Dynamic Regret of Adversarial MDPs with Unknown Transition and Linear Function Approximation

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 (RL) in episodic MDPs with adversarial full-information losses and the unknown transition. Instead of the classical static regret, we adopt *dynamic regret* as the performance measure which benchmarks the learner's performance with *changing* policies, making it more suitable for non-stationary environments. The primary challenge is to handle the uncertainties of unknown transition and unknown non-stationarity of environments simultaneously. We propose a general framework to decouple the two sources of uncertainties and show the dynamic regret bound naturally decomposes into two terms, one due to constructing confidence sets to handle the unknown transition and the other due to choosing sub-optimal policies under the unknown non-stationarity. To this end, we first employ the two-layer online ensemble structure to handle the adaptation error due to the unknown non-stationarity, which is model-agnostic. Subsequently, we instantiate the framework to three fundamental MDP models, including tabular MDPs, linear MDPs and linear mixture MDPs, and present corresponding approaches to control the exploration error due to the unknown transition. We provide dynamic regret guarantees respectively and show they are optimal in terms of the number of episodes K and the non-stationarity $\bar{P}_K$ by establishing matching lower bounds. To the best of our knowledge, this is the first work that achieves the dynamic regret exhibiting optimal dependence on K and $\bar{P}_K$ *without* prior knowledge about the non-stationarity for adversarial MDPs with unknown transition.

Introduction

Reinforcement learning studies the problem where a learner interacts with the environment sequentially and aims to improve her strategy over time (Sutton and Barto 1998). The dynamics of the environment are typically modeled as a Markov Decision Process (MDP) (Puterman 1994). Much of the existing literature on MDPs assumes the losses and dynamics of the environment are stationary over time (Jaksch, Ortner, and Auer 2010; Azar, Osband, and Munos 2017; Jin et al. 2018). However, in real-world applications, the loss functions might change over time, even could potentially be adversarial chosen by the environments (Zhou 2022).

To better capture applications with non-stationary or even adversarial losses, the seminal work of Even-Dar, Kakade, and Mansour (2009) first studies the problem of learning adversarial MDPs, where the losses can change arbitrarily. There is a line of subsequent work studying adversarial MDPs (Zimin and Neu 2013; Rosenberg and Mansour 2019; Jin et al. 2020a), which studies various settings depending on whether the transition kernel is known, and whether the feedback is full-information or bandit. We focus on the unknown transition and full-information setting where the loss function is revealed to the learner after each episode ends.

One limitation of the above studies is that they choose static regret as the performance measure, which is defined as the performance difference between the learner's policy $\pi_1, \ldots, \pi_K$ and that of the best-fixed policy, namely,

$$\mathrm{Reg}_K = \sum_{k=1}^{K} L_k(\pi_k) - \min_{\pi \in \Pi} \sum_{k=1}^{K} L_k(\pi), \tag{1}$$

where $L_k(\pi_k)$ is the expected loss of policy π_k at episode k, and Π is the set of all stochastic policies. However, the *best-fixed* policy may behave poorly in non-stationary environments. To this end, following previous studies (Zhao, Li, and Zhou 2022; Li, Zhao, and Zhou 2023), we choose *dynamic regret* as the performance measure, which benchmarks the learner's performance with changing policies, defined as

$$\text{D-Reg}_K(\pi_1^c, \ldots, \pi_K^c) = \sum_{k=1}^{K} L_k(\pi_k) - \sum_{k=1}^{K} L_k(\pi_k^c), \tag{2}$$

where $\pi_1^c, \ldots, \pi_K^c \in \Pi$ is any policy sequence, which can be chosen by taking into account complete knowledge of the online loss functions. An upper bound of dynamic regret is expected to scale with a certain variation quantity of the compared policies denoted by $P_K(\pi_1^c, \ldots, \pi_K^c)$, which reflects the degree of the environmental non-stationarity.

The dynamic regret measure in (2) is very powerful and general due to the flexibility of compared policies. For example, it immediately recovers the standard regret in (1) when choosing the single best policy in hindsight, that is, $\forall k \in [K], \pi_k^c = \pi^* = \arg\min_{\pi \in \Pi} \sum_{k=1}^{K} L_k(\pi)$. Another typical choice for the compared policies is the sequence of the best policy of each episode, namely, $\forall k \in [K], \pi_k^c = \pi_k^* = \arg\min_{\pi \in \Pi} L_k(\pi)$, which is studied in the work of Fei et al. (2020); Zhong et al. (2021) and we refer it as *worst-case* dynamic regret. A dynamic regret bound with respect to (2) implies a worst-case dynamic regret bound directly.

While the flexibility of dynamic regret makes it more appropriate in non-stationary environments, it brings great challenges at the same time. This is due to that we need to establish a universal guarantee that holds for any sequence of comparators. Few studies focus on this measure in the literature. Zhao, Li, and Zhou (2022) first investigate the dynamic regret of adversarial MDPs with full-information feedback but importantly *known* transition. In particular, Zhao, Li, and Zhou (2022) study the tabular MDPs, which are not scalable to large-scale MDPs. Later, Li, Zhao, and Zhou (2023) study the linear mixture MDPs with the *unknown* transition setting. They propose a policy optimization algorithm with the optimal dynamic regret when the non-stationarity of environments is *known*. Furthermore, they propose a meta-base two-layer framework for situations where the non-stationarity of environments is *unknown*, though their dynamic regret bound in this case suffers an additional term about the switching number of the best base-learner which can be linear of K and ruin the final bound in the worst case.

In this work, we study the dynamic regret of adversarial MDPs with the *unknown* transition and *unknown* non-stationarity of the environment. We investigate tabular MDPs as well as linear MDPs and linear mixture MDPs to deal with large-scale MDPs. For three MDP models, we propose corresponding algorithms equipped with dynamic regret guarantees respectively. We show that our dynamic regret bounds are optimal in terms of the number of episodes K and the non-stationarity measure $\bar{P}_K$ by establishing matching lower bounds. To the best of our knowledge, this is the first work that achieves the dynamic regret with optimal dependence on K and $\bar{P}_K$ for adversarial MDPs with the unknown transition and unknown non-stationarity.

Our contributions are summarized as follows. The primary challenge is to handle the uncertainties of unknown transition and unknown non-stationarity simultaneously. We propose a general framework to decouple the two sources of uncertainties and show the dynamic regret naturally decomposes into two terms, one due to constructing confidence sets to handle the unknown transition and the other due to choosing sub-optimal policies under the unknown non-stationarity. This decomposition highlights two main components RL algorithms need to perform well in non-stationary environments: exploration to deal with the unknown transition and adaptation to handle the adversarial losses. Thus, we first employ the two-layer structure to handle the adaptation error due to the unknown non-stationarity. Then we instantiate the framework to three classical MDP models and present corresponding methods to control the exploration error due to the unknown transition. Though the two-layer online ensemble structure is also used in Li, Zhao, and Zhou (2023), we use occupancy-measure-base method rather than policy optimization to update policies, leading to a dynamic regret with optimal dependence on K and $\bar{P}_K$.

The rest is organized as follows. We start with a review of related works. Then we present the setup and introduce a general framework. Next, we instantiate it to three MDPs and establish dynamic regret upper and lower bounds respectively. Finally, we conclude the paper. Due to limited space, detailed proofs will be presented in a longer version.

Related Work

RL with Adversarial Losses. Learning the static regret of RL with adversarial losses has been well-studied in the literature (Zimin and Neu 2013; Rosenberg and Mansour 2019; Jin et al. 2020a; Cai et al. 2020; Zhou, Gu, and Szepesvári 2021; Chen, Luo, and Wei 2021; He, Zhou, and Gu 2022; He et al. 2023). In general, these works can be divided into three lines based on the structure of the MDPs. The first line of work studies the tabular MDPs, yet with various settings depending on whether the transition is known, and whether the feedback is full-information or bandit. In particular, the pioneering studies of Even-Dar, Kakade, and Mansour (2009) and Yu, Mannor, and Shimkin (2009) investigate the infinite-horizon MDPs with known transition and full-information feedback. For episodic MDPs with known transition, Zimin and Neu (2013) propose the O-REPS algorithm, which achieves (near) optimal regret in both full-information and bandit feedback settings. Rosenberg and Mansour (2019) and Jin et al. (2020a) further study the harder unknown transition and bandit feedback setting. The second line of work studies the linear mixture MDPs, where the transition kernel can be parameterized as a linear function of a state-action-state feature mapping. Cai et al. (2020) first study adversarial linear mixture MDPs in unknown transition but full-information feedback setting and He, Zhou, and Gu (2022) improve the result to (near) optimal. Later, Zhao et al. (2023) extend the study to the bandit feedback setting. The last line of work considers the linear MDPs where both the transition and the loss function can be parameterized as linear functions of a given state-action feature mapping. Neu and Olkhovskaya (2021) study adversarial linear MDPs in the known transition and bandit feedback setting. Zhong and Zhang (2023) and Sherman et al. (2023) investigate adversarial linear MDPs in the unknown transition but full-information setting. Luo, Wei, and Lee (2021) make the first step to establish a sublinear regret guarantee for adversarial linear MDPs in the unknown transition and bandit feedback setting. The result is further improved by Dai et al. (2023); Sherman, Koren, and Mansour (2023); Kong et al. (2023).

Non-stationary RL. Another relevant research area focuses on non-stationary MDPs. Unlike adversarial MDPs where losses are generated in an adversarial manner, non-stationary MDPs address scenarios where losses are stochastically generated from varying distributions. Jaksch, Ortner, and Auer (2010) and Gajane, Ortner, and Auer (2018) study the piecewise stationary setting, where the transition kernel and losses are permitted to change at certain times. Subsequently, Ortner, Gajane, and Auer (2019) extend the previous setting to allow changes to occur at every step. However, these works rely on prior knowledge of the non-stationarity. To overcome this limitation, Cheung, Simchi-Levi, and Zhu (2020) introduce the Bandit-over-RL algorithm, which eliminates this requirement. In a recent breakthrough, Wei and Luo (2021) propose a black-box method capable of converting any algorithm with optimal static regret that satisfies specific conditions into another algorithm that achieves optimal dynamic regret in non-stationary environments without prior knowledge about the degree of the non-stationarity of envi-

ronments. However, it is not applicable in the adversarial setting since the approach of constructing an optimistic estimator to detect environmental change by a UCB-type algorithm can only be applied effectively in the stochastic setting.

Dynamic Regret. Dynamic Regret has been studied under the settings of bandits (Auer, Gajane, and Ortner 2019; Chen et al. 2019; Luo et al. 2022; Wang, Zhao, and Zhou 2023), online convex optimization (Zinkevich 2003; Zhang, Lu, and Zhou 2018; Zhao et al. 2020; Baby and Wang 2021), stochastic programming (Besbes, Gur, and Zeevi 2015; Chen, Wang, and Wang 2019), linear control systems (Zhao, Wang, and Zhou 2022; Yan, Zhao, and Zhou 2023) and online distribution shift (Bai et al. 2022; Qin et al. 2023; Qian et al. 2023). For adversarial MDPs, Fei et al. (2020) study the *worst-case* dynamic regret of adversarial tabular MDPs with unknown transition and full-information feedback. Zhong et al. (2021) extend the algorithm of Fei et al. (2020) to accommodate non-stationary transitions with linear function approximation. However, both their algorithms require prior knowledge about the non-stationarity as input. Even so, their dynamic regret bounds are still suboptimal. Zhao, Li, and Zhou (2022) make the first step to study the dynamic regret in (2). They investigate the dynamic regret of adversarial *tabular* MDPs with the *known* transition and present algorithms with optimal dynamic regret without prior knowledge about the non-stationarity. For the more challenging unknown transition setting studied in this work, the only previous work (Li, Zhao, and Zhou 2023) studies linear mixture MDPs with unknown transition and proposes a policy optimization algorithm with optimal dynamic regret when the non-stationarity is *known*. Furthermore, they propose a two-layer ensemble algorithm for situations where the non-stationarity of environments is *unknown*, though their dynamic regret bound in this case suffers an additional term about the switching number of the best base-learner which can be linear of K and ruin the final bound in the worst case. In this work, we obtain the dynamic regret with optimal dependence on K and $\bar{P}_K$ for adversarial tabular MDPs, linear MDPs and linear mixture MDPs with the unknown transition and unknown non-stationarity.

Problem Setup

Episodic Adversarial MDPs. An inhomogeneous MDP is denoted by a tuple $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, H, \{\mathbb{P}_h\}_{h=1}^H, \{\ell_k\}_{k=1}^K\}$, where $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$, K is the number of episodes and H is the horizon, $\mathbb{P}_h(\cdot \mid \cdot, \cdot) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition at stage h, $\ell_{k,h} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the loss function. We assume the MDP has a loop-free structure, satisfying:

- $\mathcal{S}$ is constituted by $H + 1$ disjoint layers $\mathcal{S}_1, \ldots, \mathcal{S}_{H+1}$ with $\mathcal{S} = \cup_{h=1}^{H+1} \mathcal{S}_h$ and $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$ for $i \neq j$.
- $\mathcal{S}_1, \mathcal{S}_{H+1}$ are singletons, $\mathcal{S}_1 = \{s_1\}, \mathcal{S}_{H+1} = \{s_{H+1}\}$.
- Transition can only happen between adjacent layers, i.e., $\forall h \in [H]$, if $\mathbb{P}_h(s' | s, a) > 0$, then $s \in \mathcal{S}_h$ and $s' \in \mathcal{S}_{h+1}$.

These assumptions are common in previous studies (Neu, György, and Szepesvári 2010; Neu et al. 2010; Zimin and Neu 2013). They are not necessary but simplify the notation.

The interaction protocol between the learner and the environment is given as follows. At the beginning of episode $k \in [K]$, the environment chooses a loss function ℓ_k and simultaneously the learner decides a policy $\pi_k : \mathcal{S} \times \mathcal{A} \to [0, 1]$ with $\pi_k(a|s)$ being the probability of taking action a at state s. Starting from the initial state $s_{k,1} = s_1$, the learner repeatedly sample action $a_{k,h}$ from policy $\pi_k(\cdot|s_{k,h})$ and suffers loss $\ell_k(s_{k,h}, a_{k,h})$ until reaching the terminal state $s_{k,H+1}$. We focus on the full-information setting where the entire loss ℓ_k is revealed to the learner after episode k ends. The expected loss of any policy π is denoted by $L_k(\pi) = \mathbb{E}[\sum_{h=1}^H \ell_k(s_{k,h}, a_{k,h})|\mathbb{P}, \pi]$, where the expectation is taken over the randomness of the transition $\mathbb{P}$ and the stochastic policy π. The total step is defined as $T = HK$.

Occupancy Measure. Given policy π and a transition $\mathbb{P}$, the occupancy measure $q^{\mathbb{P},\pi} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is defined as the probability of visiting state-action-state triple (s, a, s') under transition $\mathbb{P}$ and policy π, namely,

$$q^{\mathbb{P},\pi}(s, a, s') = \Pr[s_{h(s)} = s, a_{h(s)} = a, s_{h(s)+1} = s'],$$

where $h(s)$ is the index of the layer of state s (Altman 1998). An occupancy measure q satisfies the following two properties. First, each layer is visited exactly once and thus for $\forall h \in [H]$, $\sum_{s \in \mathcal{S}_h} \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}_{h+1}} q(s, a, s') = 1$. Second, the probability of entering a state when coming from the previous layer is exactly the probability of leaving from that state to the next layer, i.e., for every $s \in \mathcal{S}_h$, $\sum_{s' \in \mathcal{S}_{h-1}} \sum_{a \in \mathcal{A}} q(s', a, s) = \sum_{s' \in \mathcal{S}_{h+1}} \sum_{a \in \mathcal{A}} q(s, a, s')$. For any occupancy measure q satisfying the above two properties, it induces a transition $\mathbb{P}^q$: $\mathbb{P}^q(s' \mid s, a) = q(s, a, s')/\sum_{s'' \in \mathcal{S}_{h(s)+1}} q(s, a, s'')$, a policy π^q: $\pi^q(a \mid s) = \sum_{s' \in \mathcal{S}_{h(s)+1}} q(s, a, s')/\sum_{a' \in \mathcal{A}} \sum_{s' \in \mathcal{S}_{h(s)+1}} q(s, a', s')$. We denote by Δ the set of valid occupancy measures, that is, the set of all occupancy measures satisfying the above two properties. For a fixed transition $\mathbb{P}$, we denote by $\Delta(\mathbb{P}) \in \Delta$ the set of occupancy measures whose induced transition $\mathbb{P}^q$ is exactly $\mathbb{P}$. Further, we denote by $\Delta(\mathcal{P}) \in \Delta$ the set of occupancy measures whose induced transition $\mathbb{P}^q$ is in a set of transition kernels $\mathcal{P}$. With slight abuse of notation, we denote by $q(s, a) = \sum_{s' \in \mathcal{S}_{h(s)+1}} q(s, a, s')$ and define the norm $\|q - q'\|_1 \triangleq \sum_{s,a \in \mathcal{S} \times \mathcal{A}} |q(s, a) - q'(s, a)|$.

Dynamic Regret. It can be verified the dynamic regret in (2) can be written into a form about occupancy measures:

$$\text{D-Reg}_K(\pi_1^c, \ldots, \pi_K^c) = \sum_{k=1}^K \langle q^{\mathbb{P},\pi_k} - q^{\mathbb{P},\pi_k^c}, \ell_k \rangle. \quad (3)$$

We introduce $\pi_0^c = \pi_1^c$ and use $q_k = q^{\mathbb{P},\pi_k}, q_k^c = q^{\mathbb{P},\pi_k^c}$ to simplify the notation. In the work of Zhao, Li, and Zhou (2022), they define two types of non-stationarity measures. The first measure is related to the compared policies and is denoted as $P_K = \sum_{k=1}^K \sum_{h=1}^H \|\pi_{k,h}^c - \pi_{k-1,h}^c\|_{1,\infty}$. The second pertains to the occupancy measures and is represented by $\bar{P}_K = \sum_{k=1}^K \|q_k^c - q_{k-1}^c\|_1$. Zhao, Li, and Zhou (2022, Lemma 6) show that these two measures satisfy $\bar{P}_K \leq HP_K$. We thus focus on the $\bar{P}_K$-type upper bound, which directly implies an upper bound in terms of HP_K.

A General Framework

In this section, we propose a general framework for learning adversarial MDPs with the unknown transition kernel. In this setting, we encounter the challenge of handling uncertainties regarding the unknown transition kernel, as well as the unknown non-stationarity of environments. To tackle this challenge, we introduce a general framework that effectively separates these two sources of uncertainties. At a high level, our framework comprises two key components: the transition estimation step and the policy update step.

Transition Estimation. During the transition estimation step, at episode k, the algorithm utilizes $k-1$ previously observed trajectories to estimate the unknown transition. The underlying approach involves two key steps. Firstly, the algorithm computes the empirical transition kernel $\bar{\mathbb{P}}_k$, which serves as an estimate based on the available historical data. Subsequently, a confidence set $\mathcal{P}_k$ is maintained, which aims to contain the true transition kernel $\mathbb{P}$ with high probability. In subsequent sections, we will delve into the specific details of the transition estimation step for various types of MDPs, highlighting the tailored methodologies for each case.

Policy Update. In the policy update step, we update the policy based on the estimated transition. Following the study for dynamic regret of adversarial tabular MDPs with known transition (Zhao, Li, and Zhou 2022), we apply Online Mirror Descent (OMD) (Orabona 2019) to update the occupancy measure and adopt a two-layer online ensemble structure to address the non-stationarity of environments.

At each episode $k \in [K]$, the basic idea is to perform OMD over the clipped occupancy measure space induced by the confidence set $\mathcal{P}_k$, defined as $\Delta(\mathcal{P}_k, \alpha) = \{q \in \Delta(\mathcal{P}_k)$ and $q(s,a,s') \geq \alpha, \forall s,a,s'\}$ with $0 \leq \alpha < 1$ being the clipping parameter. The clipping operation can be regarded as forcing some amount of uniform exploration to deal with non-stationary environments. The learner updates

$$\widehat{q}_{k+1} = \arg\min_{q \in \Delta(\mathcal{P}_k, \alpha)} \eta \langle q, \ell_k \rangle + D_\psi(q \| \widehat{q}_k), \quad (4)$$

where $D_\psi(q\|q') = \sum_{s,a,s'} q(s,a,s') \log \frac{q(s,a,s')}{q'(s,a,s')}$ is the normalized KL-divergence, $\eta > 0$ is the step size, and we use the notation $\widehat{q}$ to emphasize that the occupancy measure is over the confidence set $\mathcal{P}$ and is not necessarily an occupancy measure over the true transition $\mathbb{P}$. We start by considering the ideal setting where the confidence set *only* contains the true transition kernel, that is $\Delta(\mathcal{P}_k, \alpha) = \Delta(\mathbb{P}, \alpha)$. In this ideal case, Zhao, Li, and Zhou (2022, Lemma 1) show that the dynamic regret of (4) is bounded by

$$\text{D-Reg}_K \leq \eta T + \frac{1}{\eta}(H \log(S^2 A) + \bar{P}_K \log \frac{1}{\alpha}).$$

It is clear that to obtain a favorable dynamic regret, we need to set the step size η optimally to balance the number of steps T and the non-stationary measure $\bar{P}_K$, more specifically, set $\eta \approx \sqrt{(H + \bar{P}_K)/T}$. However, we do not have prior knowledge of $\bar{P}_K$ even after the horizon ends since the compared policies can be arbitrarily chosen. To address this issue, Zhao, Li, and Zhou (2022) adopt a two-layer online ensemble method (Zhou 2012; Zhao et al. 2021).

Algorithm 1: Overall Algorithm framework

Input: step pool $\mathcal{H}$, learning rate ε, and clipping param α.

Set $\widehat{q}_{1,i}(s,a,s') = 1/(S_h \cdot A \cdot S_{h+1}), \forall (s,a,s') \in \mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}$ and $p_{1,i} = 1/N, \forall i \in [N]$.
for $k = 1$ to K **do**
  Receive $\widehat{q}_{k,i}$ from base-learner $\mathcal{B}_i$ for $i \in [N]$.
  Compute $\widehat{q}_k = \sum_{i=1}^N p_{k,i} \widehat{q}_{k,i}$ and play policy $\pi^{\widehat{q}_k}$.
  Obtain trajectory $U_k = \{(s_{k,h}, a_{k,h})\}_{h=1}^H$.
  $\mathcal{P}_k \leftarrow$ EstimateTransition(k, U_k).
  Each base-learner updates by (5).
  Update the weights by (6).
end for

They first construct a step size pool $\mathcal{H} = \{\eta_1, \ldots, \eta_N\}$ ($N = \mathcal{O}(\log T)$) to discretize the range of the optimal step size, then maintain multiple base-learners $\mathcal{B}_1, \ldots, \mathcal{B}_N$, each of which is associated with a step size $\eta_i \in \mathcal{H}$. Finally, they use Hedge algorithm (Freund and Schapire 1997; Cesa-Bianchi and Lugosi 2006) to track the unknown best base-learner. Specifically, the base learner $\mathcal{B}_i$ updates policy by

$$\widehat{q}_{k+1,i} = \arg\min_{q \in \Delta(\mathcal{P}_k, \alpha)} \eta_i \langle q, \ell_k \rangle + D_\psi(q \| \widehat{q}_{k,i}). \quad (5)$$

The meta-algorithm updates weights by

$$p_{k+1,i} \propto p_{k,i} \exp(-\varepsilon \langle \widehat{q}_{k,i}, \ell_k \rangle), \quad (6)$$

where $\varepsilon > 0$ is the learning rate of meta-algorithm, $\langle \widehat{q}_{k,i}, \ell_k \rangle$ evaluates the performance of the base-learner $\mathcal{B}_i$. The final occupancy measure is given by $\widehat{q}_{k+1} = \sum_{i=1}^N p_{k+1,i} \widehat{q}_{k+1,i}$ and the learner plays the policy $\pi = \pi^{\widehat{q}_{k+1}}$. The detailed procedures are summarized in Algorithm 1. Notice that for the general case, the obtained occupancy measure might not be in $\Delta(\mathbb{P}, \alpha)$ since the confidence set $\mathcal{P}_k$ not necessarily only contains the true transition, that is $\Delta(\mathcal{P}_k, \alpha) \neq \Delta(\mathbb{P}, \alpha)$. Nevertheless, we show Algorithm 1 enjoys a favorable dynamic regret guarantee once the occupancy measure difference due to the estimation of the transition is well controlled.

Theorem 1. *Set the step size pool as* $\mathcal{H} = \{\eta_i = 2^{i-1}\sqrt{K^{-1}\log(S^2A/H)} | i \in [N]\}$, *where* $N = \lceil \frac{1}{2}\log(1 + \frac{4K \log T}{\log(S^2A/H)}) \rceil + 1$, *the learning rate* $\varepsilon = \sqrt{(\log N)/(HT)}$ *and the clipping parameter* $\alpha = 1/T^2$. *Suppose* $\Delta(\mathbb{P}, \alpha) \in \Delta(\mathcal{P}_k, \alpha), \forall k \in [K]$, *Algorithm 1 ensures*

$$\text{D-Reg}_K = \sum_{k=1}^K \langle q_k - \widehat{q}_k, \ell_k \rangle - \sum_{k=1}^K \langle \widehat{q}_k - q_k^c, \ell_k \rangle$$

$$\leq \sum_{k=1}^K \|q_k - \widehat{q}_k\|_1 + \mathcal{O}\Big(\sqrt{T(H \log(S^2 A) + \bar{P}_K \log T)}\Big)$$

with non-stationarity measure $\bar{P}_K = \sum_{k=1}^K \|q_k^c - q_{k-1}^c\|_1$.

Remark 1. The algorithmic dependence on T can be removed by the standard doubling trick (Auer, Cesa-Bianchi, and Gentile 2002). Theorem 1 consists of two parts, the first arising from exploration, which deals with the uncertainty of transition, and the other arising from adaptation, which deals with non-stationary environments. We can handle the unknown non-stationarity with the online ensemble framework. It remains to deal with the uncertainty of transition.

The Tabular MDP Case

In this section, we apply our general algorithm to the tabular MDP case. We begin by presenting the approach for constructing the confidence set and demonstrating that the occupancy measures difference resulting from estimating the transition is well controlled. Then, we establish the upper and lower bound of the dynamic regret for tabular MDPs.

To construct the confidence set, the general idea (Burnetas and Katehakis 1997; Jaksch, Ortner, and Auer 2010) is to first compute the empirical transition $\bar{\mathbb{P}}_k$ based on the observed samples and then construct a confidence set which includes all transition functions with bounded total variation compared to the empirical transition. This ensures that the true transition kernel is included within the confidence set with high probability. Specifically, we maintain counters to record the number of visits of each state-action pair (s, a) and each state-action-state triple (s, a, s'). To reduce the computational complexity, the estimation proceeds in a doubling epoch schedule, where a new epoch starts whenever there exists a state-action pair whose counter is doubled and the estimation is only updated at the beginning of each epoch. For epoch $i > 1$, denote by $N_i(s, a)$ and $N_i(s' \mid s, a)$ the total number of visits of pair (s, a) and triple (s, a, s') before epoch i, $i(k)$ the epoch that episode k belongs to. The empirical transition $\bar{\mathbb{P}}_i$ for epoch i is given by

$$\bar{\mathbb{P}}_{i,h}(s' \mid s, a) = \frac{N_{i,h}(s' \mid s, a)}{\max\{1, N_{i,h}(s, a)\}}, \forall h \in [H].$$

The following lemma shows that the empirical transition $\bar{\mathbb{P}}_i$ is close to the true transition kernel $\mathbb{P}$ with high probability.

Lemma 1 (Jaksch, Ortner, and Auer (2010, Lemma 17))**.** *With probability at least $1 - \zeta$, it holds that*

$$\|\mathbb{P}_h(\cdot \mid s, a) - \bar{\mathbb{P}}_{i,h}(\cdot \mid s, a)\|_1 \leq \sqrt{\frac{2S_{h+1} \log(TSA/\zeta)}{\max\{1, N_{i,h}(s, a)\}}}$$

for all epochs and all $(s, a) \in \mathcal{S}_h \times \mathcal{A}, h \in [H]$.

Then, a confidence set for epoch i which includes all transition functions with bounded total variation compared to the empirical transition kernel $\bar{\mathbb{P}}_i$ is given by

$$\mathcal{P}_i = \{\hat{\mathbb{P}} \mid \|\hat{\mathbb{P}}_h(\cdot|s, a) - \bar{\mathbb{P}}_{i,h}(\cdot|s, a)\|_1 \leq \xi_{i,h}(s, a)\}. \quad (7)$$

for all $(s, a) \in \mathcal{S}_h \times \mathcal{A}, h \in [H]$ where $\xi_{i,h}(s, a)$ is set as $\sqrt{2S_{h+1} \log(TSA/\zeta)/\max\{1, N_{i,h}(s, a)\}}$ by Lemma 1.

We present the lemma below which shows the occupancy measure difference in the confidence set is well controlled.

Lemma 2. *With probability at least $1-\zeta$, for any collection of transitions $\{\mathbb{P}_k\}$ such that $\mathbb{P}_k \in \mathcal{P}_{i(k)}$ for all s, we have*

$$\sum_{k=1}^{K} \|q^{\mathbb{P}, \pi_k} - q^{\mathbb{P}_k, \pi_k}\|_1 \leq \mathcal{O}\left(S\sqrt{HAT \log(SAT/\zeta)}\right).$$

Combining Lemma 2 with Theorem 1, we obtain the following dynamic regret bound for tabular MDPs.

Theorem 2. *Set parameters as in Theorem 1, Algorithm 1 with Algorithm 2 as subroutine ensures with probability at least $1-\zeta$, the dynamic regret* D-Reg$_K$ *is upper bounded by*

$$\mathcal{O}\Big(S\sqrt{HAT \log(SAT/\zeta)} + \sqrt{T(H \log(S^2A) + \bar{P}_K \log T)}\Big).$$

Algorithm 2: EstimateTransition (Tabular MDPs)

Input: episode index k, trajectory U_k.
Output: confidence set $\mathcal{P}_k$.

```
for h = 1 to H do
    N_{i,h}(s_{i,h}, a_{i,h}) = N_{i,h}(s_{i,h}, a_{i,h}) + 1
    N_{i,h}(s_{i,h}, a_{i,h}, s_{i,h+1}) = N_{i,h}(s_{i,h}, a_{i,h}, s_{i,h+1}) + 1
    if ∃ N_{i,h}(·,·) ≥ max{1, 2N_{i-1,h}(·,·)} then
        Increase epoch index i = i + 1
        N_{i,h}(·,·,·), N_{i,h}(·,·) = N_{i-1,h}(·,·,·), N_{i-1,h}(·,·)
    end if
end for
Compute confidence set P_{i(k)} as in (7).
```

Remark 2. Setting compared policies $\pi^c_{1:K} = \pi^*(\bar{P}_K = 0)$, Theorem 2 recovers the $\widetilde{\mathcal{O}}(S\sqrt{HAT})$ best so far static regret bound of Rosenberg and Mansour (2019).

Remark 3. Setting compared policies as $\pi^c_{1:K} = \pi^*_{1:K}$, Theorem 2 implies an $\widetilde{\mathcal{O}}(S\sqrt{HAT} + \sqrt{HT(1 + P_K)})$ worst-case dynamic regret, which strictly improves the $\widetilde{\mathcal{O}}(S\sqrt{H^3AT} + H^{\frac{4}{3}}T^{\frac{2}{3}}P_K^{\frac{1}{3}})$ result of Fei et al. (2020).

We finally establish the lower bound in the theorem below.

Theorem 3. *For any online algorithm and any $\gamma \in [0, 2T]$, there exists an episodic loop-free MDP with K episodes, H layers, S states and A actions and a sequence of compared policies $\pi^c_1, \ldots, \pi^c_K$ such that $\bar{P}_K(\pi^c_1, \ldots, \pi^c_K) \leq \gamma$ and*

$$\text{D-Reg}_K \geq \Omega\big(\sqrt{HSAT} + \sqrt{T(H + \gamma)}\big)$$

under the full-information and unknown transition setting.

Remark 4. Combining Theorem 2 and Theorem 3, we see that our dynamic regret bound is (nearly) optimal regarding the dependence on H, A, K, $\bar{P}_K$, yet looses a factor of $\mathcal{O}(\sqrt{S})$. Closing this gap is an important open problem.

The Linear MDP Case

To handle large-scale MDPs, we consider the MDPs with linear function approximation. In this section, we study one of the most popular models, namely, the linear MDPs. First, we introduce the definition and present the method to construct the confidence set. Then, we establish the upper and lower bounds for the dynamic regret of linear MDPs.

Definition 1 (Linear MDPs)**.** An MDP instance $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, H, \{\mathbb{P}_h\}_{h=1}^H, \{\ell_k\}_{k=1}^K\}$ is called an inhomogeneous, episode B-bounded linear MDP, if there exist a *known* feature mapping $\phi(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and *unknown* measures $\mu^*_h \in \mathbb{R}^{S \times d}$ such that for any h and any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ (i) $\mathbb{P}_h(s' \mid s, a) = \mu^*_h(s')\phi(s, a)$; (ii) $\|\phi(s, a)\|_2 \leq 1$, $\|v^\top \mu^*_h\|_2 \leq B$ for $\forall v \in \mathbb{R}^S, \|v\|_\infty \leq 1$.

Remark 5. Different from previous works on linear MDPs (Jin et al. 2020b; He et al. 2023), we do not assume the losses admit a linear structure, which is more general.

We estimate the unknown transition via solving a *multivariate* linear regression problem. Specifically, denote by $\delta_s \in \{0, 1\}^S$ the Dirac measure at s, namely, an one-hot

vector with 1 at the s-th entry and note that $\mathbb{P}_h(\cdot \mid s,a) = \mu_h^*(\cdot)\phi(s,a)$ by definition. For any episode i, $s_{i,h+1}$ is sampled from $\mathbb{P}_h(\cdot \mid s_{i,h}, a_{i,h})$ conditioned on $(s_{i,h}, a_{i,h})$, thus $\delta_{s_{i,h+1}}$ is an unbiased estimate of $\mathbb{P}_h(\cdot \mid s_{i,h}, a_{i,h})$ conditioned on $(s_{i,h}, a_{i,h})$. This leads us to estimate μ_h^* by solving the following ridge linear regression:

$$\mu_{k,h} = \arg\min_{\mu \in \mathbb{R}^{S\times d}} \sum_{i=0}^{k-1} \left\| \mu\phi\left(s_{i,h}, a_{i,h}\right) - \delta_{s_{i,h+1}} \right\|_2^2 + \lambda \|\mu\|_F^2.$$

The closed-form solution of $\mu_{k,h}$ is $\mu_{k,h} = \Gamma_{k,h}\Lambda_{k,h}^{-1}$ with

$$\begin{aligned} \Lambda_{k,h} &= \sum_{i=1}^{k-1} \phi(s_{i,h}, a_{i,h})\phi(s_{i,h}, a_{i,h})^\top + \lambda I, \\ \Gamma_{k,h} &= \sum_{i=1}^{k-1} \delta_{s_{i,h+1}} \phi\left(s_{i,h}, a_{i,h}\right)^\top \end{aligned} \tag{8}$$

We present the following lemma, which shows that $\mu_{k,h}$ is close to the true parameter μ_h^* with high probability.

Lemma 3. *Let $\zeta \in (0,1)$, $\forall k \in \mathbb{N}$ and simultaneously $\forall h \in [H]$, with probability at least $1-\zeta$, it holds that $\mu_h^* \in \mathcal{C}_{k,h}$, where $\mathcal{C}_{k,h} = \{\mu \in \mathbb{R}^{S\times d} \mid \|\mu^\top(\cdot) - \mu_{k,h}^\top(\cdot)\|_{\Lambda_{k,h}} \le \beta_{k,h}\}$ with $\beta_{k,h} = B\sqrt{\lambda} + \sqrt{2\log\left(H/\zeta\right) + \log\left(\det\left(\Lambda_{k,h}\right)/\lambda^d\right)}$.*

We construct the confidence set $\mathcal{P}_k = \{\mathcal{P}_{k,h}\}_{h=1}^H$ below.

$$\mathcal{P}_k = \{\hat{\mathbb{P}} \mid \exists \mu \in \mathcal{C}_{k,h}, \hat{\mathbb{P}}_h\left(s' \middle| s,a\right) = \mu(s')\phi\left(s,a\right)\} \tag{9}$$

for all $(s,a,s') \in \mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}, h \in [H]$. The detailed algorithm is presented in Algorithm 3. We show the following lemma which bounds the occupancy measure difference.

Lemma 4. *With probability at least $1-\zeta$, for any collection of transitions $\{\mathbb{P}_k\}$ such that $\mathbb{P}_k \in \mathcal{P}_k$ for all s, we have*

$$\sum_{k=1}^K \|q^{\mathbb{P},\pi_k} - q^{\mathbb{P}_k,\pi_k}\|_1 \le \mathcal{O}\Big(dHS\sqrt{K\log^2(K/\zeta)}\Big).$$

Theorem 4. *Set parameters as in Theorem 1, Algorithm 1 with Algorithm 3 as subroutine ensures with probability at least $1-\zeta$, the dynamic regret D-Reg_K is upper bounded by*

$$\mathcal{O}\Big(dHS\sqrt{K\log^2(K/\zeta)} + \sqrt{T(H\log\left(S^2A\right) + \bar{P}_K\log T)}\Big).$$

Remark 6. Compared with the result in Theorem 2, we replace $\widetilde{\mathcal{O}}(HS\sqrt{AK})$ term with $\widetilde{\mathcal{O}}(dHS\sqrt{K})$, which enjoys better guarantees when $d \le \sqrt{A}$. Our result is independent of A though still has an undesirable dependence on S.

We finally establish the lower bound in the theorem below.

Theorem 5. *For any online algorithm and any $\gamma \in [0, 2T]$, there exists an episodic loop-free linear MDP with K episodes, H layers, S states and A actions and a policy sequence $\pi_1^c, \ldots, \pi_K^c$ such that $\bar{P}_K(\pi_1^c, \ldots, \pi_K^c) \le \gamma$ and*

$$\text{D-Reg}_K \ge \Omega\big(dH\sqrt{K} + \sqrt{T(H+\gamma)}\big)$$

under the full-information and unknown transition setting.

Algorithm 3: EstimateTransition (Linear MDPs)

Input: Episode index k, trajectory U_k.
Output: Confidence set $\mathcal{P}_k$.
1: **for** $h = 1$ to H **do**
2: Set $\Lambda_{k,h}$ and $\Gamma_{k,h}$ as in (8), $\mu_{k,h} = \Gamma_{k,h}\Lambda_{k,h}^{-1}$.
3: **end for**
4: Set the confidence set as (9).

Remark 7. Combining Theorem 4 and 5, we claim that our dynamic regret bound is (nearly) optimal regarding the dependence on d, H, K, $\bar{P}_K$, yet looses a factor of $\mathcal{O}(S)$. The main challenge to eliminate the dependence on S is that though the transition $\mathbb{P}$ admits a linear structure, it is not the case for the occupancy measure $q^{\mathbb{P},\pi_k}$, which has a complicated recursive form. Investigating the possibility of eliminating the dependence on S is left as future work.

The Linear Mixture MDP Case

In this section, we consider another popular function approximation model, the linear mixture MDPs. We first introduce the linear mixture MDPs and then present the method to construct the confidence set. Finally, we establish the upper and lower bound for the dynamic regret of linear mixture MDPs.

Definition 2 (Linear mixture MDPs)**.** An MDP instance $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, H, \{\mathbb{P}_h\}_{h=1}^H, \{\ell_k\}_{k=1}^K\}$ is called an inhomogeneous, episode B-bounded linear mixture MDP, if there exist a *known* feature mapping $\phi(s' \mid s,a) : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}^d$ and an *unknown* vector $\theta_h^* \in \mathbb{R}^d$ such that for any $h \in [H]$ and any $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$: (i) $\mathbb{P}_h(s' \mid s,a) = \phi(s' \mid s,a)^\top\theta_h^*$; (ii) $\|\phi(s' \mid s,a)\|_2 \le 1, \|\theta_h^*\|_2 \le B$.

To estimate the unknown transition parameter θ^*, value-targeted regression (VTR) is a popular approach in previous works studying linear mixture MDPs (Ayoub et al. 2020; Zhou, Gu, and Szepesvári 2021; He, Zhou, and Gu 2022; Ji et al. 2023). Specifically, for any function $V : \mathcal{S} \to [0,H]$, let $\phi_V(s,a) = \sum_{s'}\phi(s' \mid s,a)V(s')$. With the observation that $\mathbb{P}_h(\cdot \mid s,a)^\top V = \langle \phi_V(s,a), \theta_h^*\rangle$, existing works regard it as solving a "linear bandit" problem (Abbasi-Yadkori, Pál, and Szepesvári 2011; Lattimore and Szepesvári 2020) where the context is $\phi_V(s_{k,h}, a_{k,h})$ and the noise is $V(s_{k,h+1}) - [\mathbb{P}_h V](s_{k,h}, a_{k,h})$. Thus, this approach can be viewed as a "model-free" method as it bypasses the need for fully learning the transition $\mathbb{P}$ and only requires $[\hat{\mathbb{P}}_h V](s,a) = [\mathbb{P}_h V](s,a)$. However, this approach makes it hard to control the occupancy measure difference. To overcome this challenge, following previous study (Zhao et al. 2023), we employ a "model-based" method that fully utilizes the transition information to learn the parameter θ^*.

Denote by $\delta_s \in \{0,1\}^S$ the Dirac measure at s, namely, an one-hot vector with 1 at the s-th entry. Since $\mathbb{P}_h(\cdot|s,a) = \phi(\cdot \mid s,a)^\top\theta_h^*$, we use $\phi(\cdot \mid s_{k,h}, a_{k,h})$ as the context and $\delta_{s_{k,h+1}}$ as the regression target to learn the transition parameter θ_h^*. One may consider using all the state information to learn θ_h^*, as $\phi(\cdot \mid s,a)$ and $\delta_s(s')$ is known for any (s,a,s'). However, there is still one obstacle to be addressed. Particularly, let $\epsilon_{i,h} = \mathbb{P}_h(\cdot \mid s_{i,h}, a_{i,h}) - \delta_{s_{i,h+1}}$ be the noise at

episode i and stage h. Then it is clear that $\epsilon_{i,h} \in [-1,1]^S$, $\mathbb{E}_{i,h}[\epsilon_{i,h}] = \mathbf{0}$ and $\sum_{s\in\mathcal{S}_{h+1}} \epsilon_{i,h}(s) = 0$. Therefore, the noise $\epsilon_{i,h}$ is not independent across different states. To further address this issue, we use the transition information of only one state $s'_{i,h+1}$ in the next layer. The most direct idea is choosing the true next state $s_{i,h+1}$ experienced by the learner as $s'_{i,h+1}$. However, it is hard for this approach to control the estimation error of all states. Instead, we choose the next state $s'_{i,h+1}$ with the largest uncertainty. More specifically, we construct the estimator $\theta_{k,h}$ of θ^*_h by finding the minimizer of the following objective function:

$$\sum_{i=1}^{k-1} \left[\left\langle \phi\left(s'_{i,h+1}|s_{i,h},a_{i,h}\right),\theta\right\rangle - \delta_{s_{i,h+1}}\left(s'_{i,h+1}\right)\right]^2 + \lambda\|\theta\|_2^2,$$

and we choose $s'_{i,h+1}$ as

$$s'_{i,h+1} \in \arg\max_{s\in\mathcal{S}_{h+1}} \|\phi(s \mid s_{i,h},a_{i,h})\|_{M_{i,h}^{-1}}. \quad (10)$$

The closed-form solution of $\theta_{k,h}$ is $\theta_{k,h} = M_{k,h}^{-1} b_{k,h}$ with

$$M_{k,h} = \sum_{i=1}^{k-1} \phi(s'_{i,h+1}|s_{i,h},a_{i,h})\phi(s'_{i,h+1}|s_{i,h},a_{i,h})^\top + \lambda I,$$
$$b_{k,h} = \sum_{i=1}^{k-1} \delta_{s_{i,h+1}}(s'_{i,h+1})\phi(s'_{i,h+1} \mid s_{i,h},a_{i,h}). \quad (11)$$

We present the following lemma which shows $\theta_{k,h}$ is close to the true transition parameter θ^*_h with high probability.

Lemma 5. *Let $\zeta \in (0,1)$, then $\forall k \in \mathbb{N}$, and simultaneously $\forall h \in [H]$, with probability at least $1-\zeta$, it holds that $\theta^*_h \in \mathcal{C}_{k,h}$, where $\mathcal{C}_{k,h} = \{\theta \in \mathbb{R}^d \mid \|\theta - \theta_{k,h}\|_{M_{k,h}} \le \beta_{k,h}\}$ with $\beta_{k,h} = B\sqrt{\lambda} + \sqrt{2\log(H/\zeta) + \log(\det(M_{k,h})/\lambda^d)}$.*

Then the confidence set $\mathcal{P}_k = \{\mathcal{P}_{k,h}\}_{h=1}^H$ is given by

$$\mathcal{P}_k = \{\hat{\mathbb{P}} \mid \exists\theta \in \mathcal{C}_{k,h}, \hat{\mathbb{P}}_h(s'|s,a) = \phi(s'|s,a)^\top \theta\} \quad (12)$$

for all $(s,a,s') \in \mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}, h \in [H]$. The detailed algorithm is presented in Algorithm 4. We provide the following lemma which shows the occupancy measure difference in the confidence set can be upper bounded.

Lemma 6. *With probability at least $1-\zeta$, for any collection of transitions $\{\mathbb{P}_k\}$ such that $\mathbb{P}_k \in \mathcal{P}_k$ for all s, we have*

$$\sum_{k=1}^K \|q^{\mathbb{P},\pi_k} - q^{\mathbb{P}_k,\pi_k}\|_1 \le \mathcal{O}\left(dHS\sqrt{K\log^2(K/\zeta)}\right).$$

Remark 8. Compared with the $\widetilde{\mathcal{O}}(dS^2\sqrt{K})$ occupancy measure difference in Zhao et al. (2023, Lemma 2), our result is better since $H \le S$ by the layered structure of MDPs.

Combining Lemma 6 with Theorem 1, we obtain the following dynamic regret bound for the linear mixture MDPs.

Theorem 6. *Set parameters as in Theorem 1, Algorithm 1 with Algorithm 4 as subroutine ensures with probability at least $1-\zeta$, the dynamic regret* D-Reg$_K$ *is upper bounded by*

$$\mathcal{O}\Big(dHS\sqrt{K\log^2(K/\zeta)} + \sqrt{T(H\log(S^2A) + \bar{P}_K\log T)}\Big).$$

Algorithm 4: EstimateTransition (Linear Mixture MDPs)

Input: Episode index k, trajectory U_k.
Output: Confidence set $\mathcal{P}_k$.
1: **for** $h = 1$ to H **do**
2: Set $s'_{k,h+1}$ as in (10), $M_{k,h}$ and $b_{k,h}$ as in (11).
3: Compute $\theta_{k,h} = M_{k,h}^{-1} b_{k,h}$.
4: **end for**
5: Set the confidence set as (12).

Remark 9. Compared with the result in Theorem 2, we replace the $\widetilde{\mathcal{O}}(HS\sqrt{AK})$ in the first term with $\widetilde{\mathcal{O}}(dHS\sqrt{K})$, which enjoys a better guarantee when $d \le \sqrt{A}$. Our result is independent of A though still has an undesirable dependence on S. Similar dependence on S also appears in the study of static regret of linear mixture MDPs (Zhao et al. 2023).

We finally establish the lower bound in the theorem below.

Theorem 7. *For any online algorithm and any $\gamma \in [0, 2T]$, there exists an episodic loop-free linear mixture MDP with K episodes, H layers, S states and A actions and a policy sequence $\pi_1^c, \ldots, \pi_K^c$ such that $\bar{P}_K(\pi_1^c, \ldots, \pi_K^c) \le \gamma$ and*

$$\text{D-Reg}_K \ge \Omega\big(dH\sqrt{K} + \sqrt{T(H+\gamma)}\big)$$

under the full-information and unknown transition setting.

Remark 10. Combining Theorem 6 and Theorem 7, we claim that our dynamic regret bound is optimal in d, H, K, $\bar{P}_K$, yet looses a factor of $\mathcal{O}(S)$. The reason is the same as that for linear MDPs, since though the transition $\mathbb{P}$ admits a linear structure, it is not the case for the occupancy measure $q^{\mathbb{P},\pi_k}$, which has a complicated recursive form. Eliminating the dependence on S is an important future direction.

Conclusion and Future Work

In this work, we investigate the dynamic regret of adversarial MDPs with the unknown transition and unknown non-stationarity of the environments. We propose a general framework to decouple the two sources of uncertainties and show the dynamic regret bound naturally decomposes into two terms, one due to constructing confidence sets to handle the unknown transition and the other due to choosing sub-optimal policies under the unknown non-stationarity. To this end, we first employ the two-layer ensemble structure to handle the adaptation error due to the unknown non-stationarity. Subsequently, we instantiate the framework to three fundamental MDP models and present corresponding approaches to control the exploration error due to the unknown transition. We provide dynamic regret guarantees respectively and show they are optimal in terms of the number of episodes K and the non-stationarity measure $\bar{P}_K$ by establishing matching lower bounds. To the best of our knowledge, this is the first work that achieves the dynamic regret with optimal dependence on K and $\bar{P}_K$ for this problem. There are several important future works to study. First, the dependence on S in our dynamic regret bound is not optimal. How to close the gap is an open problem. Second, how to extend our results to the more challenging bandit feedback setting is an important and interesting future direction.

Acknowledgements

This research was supported by National Key R&D Program of China (2022ZD0114800) and NSFC (62361146852, 61921006, 62206125). Peng Zhao was supported in part by the Xiaomi Foundation.

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems 24 (NIPS)*, 2312–2320.

Altman, E. 1998. *Constrained Markov decision processes*. Routledge.

Auer, P.; Cesa-Bianchi, N.; and Gentile, C. 2002. Adaptive and Self-Confident On-Line Learning Algorithms. *Journal of Computer and System Sciences*, 64(1): 48–75.

Auer, P.; Gajane, P.; and Ortner, R. 2019. Adaptively Tracking the Best Bandit Arm with an Unknown Number of Distribution Changes. In *Proceedings of the 32nd Conference on Learning Theory (COLT)*, 138–158.

Ayoub, A.; Jia, Z.; Szepesvári, C.; Wang, M.; and Yang, L. 2020. Model-Based Reinforcement Learning with Value-Targeted Regression. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 463–474.

Azar, M. G.; Osband, I.; and Munos, R. 2017. Minimax Regret Bounds for Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 263–272.

Baby, D.; and Wang, Y. 2021. Optimal Dynamic Regret in Exp-Concave Online Learning. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, 359–409.

Bai, Y.; Zhang, Y.-J.; Zhao, P.; Sugiyama, M.; and Zhou, Z.-H. 2022. Adapting to Online Label Shift with Provable Guarantees. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 29960–29974.

Besbes, O.; Gur, Y.; and Zeevi, A. J. 2015. Non-Stationary Stochastic Optimization. *Operations Research*, 1227–1244.

Burnetas, A. N.; and Katehakis, M. N. 1997. Optimal Adaptive Policies for Markov Decision Processes. *Mathematics of Operations Research*, 22(1): 222–255.

Cai, Q.; Yang, Z.; Jin, C.; and Wang, Z. 2020. Provably Efficient Exploration in Policy Optimization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 1283–1294.

Cesa-Bianchi, N.; and Lugosi, G. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

Chen, L.; Luo, H.; and Wei, C.-Y. 2021. Minimax Regret for Stochastic Shortest Path with Adversarial Costs and Known Transition. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, 1180–1215.

Chen, X.; Wang, Y.; and Wang, Y.-X. 2019. Non-stationary Stochastic Optimization under $L_{p,q}$-Variation Measures. *Operations Research*, 67(6): 1752–1765.

Chen, Y.; Lee, C.-W.; Luo, H.; and Wei, C.-Y. 2019. A New Algorithm for Non-stationary Contextual Bandits: Efficient, Optimal and Parameter-free. In *Proceedings of the 32nd Conference on Learning Theory (COLT)*, 696–726.

Cheung, W. C.; Simchi-Levi, D.; and Zhu, R. 2020. Reinforcement Learning for Non-Stationary Markov Decision Processes: The Blessing of (More) Optimism. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 1843–1854.

Dai, Y.; Luo, H.; Wei, C.; and Zimmert, J. 2023. Refined Regret for Adversarial MDPs with Linear Function Approximation. *ArXiv preprint*, arXiv:2301.12942.

Even-Dar, E.; Kakade, S. M.; and Mansour, Y. 2009. Online Markov Decision Processes. *Mathematics of Operations Research*, 726–736.

Fei, Y.; Yang, Z.; Wang, Z.; and Xie, Q. 2020. Dynamic Regret of Policy Optimization in Non-Stationary Environments. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 6743–6754.

Freund, Y.; and Schapire, R. E. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1): 119–139.

Gajane, P.; Ortner, R.; and Auer, P. 2018. A Sliding-Window Algorithm for Markov Decision Processes with Arbitrarily Changing Rewards and Transitions. *ArXiv preprint*, arXiv:1805.10066.

He, J.; Zhao, H.; Zhou, D.; and Gu, Q. 2023. Nearly Minimax Optimal Reinforcement Learning for Linear Markov Decision Processes. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 12790–12822.

He, J.; Zhou, D.; and Gu, Q. 2022. Near-optimal Policy Optimization Algorithms for Learning Adversarial Linear Mixture MDPs. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 4259–4280.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 1563–1600.

Ji, K.; Zhao, Q.; He, J.; Zhang, W.; and Gu, Q. 2023. Horizon-free Reinforcement Learning in Adversarial Linear Mixture MDPs. *ArXiv preprint*, arXiv:2305.08359.

Jin, C.; Allen-Zhu, Z.; Bubeck, S.; and Jordan, M. I. 2018. Is Q-Learning Provably Efficient? In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 4868–4878.

Jin, C.; Jin, T.; Luo, H.; Sra, S.; and Yu, T. 2020a. Learning Adversarial Markov Decision Processes with Bandit Feedback and Unknown Transition. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 4860–4869.

Jin, C.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2020b. Provably Efficient Reinforcement Learning with Linear Function Approximation. In *Proceedings of the 33rd Conference on Learning Theory (COLT)*, 2137–2143.

Kong, F.; Zhang, X.; Wang, B.; and Li, S. 2023. Improved Regret Bounds for Linear Adversarial MDPs via Linear Optimization. *ArXiv preprint*, arXiv:2302.06834.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Li, L.-F.; Zhao, P.; and Zhou, Z.-H. 2023. Dynamic Regret of Adversarial Linear Mixture MDPs. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, to appear.

Luo, H.; Wei, C.; and Lee, C. 2021. Policy Optimization in Adversarial MDPs: Improved Exploration via Dilated Bonuses. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 22931–22942.

Luo, H.; Zhang, M.; Zhao, P.; and Zhou, Z.-H. 2022. Corralling a Larger Band of Bandits: A Case Study on Switching Regret for Linear Bandits. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, 3635–3684.

Neu, G.; György, A.; and Szepesvári, C. 2010. The Online Loop-free Stochastic Shortest-Path Problem. In *Proceedings of 23rd Conference on Learning Theory (COLT)*, 231–243.

Neu, G.; György, A.; Szepesvári, C.; and Antos, A. 2010. Online Markov Decision Processes under Bandit Feedback. In *Advances in Neural Information Processing Systems 24 (NIPS)*, 1804–1812.

Neu, G.; and Olkhovskaya, J. 2021. Online Learning in MDPs with Linear Function Approximation and Bandit Feedback. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 10407–10417.

Orabona, F. 2019. A Modern Introduction to Online Learning. *ArXiv preprint*, 1912.13213.

Ortner, R.; Gajane, P.; and Auer, P. 2019. Variational Regret Bounds for Reinforcement Learning. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, 81–90.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley.

Qian, Y.-Y.; Bai, Y.; Zhang, Z.-Y.; Zhao, P.; and Zhou, Z.-H. 2023. Handling New Class in Online Label Shift. In *Proceedings of the 23rd IEEE International Conference on Data Mining (ICDM)*, to appear.

Qin, T.; Li, L.-F.; Wang, T.-Z.; and Zhou, Z.-H. 2023. Tracking Treatment Effect Heterogeneity in Evolving Environments. *Machine Learning*, in press.

Rosenberg, A.; and Mansour, Y. 2019. Online Convex Optimization in Adversarial Markov Decision Processes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 5478–5486.

Sherman, U.; Cohen, A.; Koren, T.; and Mansour, Y. 2023. Rate-Optimal Policy Optimization for Linear Markov Decision Processes. *ArXiv preprint*, arXiv:2308.14642.

Sherman, U.; Koren, T.; and Mansour, Y. 2023. Improved Regret for Efficient Online Reinforcement Learning with Linear Function Approximation. *ArXiv preprint*, arXiv:2301.13087.

Sutton, R. S.; and Barto, A. G. 1998. *Reinforcement learning: An introduction*. MIT Press.

Wang, J.; Zhao, P.; and Zhou, Z.-H. 2023. Revisiting Weighted Strategy for Non-stationary Parametric Bandits. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 7913–7942.

Wei, C.-Y.; and Luo, H. 2021. Non-stationary Reinforcement Learning without Prior Knowledge: An Optimal Black-box Approach. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, 4300–4354.

Yan, Y.-H.; Zhao, P.; and Zhou, Z.-H. 2023. Online Nonstochastic Control with Partial Feedback. *Journal of Machine Learning Research*, 24(273): 1–50.

Yu, J. Y.; Mannor, S.; and Shimkin, N. 2009. Markov Decision Processes with Arbitrary Reward Processes. *Mathematics of Operations Research*, 737–757.

Zhang, L.; Lu, S.; and Zhou, Z.-H. 2018. Adaptive Online Learning in Dynamic Environments. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 1330–1340.

Zhao, C.; Yang, R.; Wang, B.; and Li, S. 2023. Learning Adversarial Linear Mixture Markov Decision Processes with Bandit Feedback and Unknown Transition. In *11th International Conference on Learning Representations (ICLR)*.

Zhao, P.; Li, L.-F.; and Zhou, Z.-H. 2022. Dynamic Regret of Online Markov Decision Processes. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 26865–26894.

Zhao, P.; Wang, Y.-X.; and Zhou, Z.-H. 2022. Non-stationary Online Learning with Memory and Non-stochastic Control. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2101–2133.

Zhao, P.; Zhang, Y.-J.; Zhang, L.; and Zhou, Z.-H. 2020. Dynamic Regret of Convex and Smooth Functions. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 12510–12520.

Zhao, P.; Zhang, Y.-J.; Zhang, L.; and Zhou, Z.-H. 2021. Adaptivity and Non-stationarity: Problem-dependent Dynamic Regret for Online Convex Optimization. *ArXiv preprint*, 2112.14368.

Zhong, H.; Yang, Z.; Wang, Z.; and Szepesvári, C. 2021. Optimistic Policy Optimization is Provably Efficient in Non-stationary MDPs. *ArXiv preprint*, arXiv:2110.08984.

Zhong, H.; and Zhang, T. 2023. A Theoretical Analysis of Optimistic Proximal Policy Optimization in Linear Markov Decision Processes. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, to appear.

Zhou, D.; Gu, Q.; and Szepesvári, C. 2021. Nearly Minimax Optimal Reinforcement Learning for Linear Mixture Markov Decision Processes. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, 4532–4576.

Zhou, Z.-H. 2012. *Ensemble methods: foundations and algorithms*. CRC Press.

Zhou, Z.-H. 2022. Open-environment Machine Learning. *National Science Review*, 9(8).

Zimin, A.; and Neu, G. 2013. Online Learning in Episodic Markovian Decision Processes by Relative Entropy Policy Search. In *Advances in Neural Information Processing Systems 26 (NIPS)*, 1583–1591.

Zinkevich, M. 2003. Online Convex Programming and Generalized Infinitesimal Gradient Ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, 928–936.