit, which shows the guarantee of the confidence set.

Lemma 8 (Zhou et al. [31, Lemma 5]). Let \( \hat{G}_{k,h} \) be defined in (10) and set \( \hat{\beta}_k \) as
\[
\hat{\beta}_k = 8 \sqrt{d \log(1 + k/\lambda)} \log(4k^2H/\delta) + 4 \sqrt{d} \log(4k^2H/\delta) + \sqrt{KB}.
\]
Then, with probability at least \( 1 - 3\delta \), we have that simultaneously for all \( h \in [H] \) and \( k \in [K] \),
\[
\theta^0_h \in \left\{ \theta \mid \left\| \Sigma_{k,h}^{1/2} (\theta - \hat{\theta}_k) \right\|_2 \leq \hat{\beta}_k \right\}, \quad \left\| \overline{V}_{k,h} V_{k,h+1} \right\|_{2, \hat{\beta}_k} \leq E_{k,h},
\]
where \( E_{k,h} \) is defined as follows:
\[
\text{min} \left\{ H^2, 2H \hat{\beta}_k \| \Sigma_{k,h}^{-1/2} \phi V_{k,h+1} (s_{k,h}, a_{k,h}) \|_2 \right\} + \text{min} \left\{ H^2, \hat{\beta}_k \| \Sigma_{k,h}^{-1/2} \phi V_{k,h+1} (s_{k,h}, a_{k,h}) \|_2 \right\}
\]
with
\[
\hat{\beta}_k = 8d \sqrt{\log(1 + k/\lambda)} \log(4k^2H/\delta) + 4 \sqrt{d} \log(4k^2H/\delta) + \sqrt{KB},
\]
\[
\hat{\beta}_k = 8 \sqrt{dH^4 \log(1 + kH^4/(d\lambda)) \log(4k^2H/\delta) + 4H^2 \log(4k^2H/\delta) + \sqrt{KB}}.
\]
We denote by \( \mathcal{E} \) the event when the result of Lemma 8 holds, we have \( \text{Pr}(\mathcal{E}) \geq 1 - 3\delta \).

Then we present the following lemma that shows the value function difference can be decomposed into two parts, one is a martingale sequence and the other is the estimation error.

Lemma 9. For all \( k \in [K], h \in [H], \) it holds that
\[
V_{k,h}(s_k, h) - V_{k,h}^{\pi_k}(s_k, h) = \sum_{h' = h}^{H} \left( M_{k,h',1} + M_{k,h',2} - \mu_{k,h'}(s_{k,h'}, a_{k,h'}) \right),
\]

20
where the third equality holds by the fact $E_M = E_{a \sim \pi_k(\cdot \mid s_k, h)}[Q_{k, h}(s_k, h, a) - Q_{k, h}^\pi(s_k, h, a)] - (Q_{k, h}(s_k, h, a_k, h) - Q_{k, h}^\pi(s_k, h, a, h))$

$M_{k, h, 1} = \mathbb{E}_{a \sim \pi_k(\cdot \mid s_k, h)}[Q_{k, h}(s_k, h, a) - Q_{k, h}^\pi(s_k, h, a)] - (Q_{k, h}(s_k, h, a_k, h) - Q_{k, h}^\pi(s_k, h, a, h))$

$M_{k, h, 2} = [\mathbb{P}_h(V_{k, h + 1} - V_{k, h}^\pi)](s_k, h, a_k, h) - (V_{k, h + 1}(s_k, h + 1) - V_{k, h + 1}^\pi(s_k, h + 1))$

**Proof.** By the definition $V_{k, h}(s_k, h) = \mathbb{E}_{a \sim \pi_k(\cdot \mid s_k, h)}[Q_{k, h}(s_k, h, a)]$, we have

$$V_{k, h}(s_k, h) - V_{k, h}^\pi(s_k, h,h) = \mathbb{E}_{a \sim \pi_k(\cdot \mid s_k, h)}[Q_{k, h}(s_k, h, a) - Q_{k, h}^\pi(s_k, h, a)]$$

$$= \mathbb{E}_{a \sim \pi_k(\cdot \mid s_k, h)}[Q_{k, h}(s_k, h, a) - Q_{k, h}^\pi(s_k, h, a)] - (Q_{k, h}(s_k, h, a_k, h) - Q_{k, h}^\pi(s_k, h, a, h))$$

$$+ \mathbb{E}_{a \sim \pi_k(\cdot \mid s_k, h)}[Q_{k, h}(s_k, h, a) - Q_{k, h}^\pi(s_k, h, a)] - (Q_{k, h}(s_k, h, a_k, h) - Q_{k, h}^\pi(s_k, h, a, h))$$

$$= \mathbb{E}_{a \sim \pi_k(\cdot \mid s_k, h)}[Q_{k, h}(s_k, h, a) - Q_{k, h}^\pi(s_k, h, a)] - (Q_{k, h}(s_k, h, a_k, h) - Q_{k, h}^\pi(s_k, h, a, h))$$

$$+ \mathbb{P}_h(V_{k, h + 1} - V_{k, h}^\pi)(s_k, h, a_k, h) - \delta_{k,h}(s_k, h, a_k, h)$$

$$= \mathbb{E}_{a \sim \pi_k(\cdot \mid s_k, h)}[Q_{k, h}(s_k, h, a) - Q_{k, h}^\pi(s_k, h, a)] - (Q_{k, h}(s_k, h, a_k, h) - Q_{k, h}^\pi(s_k, h, a, h))$$

$$+ \mathbb{P}_h(V_{k, h + 1} - V_{k, h}^\pi)(s_k, h, a_k, h) - \delta_{k,h}(s_k, h, a_k, h)$$

where the third equality holds by the fact $Q_{k, h} = r_{k,h} + \mathbb{P}_h V_{k, h + 1} - \delta_{k, h}$ and $Q_{k, h}^\pi = r_{k,h} + \mathbb{P}_h V_{k, h + 1}$. Summing up the above equation from $h$ to $H$ recursively finishes the proof.

Note $M_{k, h, 1}$ is the noise from the stochastic policy and $M_{k, h, 2}$ is the noise from the state transition. Let $M_{k, h} = M_{k, h, 1} + M_{k, h, 2}, \forall k \in [K], h \in [H]$, we define two following high probability events:

$\mathcal{E}_1 = \left\{ \forall h \in [H], \sum_{k=1}^{K} \sum_{h'=h}^{H} M_{k, h'} \leq 4 \sqrt{H^3 K \log \frac{H}{\delta}} \right\}$

$\mathcal{E}_2 = \left\{ \sum_{k=1}^{K} \sum_{h=1}^{H} M_{k, h, 2} \leq \sqrt{8H^3 K \log \frac{1}{\delta}} \right\}$

According to the Azuma-Hoeffding inequality, we have $\Pr(\mathcal{E}_1) \geq 1 - \delta$ and $\Pr(\mathcal{E}_2) \geq 1 - \delta$.

Then, we show the model prediction error $\delta_{k, h}$ can be upper and lower bounded both.

**Lemma 10.** Define prediction error $\delta_{k, h} = r_{k,h} + \mathbb{P}_h V_{k, h + 1} - Q_{k, h}$, on the event $\mathcal{E}$, it holds that

$$-2\bar{\beta}_k\|\tilde{S}_{k, h}^{-1/2} \phi_{V_{k, h + 1}}(\cdot, \cdot)\|_2 \leq \delta_{k, h}(\cdot, \cdot) \leq 0, \forall k \in [K], h \in [H]$$

**Proof.** First, we prove the left-hand side inequality. By the definition of $\delta_{k, h}$, we have

$$-\delta_{k, h}(s, a)$$

$$= \delta_{k, h}(s, a)$$

$$\leq r_{k,h}(s, a) + \langle \hat{\theta}_{k, h}, \phi_{V_{k, h + 1}}(s, a) \rangle + \hat{\beta}_k\|\tilde{S}_{k, h}^{-1/2} \phi_{V_{k, h + 1}}(s, a)\|_2 - (r_{k,h} + \mathbb{P}_h V_{k, h + 1})(s, a)$$

$$= \langle \hat{\theta}_{k, h} - \theta_{k, h}', \phi_{V_{k, h + 1}}(s, a) \rangle + \hat{\beta}_k\|\tilde{S}_{k, h}^{-1/2} \phi_{V_{k, h + 1}}(s, a)\|_2 - 2\bar{\beta}_k\|\tilde{S}_{k, h}^{-1/2} \phi_{V_{k, h + 1}}(s, a)\|_2$$

where the first inequality holds by the configuration of $Q_{k, h}$ in Algorithm 1, the second equality holds by the definition of linear mixture MDP such that $\mathbb{P}_h V_{k, h + 1}(s, a) = \langle \phi_{V_{k, h + 1}}(s, a), \theta_{k, h}' \rangle$ and the last inequality holds by the configuration of confidence set in Lemma 8.

Next, we prove the right-hand side inequality. By the definition of $\delta_{k, h}$, we have

$$\delta_{k, h}(s, a)$$

$$= (r_{k,h} + \mathbb{P}_h V_{k, h + 1})(s, a) - Q_{k, h}(s, a)$$

$$\leq (r_{k,h} + \mathbb{P}_h V_{k, h + 1})(s, a) - [r_{k,h}(s, a) + \langle \hat{\theta}_{k, h}, \phi_{V_{k, h + 1}}(s, a) \rangle + \hat{\beta}_k\|\tilde{S}_{k, h}^{-1/2} \phi_{V_{k, h + 1}}(s, a)\|_2]_{[0, H + 1]}$$
where the second inequality holds by the configuration of $Q_{k,h}$ in Algorithm 1, the second equality holds by the definition of linear mixture MDP such that $[\mathbb{P}_h V_{k,h+1}](s, a) = (\phi_{V_{k,h+1}}(s, a), \theta_h^*)$ and the last inequality holds by the configuration of confidence set in Lemma 8.

Next, we show the prediction error can be bounded by the cumulative estimate variance.

**Lemma 11.** Define prediction error $\iota_{k,h} = r_{k,h} + \mathbb{P}_h V_{k,h+1} - Q_{k,h}$, on the event $\mathcal{E}$, it holds that

$$- \sum_{k=1}^{K} \sum_{h=1}^{H} \iota_{k,h}(s_{k,h}, a_{k,h}) \leq 2\beta_K \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \hat{\sigma}_{k,h}^2 \sqrt{2Hd \log(1 + K/\lambda)}}.$$

**Proof.** By Lemma 10 and the definition of $\iota_{k,h} = r_{k,h} + \mathbb{P}_h V_{k,h+1} - Q_{k,h}$, we have

$$- \sum_{k=1}^{K} \sum_{h=1}^{H} \iota_{k,h}(s_{k,h}, a_{k,h}) \leq \sum_{k=1}^{K} \sum_{h=1}^{H} 2\beta_k \|\hat{\Sigma}_{k,h}^{-1/2} \phi_{V_{k,h+1}}(s_{k,h}, a_{k,h})\|_2 H$$

$$\leq \sum_{k=1}^{K} \sum_{h=1}^{H} 2\beta_k \sigma_{k,h} \min \left\{ \left\| \hat{\Sigma}_{k,h}^{-1/2} \phi_{V_{k,h+1}}(s_{k,h}, a_{k,h}) / \sigma_{k,h} \right\|_2, 1 \right\}$$

$$\leq 2\beta_K \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \sigma_{k,h}^2 \sqrt{2Hd \log(1 + K/\lambda)}}.$$

where the second inequality holds by $2\beta_k \sigma_{k,h} \geq \sqrt{\alpha H / \sqrt{d}} = H$, the third inequality is by Cauchy-Schwarz inequality and the last inequality holds by the elliptical potential lemma in Lemma 16. ■

Define the event

$$\mathcal{E}_3 = \left\{ \sum_{k=1}^{K} \sum_{h=1}^{H} [V_{h+1}(s_{k,h})]^2(s_{h+1}^k, a_{h+1})^k \leq 3(HK + H^3 \log(1/\delta)) \right\},$$

by the law of total variance in Lemma 15, we have $\Pr(\mathcal{E}_3) \geq 1 - \delta$. Then, we have the following lemma which bounds the estimated variance of the value function.

**Lemma 12** (He et al. [33, Lemma 6.5]). On the events $\mathcal{E} \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, it holds that

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \sigma_{k,h}^2 \leq \frac{2H^3 K^2}{d} + 179H^2 K^2 + 165d^3 H^4 \log^2 \left( 4K^2 H / \delta \right) \log^2(1 + K H^4 / \lambda) + 2062d^2 H^5 \log^2 \left( 4K^2 H / \delta \right) \log^2(1 + K / \lambda).$$

Now, we are ready to prove Theorem 2.

**Proof of Theorem 2.** On the events $\mathcal{E} \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, for any $h \in [H]$, it holds that

$$\sum_{k=1}^{K} V_{k,h}(s_{k,h}) - \sum_{k=1}^{K} V_{k,h}(s_{k,h})$$
where the first equality holds by Lemma 9, the second inequality holds by Lemma 11, and the last inequality holds by Lemma 12. This finishes the proof.

C.7 Proof of Theorem 3

Proof. By Lemma 5, we can rewrite the dynamic regret as follows.

\[
\text{D-Regret}(K) = \sum_{k=1}^{K} V_{k,1}^π(s_{k,1}) - \sum_{k=1}^{K} V_{k,1}^{π_κ}(s_{k,1})
\]

\[
= \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{π_κ}[(Q_{k,h}(s_h,\cdot), π_{k,h}(\cdot | s_h) - π_{k,h}(\cdot | s_h)]
\]

\[
+ \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{π_κ}[(t_{k,h}(s_h,a_h) - t_{k,h}(s_h,a_h)) + \mathcal{M}_{K,H}.
\]

By Lemma 10, we have $t_{k,h}(s,a) \leq 0$ for any $k \in [K], h \in [H], s \in S, a \in A$. Thus, we have

\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{π_κ}[(t_{k,h}(s_h,a_h)] \leq 0.
\]

By Theorem 2, we have

\[
\sum_{k=1}^{K} \sum_{h=1}^{H} (-t_{k,h}(s_h,a_h)) + \mathcal{M}_{K,H} \leq \tilde{O}(\sqrt{dH^4K + d^2H^3K}).
\]

It remains to bound the first term. Note our algorithm is indeed updated based on the estimated action-value function $Q_{k,h}$. By Theorem 1, set $γ = 1/KH$ and note that $\log(1/(1-γ)) \leq γ/(1-γ)$ for all $γ > 0$. Then, we have

\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{π_κ}[(Q_{k,h}(s_h,\cdot), π_{κ,h}(\cdot | s_h) - π_{κ,h}(\cdot | s_h)]
\]

\[
\leq \frac{ηKH^3}{2} + \frac{1}{η}(H \log A + (1 + H) \log \frac{A}{γ} P_T + 2)
\]

It is clear that the optimal step size is $η^* = \sqrt{P_T \log A / (KH^2)$. Note our step size is set as $η = \min\{\sqrt{(P_T \log A) / K}, 1\}/H$ to ensure $η \leq 1/H$. We consider the following two cases:

**Case 1: $η^* \leq 1/H$.** In this case, our step size is set as $η = \sqrt{(P_T \log A) / (KH^2)$. We have

\[
\frac{ηKH^3}{2} + \frac{1}{η}(H \log A + (1 + H) \log \frac{A}{γ} P_T + 2) \leq \tilde{O}(KH^4(1 + P_T)).
\]

**Case 2: $η^* > 1/H$.** In this case, our step size is set as $η = 1/H$. Therefore, we have

\[
\frac{ηKH^3}{2} + \frac{1}{η}(H \log A + (1 + H) \log \frac{A}{γ} P_T + 2) \leq \tilde{O}(H^2P_T),
\]
where the equality holds by \( P_T \geq K \) in this case. Combining these two cases, we have
\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{c_k^h}[\epsilon(Q_k,h(s_h,\cdot), \pi_{k,h}(\cdot | s_h) - \pi_{k,h}(\cdot | s_h))] \leq \tilde{O}(\sqrt{H^4(K + P_T)(1 + P_T)}). \tag{25}
\]
Combining (22), (23), (24) and (25), we obtain
\[
\text{D-Regret}(K) \leq \tilde{O} \left( \sqrt{dH^4K + d^2H^3K + \sqrt{H^4(K + P_T)(1 + P_T)}} \right).
\]
This finishes the proof.

C.8 Proof of Theorem 4

Proof. At a high level, we prove this lower bound by noting that optimizing the dynamic regret of linear mixture MDPs is harder than (i) optimizing the static regret of linear mixture MDPs with the unknown transition kernel, (ii) optimizing the dynamic regret of linear mixture MDPs with the known transition kernel, both. Thus, we can consider the lower bound of these two problems separately and combine them to obtain the lower bound of the dynamic regret of linear mixture MDPs with the unknown transition kernel.

First, we consider the lower bound of optimizing the static regret of adversarial linear mixture MDPs with the unknown transition kernel. From lower bound in He et al. [33, Theorem 5.3], we have the following lower bound in this case since the dynamic regret is no smaller than the static regret.

\[
\text{D-Regret}(K) \geq \Omega(\sqrt{d^2H^3K}). \tag{26}
\]

Then, we consider the lower bound of optimizing the dynamic regret of adversarial linear mixture MDPs with the known transition kernel. We note that Zimin and Neu [12] show the lower bound of the static regret for adversarial episodic loop-free SSP with known transition kernel is \( \Omega(H\sqrt{K \log(SA)}) \), we utilize this lower bound to establish our lower bound as the episodic loop-free SSP is a special case of linear mixture MDPs with \( d = S^2A \). We consider two cases:

Case 1: \( \Gamma \leq 2H \). In this case, we can directly utilize the established lower bound of static regret as a natural lower bound of dynamic regret,
\[
\text{D-Regret}(K) \geq \Omega(H\sqrt{K \log(SA)}). \tag{27}
\]

Case 2: \( \Gamma > 2H \). Without loss of generality, we assume \( L = \lceil \Gamma/2H \rceil \) divides \( K \) and split the whole episodes into \( L \) pieces equally. Next, we construct a special policy sequence such that the policy sequence is fixed within each piece and only changes in the split point. Since the sequence changes at most \( L - 1 \leq \Gamma/2H \) times and the path length of the policy sequence at each change point is at most \( 2H \), the total path length in \( K \) episodes does not exceed \( \Gamma \). As a result, we have
\[
\text{D-Regret}(K) \geq LH\sqrt{K/L \log(SA)} = H\sqrt{KL \log(SA)} \geq \Omega(\sqrt{KH^2 \log(SA)}). \tag{28}
\]
Combining (27) and (28), we have the following lower bound for the dynamic regret of adversarial linear mixture MDPs with the known transition kernel,
\[
\text{D-Regret}(K) \geq \Omega(\max\{H\sqrt{K \log(SA)}, \sqrt{KH^2 \log(SA)}\}) \geq \Omega(\sqrt{KH(H + \Gamma) \log(SA)}), \tag{29}
\]
where the last inequality holds by \( \max\{a, b\} \geq (a + b)/2 \).

Combining two lower bounds (26) and (29), we have the lower bound of the dynamic regret of adversarial linear mixture MDPs with the unknown transition kernel,
\[
\text{D-Regret}(K) \geq \Omega(\max\{\sqrt{d^2H^3K}, \sqrt{KH(H + \Gamma) \log(SA)}\}) \geq \Omega(\sqrt{d^2H^3K} + \sqrt{KH(H + \Gamma) \log(SA)}).
\]
This finishes the proof.
C.9 Proof of Theorem 5

Proof. By Lemma 5, we can rewrite the dynamic regret as follows.

\[
\text{D-Regret}(K) = \sum_{k=1}^{K} V_{k,1}^{\pi_k} - \sum_{k=1}^{K} V_{k,1}^{\pi_k} \\
= \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k} [(Q_{k,h}(s_h, \cdot), \pi_k^* (\cdot | s_h) - \pi_{k,h}(\cdot | s_h))] \\
+ \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k} [\pi_k^* (\cdot | s_h) - \pi_{k,h}(\cdot | s_h)] + \mathcal{M}_{K,H}.
\]

By Lemma 10, we have \( t_{k,h}(s,a) \leq 0 \) for any \( k \in [K] \), \( h \in [H] \), \( s \in \mathcal{S} \), \( a \in \mathcal{A} \). Thus, we have

\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k} [\pi_k^* (\cdot | s_h) - \pi_{k,h}(\cdot | s_h)] \leq 0.
\]

By Theorem 2, we have

\[
\sum_{k=1}^{K} \sum_{h=1}^{H} (-t_{k,h}(s_h, a_k)) + \mathcal{M}_{K,H} \leq \tilde{O}\left(\sqrt{dH^4K + d^2H^3} \right).
\]

It remains to bound the first term. We decompose this term as follows.

\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k} [(Q_{k,h}(s_h, \cdot), \pi_k^* (\cdot | s_h) - \pi_{k,h}(\cdot | s_h))] \\
= \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k} [(Q_{k,h}(s_h, \cdot), \pi_k^* (\cdot | s_h) - \pi_{k,h}(\cdot | s_h))] \\
+ \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k} [\pi_k^* (\cdot | s_h) - \pi_{k,h}(\cdot | s_h)],
\]

where the decomposition holds for any base-learner \( i \in N \). Next, we bound the two terms separately.

Upper bound of base regret. From Theorem 1, we have

\[
\text{base-regret} \leq \frac{\eta_i KH^3}{2} + \frac{1}{\eta_i} \left( H \log A + (1 + H) \log A \gamma P_T + KH \log \frac{1}{1 - \gamma} \right).
\]

Set \( \gamma = 1/KH \) and note that \( \log(1/(1-\gamma)) \leq \gamma/(1-\gamma) \) for all \( \gamma > 0 \). Then, we have

\[
\text{base-regret} \leq \frac{\eta_i KH^3}{2} + \frac{1}{\eta_i} \left( H \log A + (1 + H) P_T \log(KHA) + 2 \right) \\
\leq \frac{\eta_i KH^3}{2} + \frac{2H}{\eta_i} \left( \log A + P_T \log(KHA) \right).
\]

It is clear that the optimal learning rate \( \eta^* = \sqrt{4(\log A + P_T \log(KHA))/(KH^2)} \). By the definition \( P_T = \sum_{k=1}^{K} \sum_{h=1}^{H} \| \pi_k^* - \pi_{k-1,h}^* \|_1 \), it holds that \( 0 \leq P_T \leq 2KH \). Therefore, the range of the optimal learning rate is

\[
\eta_{\text{min}} = \sqrt{\frac{4 \log A}{KH^2}}, \text{ and } \eta_{\text{max}} = \sqrt{\frac{4 \log A + 8KH \log(KHA)}{KH^2}}.
\]

From the construction of the step size pool \( \mathcal{H} = \{ \eta_i = (2^i/H)^{\sqrt{(\log A)}/K} \mid i \in [N] \} \) with \( N = \left\lfloor \frac{1}{2} \log_{\frac{K}{\log \pi^*}} \right\rfloor \), we know that the step size therein is monotonically increasing, in particular

\[
\eta_1 = \sqrt{\frac{4 \log A}{KH^2}}, \text{ and } \eta_N = \frac{1}{H}.
\]

In the following, we consider two cases:
Case 1: $\eta^* \in [\eta_1, \eta_N]$. In this case, we ensure there exists $i^* \in N$ such that $\eta_{i^*} \leq \eta^* \leq 2\eta_{i^*}$. Note the decomposition in (30) holds for any base-learner. Therefore, we choose the base-learner whose step size is $\eta_{i^*}$ and have
\[
\text{base-regret} \leq \frac{\eta_1 \eta_{i^*} K H^3}{2} + \frac{2H}{\eta_{i^*}} (\log A + P_T \log(KHA)) \\
\leq \frac{\eta^* K H^3}{2} + \frac{4H}{\eta^*} (\log A + P_T \log(KHA)) \\
= 3\sqrt{K H^4 (\log A + P_T \log(KHA))},
\]
where the second inequality holds by the condition that $\eta_{i^*} \leq \eta^* \leq 2\eta_{i^*}$ and the last equality holds by substituting $\eta^* = \sqrt{4(\log A + P_T \log(KHA))/(K H^2)}$.

Case 2: $\eta^* > \eta_N$. In this case, we know that $4(\log A + P_T \log(KHA)) > K$. Therefore, we choose the base-learner whose step size is $\eta_N$ and have
\[
\text{base-regret} \leq \frac{\eta_N K H^3}{2} + \frac{2H}{\eta_N} (\log A + P_T \log(KHA)) \\
= \frac{K H^2}{2} + 2H^2 (\log A + P_T \log(KHA)) \\
\leq 4H^2 (\log A + P_T \log(KHA)),
\]
where the last inequality holds by the condition that $4(\log A + P_T \log(KHA)) > K$.

Summing over the two upper bounds yields
\[
\text{base-regret} \leq 3\sqrt{K H^4 (\log A + P_T \log(KHA))} + 4H^2 (\log A + P_T \log(KHA)) \\
\leq 4\sqrt{H^4(K + \log A + P_T \log(KHA))(\log A + P_T \log(KHA))} \\
= \hat{O}(\sqrt{H^4(K + P_T)(1 + P_T)}),
\]
where the inequality holds by $\sqrt{a + \sqrt{b}} \leq \sqrt{2(a + b)}, \forall a, b \geq 0$.

Upper bound of meta regret. For meta-regret, we have
\[
\text{meta-regret} = K \sum_{k=1}^{H} \sum_{h=1}^{H} \mathbb{E}_{\pi_k} \left[ (Q_{k,h}(s_h, \cdot), \pi_k(s_h) - \pi_{k,h}(\cdot | s_h)) \right] \\
= K \sum_{k=1}^{H} \sum_{h=1}^{H} \mathbb{E}_{\pi_k} \left[ (e^{i_k,h}(s_h) - p_{k,h}(s_h), Q_{k,h}(s_h, \cdot), \pi_k(s_h) - \pi_{k,h}(\cdot | s_h)) \right] \\
\leq K \sum_{k=1}^{H} \sum_{h=1}^{H} \mathbb{E}_{\pi_k} \left[ (e^{i_k,h}(s_h) - p_{k,h}(s_h), Q_{k,h}(s_h, \cdot), \pi_k(s_h) - \pi_{k,h}(\cdot | s_h)) \right] \\
\leq K \sum_{k=1}^{H} \sum_{h=1}^{H} \mathbb{E}_{\pi_k} \left[ ||e^{i_k,h}(s_h) - e^{i_{k-1,h}}(s_h)||_1 \right] \\
\leq 2HS_T,
\]
where the first inequality holds by the definition that $i_{k,h} = \arg \max_{i \in [N]} (Q_{k,h}(s_h, \cdot), \pi_k(s_h) - \pi_{k,h}(\cdot | s_h))$, the second inequality holds due to $p_{k,h}(s_h) = e^{i_{k-1,h}}(s_h)$, and the last equality holds by the definition $S_T = \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k} \mathbb{I}[i_{k,h}(s_h) = i_{k-1,h}(s_h)]$.

Combining (31) and (32), by $\sqrt{a + \sqrt{b}} \leq \sqrt{2(a + b)}, \forall a, b \geq 0$, we have
\[
\text{D-Regret}(K) \leq \hat{O}\left( \sqrt{dH^4 K + d^2 H^3 K} + \sqrt{H^4(K + P_T)(1 + P_T) + H^2 S_T^2} \right).
\]
This finishes the proof. 

\[\Box\]
D Supporting Lemmas

In this section, we introduce the supporting lemmas used in the proofs.

First, we introduce the performance difference lemma which connects the difference between two policies to the difference between their expected total rewards through the Q-function.

**Lemma 13** (Cai et al. [32, Lemma 3.2]). For any policies $\pi, \pi' \in \Delta(A | S, H)$, it holds that

$$V_{\pi'}(s_{K,1}) - V_{\pi}(s_{K,1}) = \mathbb{E}_{\pi'} \left[ \sum_{h=1}^{H} Q_{K,h}(s_{h},\cdot), \pi'_h(\cdot | s_h) - \pi_h(\cdot | s_h) \right] s_1 = s_{K,1}.$$ 

Then, we introduce the following lemmas which show the “one-step descent” guarantee.

**Lemma 14** (Cai et al. [32, Lemma 3.3]). For any distributions $p^*, p \in \Delta(A)$, state $s \in S$, and function $Q : S \times A \to [0, H]$, it holds for $p' \in \Delta(A)$ with $p'(\cdot) \propto p(\cdot) \cdot \exp(\eta \cdot Q(s, \cdot))$ that

$$\langle Q(s, \cdot), p^*(\cdot) - p(\cdot) \rangle \leq \eta H^2/2 + \eta^{-1} \cdot (D_{KL}(p^*(\cdot)||p(\cdot)) - D_{KL}(p^*(\cdot)||p'(\cdot))).$$

Next, we introduce the law of total variance, which bounds the variance of the value function.

**Lemma 15** (Jin et al. [55, Lemma C.5]). With probability at least $1 - \delta$, it holds that

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \left[ \mathbb{V}_h V_{K+1}^{\pi^*} \right] (s_{h}^k, a_{h}^k) \leq 3 \left( HK + H^3 \log \frac{1}{\delta} \right).$$

Finally, we introduce the elliptical potential lemma, which is a key lemma in online linear regression.

**Lemma 16** (Abbasi-Yadkori et al. [51, Lemma 11]). Let $\{x_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$ space, $V_0 = \lambda I$ and define $V_t = V_0 + \sum_{i=1}^{t} x_i x_i^\top$. If $\|x_i\|_2 \leq L, \forall i \in \mathbb{Z}_+$, then for each $t \in \mathbb{Z}_+$,

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