Dynamic Regret of Online Markov Decision Processes

Peng Zhao [1] **Long-Fei Li** [1] **Zhi-Hua Zhou** [1]

Abstract

We investigate online Markov Decision Processes (MDPs) with adversarially changing loss functions and known transitions. We choose *dynamic regret* as the performance measure, defined as the performance difference between the learner and any sequence of feasible *changing* policies. The measure is strictly stronger than the standard static regret that benchmarks the learner's performance with a fixed compared policy. We consider three foundational models of online MDPs, including episodic loop-free Stochastic Shortest Path (SSP), episodic SSP, and infinite-horizon MDPs. For the three models, we propose novel online ensemble algorithms and establish their dynamic regret guarantees respectively, in which the results for episodic (loop-free) SSP are provably minimax optimal in terms of time horizon and certain non-stationarity measure.

1. Introduction

Markov Decision Processes (MDPs) are widely used to model decision-making problems, where a learner interacts with the environments sequentially and aims to improve the learned strategy over time. The MDPs model is very general and encompasses a variety of applications, including games (Silver et al., 2016), robotic control (Schulman et al., 2015), autonomous driving (Kendall et al., 2019), etc.

In this paper, we focus on the online MDPs framework with adversarially changing loss functions and known transitions, which has attracted increasing attention in recent years due to its generality (Even-Dar et al., 2009; Zimin & Neu, 2013; Rosenberg & Mansour, 2019a; Jin et al., 2020a; Chen et al., 2021a). Let T be the total time horizon. The general procedures of the online MDPs are as follows: at each round $t \in [T]$, the learner observes the current state x_t and decides a policy $\pi_t : X \times A \to [0, 1]$, where $\pi_t(a|x)$ is the probability of taking action $a \in A$ at state $x \in X$. Then, the learner draws and executes an action a_t from $\pi_t(\cdot|x_t)$ and suffers a loss $\ell_t(x_t, a_t)$. The environments subsequently transit to the next state x_{t+1} according to the transition kernel $P(\cdot|x_t, a_t)$. We focus on the full-information setting where the entire loss function is revealed to the learner. The standard measure for online MDPs is the *regret* defined as the performance difference between learner's policy and that of the best fixed policy in hindsight, namely,

$$\textsc{Reg}_T = \sum_{t=1}^{T} \ell_t\left(x_t, \pi_t(x_t)\right) - \min_{\pi \in \Pi} \sum_{t=1}^{T} \ell_t\left(x_t, \pi(x_t)\right), \quad (1)$$

where Π is a certain policy class. There are many efforts devoted to optimizing the measure, yielding fruitful results (Even-Dar et al., 2009; Neu et al., 2012; Zimin & Neu, 2013; Neu et al., 2014; Rosenberg & Mansour, 2019a; 2021; Chen et al., 2021a). However, one caveat in the performance measure in Eq. (1) is that the measure only benchmarks the learner's performance with a *fixed* strategy, so it is usually called the *static regret* in the literature. The fact makes the static regret metric not suitable to guide the algorithm design for online decision making in non-stationary environments, which is often the case in many real-world decision-making applications such as online recommendations and autonomous driving. In particular, in online MDPs model the loss functions encountered by the learner can be adversarially changing, it is thus unrealistic to assume the existence of a single fixed strategy that can perform well over the horizon. To this end, in this paper we introduce the *dynamic regret* as the metric to guide the algorithm design for online MDPs, which competes the learner's performance against a sequence of changing policies, defined as

$$\text{D-}\textsc{Reg}_T(\pi^{\mathsf{c}}_{1:T}) = \sum_{t=1}^{T} \ell_t(x_t, \pi_t(x_t)) - \sum_{t=1}^{T} \ell_t(x_t, \pi^{\mathsf{c}}_t(x_t)), \quad (2)$$

where $\pi^{\mathsf{c}}_1, \ldots, \pi^{\mathsf{c}}_T$ is any sequence of compared policies in the policy class Π, which can be chosen with the complete foreknowledge of the online loss functions. We use $\pi^{\mathsf{c}}_{1:T}$ as a shorthand of the compared policies. An upper bound of dynamic regret usually scales with a certain variation quantity of the compared policies denoted by $P_T(\pi^{\mathsf{c}}_1, \ldots, \pi^{\mathsf{c}}_T)$ that can reflect the non-stationarity of environments.

[1] National Key Laboratory for Novel Software Technology, Nanjing University. Correspondence to: Zhi-Hua Zhou <zhouzh@lamda.nju.edu.cn>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

Table 1: Summary of our main results. For three models of online MDPs (episodic loop-free SSP, episodic SSP, and infinite-horizon MDPs), we establish their dynamic regret guarantees. Our obtained dynamic regret bounds immediately recover the best known static regret presented in the last column, when choosing a fixed compared policy and the non-stationarity measure P_T or $\bar{P}_K$ then equals to zero. Note that all our results are achieved by *parameter-free* algorithms in the sense that they do not require the knowledge of unknown quantities related to the environmental non-stationarity.

MDP Model	Ours Result (dynamic regret)	Previous Work (static regret)
Episodic loop-free SSP (Section 2)	$\widetilde{\mathcal{O}}(H\sqrt{K(1+P_T)})$ [Theorem 1]	$\widetilde{\mathcal{O}}(H\sqrt{K})$ (Zimin & Neu, 2013)
Episodic SSP (Section 3)	$\widetilde{\mathcal{O}}(\sqrt{B_K(H_* + \bar{P}_K)} + \bar{P}_K)$ [Theorem 3]	$\widetilde{\mathcal{O}}(\sqrt{H^{\pi^*}DK})$ (Chen et al., 2021a)
Infinite-horizon MDPs (Section 4)	$\widetilde{\mathcal{O}}(\sqrt{\tau T(1+\tau P_T)} + \tau^2 P_T)$ [Theorem 6]	$\widetilde{\mathcal{O}}(\sqrt{\tau T})$ (Zimin & Neu, 2013)

The dynamic regret measure in Eq. (2) is in fact very general due to the flexibility of compared policies. For example, it immediately recovers the standard regret notion defined in Eq. (1) when choosing the single best compared policy in hindsight, namely, choosing $\pi^{\mathsf{c}}_{1:T} = \pi^* \in \arg\min_{\pi\in\Pi} \sum_{t=1}^{T} \ell_t(x_t, \pi(x_t))$. Hence, any dynamic regret upper bound directly implies a static regret upper bound by substituting a fixed compared policy. Another typical choice is setting the compared policies as the sequence of the best policy of each round, namely, choosing $\pi^{\mathsf{c}}_t = \pi^*_t \in \arg\min_{\pi\in\Pi} \ell_t(x_t, \pi(x_t))$, and the resulting dynamic regret measure is sometimes referred to as the *worst-case* dynamic regret in the literature (Zhang et al., 2018). It is noteworthy to emphasize that the dynamic regret measure in Eq. (2) does not assume prior information of the compared policies, which is certainly also unknown to the online algorithms. As a result, the measure is also called *universal* dynamic regret (or *general* dynamic regret) in the sense that the regret bound holds for any feasible compared policies. Both static regret and the worst-case dynamic regret are two special cases of the universal dynamic regret by configuring different choices of compared policies.

In this paper, focusing on the dynamic regret measure presented in Eq. (2), we investigate three foundational and well-studied models of online MDPs: (i) episodic loop-free Stochastic Shortest Path (SSP) (Zimin & Neu, 2013), (ii) episodic SSP (Rosenberg & Mansour, 2021; Chen et al., 2021a), and (iii) infinite-horizon MDPs (Even-Dar et al., 2009). The first two SSP models belong to episodic MDPs, in which the learner interacts with environments in episodes and the goal is to reach a goal state with minimum total loss. The distinction lies in that the learner is guaranteed to reach the goal state within a fixed number of steps in the loop-free SSP model; by contrast, the horizon length in general SSP model depends on the learner's policies, which could potentially be infinite (if the goal is not reached). In infinite-horizon MDPs, there is no goal state and the horizon can be never end and the goal of the learner is to minimize the average loss over time. For all those three models, we propose novel online algorithms and provide the corresponding expected dynamic regret guarantees. We also establish several lower bound results and show that the obtained upper bounds for episodic loop-free SSP and general SSP are *minimax optimal* in terms of time horizon and non-stationarity measure. Notably, all our algorithms are *parameter-free* in the sense that they do not require knowing the non-stationarity quantity ahead of time. Table 1 summarizes our main results.

Technical contributions. Similar to prior studies of non-stationary online learning (Hazan & Seshadhri, 2009; Daniely et al., 2015; Zhang et al., 2018; Zhao et al., 2020b), our proposed algorithms fall into the online ensemble framework with a meta-base two-layer structure. While the framework is standard in modern online learning, several important new ingredients are required to achieve minimax dynamic regret guarantees for online MDPs. We highlight the main technical challenges and contributions as follows.

- For all three models, algorithms are performed over the "occupancy measure" space, so dynamic regret inevitably scales with the variation of occupancy measures induced by compared policies, making it necessary to establish relationships between the variation of occupancy measures and that of compared policies.
- Achieving the minimax dynamic regret bound for episodic (non-loop-free) SSP is one of the most challenging parts of this paper due to the complicated structure of this model and also the requirement of handling dual uncertainties of unknown horizon length and unknown non-stationarity. This motivates a novel groupwise scheduling for base-learners and a new weighted entropy regularizer for the meta-algorithm. Additionally, appropriate correction terms in the feedback loss and carefully designed step sizes for both base-algorithm and meta-algorithm are also important.
- For learning in infinite-horizon MDPs, we present a reduction to the problem of minimizing dynamic regret of the switching-cost expert problem, which is new to the best of our knowledge.

Notations. We present several general notations used throughout the paper. We use $\ell \in \mathbb{R}^d_{[a,b]}$ to denote a vector whose each element satisfies $\ell_i \in [a,b]$ for $i \in [d]$. For a vector $a \in \mathbb{R}^d$, a^2 denotes the vector $(a_1^2, \ldots, a_d^2)^\top \in \mathbb{R}^d$. Besides, $e_i \in \mathbb{R}^d$ denotes the i-th standard basis vector. For a convex function ψ, its induced Bregman divergence is defined as $D_\psi(u,w) = \psi(u) - \psi(w) - \langle \nabla\psi(w), u - w\rangle$. Given two policies π and π', $\|\pi - \pi'\|_{1,\infty} = \max_x \|\pi(\cdot|x) - \pi'(\cdot|x)\|_1$. $\widetilde{\mathcal{O}}(\cdot)$ omits the logarithmic factors on horizon T.

Organization. The rest of the paper is organized as follows. In Section 2 and Section 3, we establish the minimax dynamic regret for episodic loop-free and general SSP respectively. In Section 4, we provide dynamic regret bound for infinite-horizon MDPs. Section 5 concludes the paper and discusses the future work. We defer the related works to Appendix A and proofs to remaining appendices.

2. Episodic Loop-free SSP

This section presents our results for episodic loop-free SSP, a foundational and conceptually simple model of online MDPs. We first introduce the problem setup, and then establish the minimax dynamic regret bound.

2.1. Problem Setup

An episodic online MDP is specified by a tuple $M = (X, g, A, P, \{\ell_k\}_{k=1}^K)$, where X and A are the finite state and action spaces, $g \notin X$ is the goal state, $P : X \times A \times X \cup \{g\} \to [0,1]$ is the transition kernel, K is the number of episodes and $\ell_k \in \mathbb{R}^{|X||A|}_{[0,1]}$ is the loss function in episode $k \in [K]$. An episodic loop-free SSP is an instance of episodic online MDP and further satisfies the following conditions: state space $X \cup \{g\}$ can be decomposed into $H+1$ non-intersecting layers denoted by $X_0, \ldots, X_{H-1}, g$ such that $X_0 = \{x_0\}$ and g are singletons, and transitions are only possible between the consecutive layers. Notice that the total horizon is $T = KH$.

The learning protocol of episodic loop-free SSP proceeds in K episodes. In each episode $k \in [K]$, environments decide a loss $\ell_k : X \times A \to [0,1]$, and simultaneously the learner starts from state x_0 and moves forward across consecutive layers until reaching the goal state g. We focus on the full-information setting, i.e., the loss is revealed to the learner after the episode ends. Notably, no statistical assumption is imposed on the loss sequence, which means the online loss functions can be chosen in an adversarial manner.

Occupancy measure. Existing studies reveal the importance of the concept "occupancy measure" in handling online MDPs (Zimin & Neu, 2013; Rosenberg & Mansour, 2019a), which deeply connects the problem of online MDPs with online convex optimization. Given a policy π and transition kernel P, the occupancy measure $q^\pi \in \mathbb{R}^{|X||A|}_{[0,1]}$ is defined as the probability of visiting state-action pair (x,a) by executing the policy π, i.e., $q^\pi(x,a) = \Pr\left[x_{l(x)} = x, a_{l(x)} = a | P, \pi\right]$, where $l(x)$ is the index of the layer that x belongs to. For an episode loop-free SSP instance M, its occupancy measure space is defined as $\Delta(M) = \{q \in \mathbb{R}^{|X||A|}_{[0,1]} \mid \sum_{x \in X_l} \sum_{a \in A} q(x,a) = 1, \forall l = \{0\} \cup [H-1]$ and $\sum_{x' \in X_{l(x)-1}} \sum_{a' \in A} P(x|x',a') q(x',a') = \sum_{a \in A} q(x,a), \forall x \in X \setminus \{x_0\}\}$, For any occupancy measure $q \in \Delta(M)$, it induces a policy π such that $\pi(a|x) \propto q(x,a)$, $\forall x \in X, a \in A$. Existing study shows that there exists a unique induced policy for all occupancy measures in $\Delta(M)$ and vice versa (Zimin & Neu, 2013). Then, the expected loss of any policy π at episode k can be written as $\mathbb{E}\left[\sum_{l=0}^{H-1} \ell_k(x_l, a_l \mid P, \pi)\right] = \sum_{l=0}^{H-1} \sum_{x \in X_l} \sum_{a \in A} q^\pi(x,a) \ell_t(x,a) = \langle q^\pi, \ell_k \rangle$, where the expectation is taken over the randomness of the policy and transition kernel. Note the total horizon T of episodic loop-free SSP can be divided into K episodes, each with horizon length H, i.e., $T = KH$. Denote by $\pi_{k,l}$ the policy at layer $l \in \{0\} \cup [H-1]$ in episode $k \in [K]$, the policy sequence $\pi_1, \ldots, \pi_T$ in Eq. (1) can be represented by $\pi_{1,0}, \ldots, \pi_{1,H-1}, \pi_{2,0}, \ldots, \pi_{K,H-1}$. We use the notation π_k as a shorthand of $\pi_{k,0:H-1}$ for notational simplicity. Then we can rewrite the expected static regret in Eq. (1) as $\mathbb{E}\left[\text{REG}_K\right] = \sum_{k=1}^K \langle q^{\pi_k}, \ell_k \rangle - \min_{q \in \Delta(M)} \sum_{k=1}^K \langle q, \ell_k \rangle$.

Dynamic regret. Similar to the derivation for static regret, we can also rewrite the expected dynamic regret in Eq. (2) into a form with respect to the occupancy measure as

$$\mathbb{E}\left[\text{D-REG}_K(\pi^{\mathsf{c}}_{1:K})\right] = \sum_{k=1}^K \langle q^{\pi_k}, \ell_k \rangle - \sum_{k=1}^K \langle q^{\pi^{\mathsf{c}}_k}, \ell_k \rangle, \quad (3)$$

where $q^{\pi^{\mathsf{c}}_k}$ is the occupancy measure of the compared policy π^{c}_k for all $k \in [K]$. The non-stationarity measure is naturally defined as $P_T = \sum_{k=2}^K \sum_{l=0}^{H-1} \|\pi^{\mathsf{c}}_{k,l} - \pi^{\mathsf{c}}_{k-1,l}\|_{1,\infty}$.

2.2. Dynamic Regret

Before presenting our algorithm for dynamic regret, we first briefly review the O-REPS algorithm of Zimin & Neu (2013) developed for minimizing the static regret. The key idea of O-REPS is to perform online mirror descent over the occupancy measure space $\Delta(M)$, specifically, at episode $k+1$, the learner updates the prediction by

$$q_{k+1} = \mathop{\arg\min}_{q \in \Delta(M)} \eta \langle q, \ell_k \rangle + D_\psi(q, q_k),$$

where $\eta > 0$ is step size, $\psi(q) = \sum_{x,a} q(x,a) \log q(x,a)$ is the standard negative entropy. Zimin & Neu (2013) prove O-REPS enjoys an $\mathcal{O}(H\sqrt{K \log(|X||A|)})$ static regret.

By slightly modifying the algorithm, in following lemma we show O-REPS over a clipped occupancy measure space can achieve dynamic regret guarantees. Specifically, define the clipped space as $\Delta(M,\alpha) = \{q \mid q \in \Delta(M), \text{ and } q(x,a) \geq \alpha, \forall x, a\}$ with $0 < \alpha < 1$ being the clipping parameter, we prove that performing O-REPS over $\Delta(M,\alpha)$ ensures the following dynamic regret.

Lemma 1. *Set $q_1 = \arg\min_{q \in \Delta(M,\alpha)} \psi(q)$. For any compared policies $\pi^{\mathsf{c}}_1, \ldots, \pi^{\mathsf{c}}_K \in \{\pi \mid q^\pi \in \Delta(M,\alpha)\}$, O-REPS over a clipped space $\Delta(M,\alpha)$ ensures*

$$\sum_{k=1}^{K} \langle q_k - q^{\pi^{\mathsf{c}}_k}, \ell_k \rangle \leq \eta T + \frac{1}{\eta}\left(H \log \frac{|X||A|}{H} + \bar{P}_T \log \frac{1}{\alpha}\right),$$

where $\bar{P}_T = \bar{P}_T(\pi^{\mathsf{c}}_1, \ldots, \pi^{\mathsf{c}}_K) = \sum_{k=2}^{K} \|q^{\pi^{\mathsf{c}}_k} - q^{\pi^{\mathsf{c}}_{k-1}}\|_1$ is the path-length of occupancy measures.

To achieve a favorable dynamic regret, we need to set the step size η optimally to balance time horizon T and the path-length of occupancy measures $\bar{P}_T$. However, we actually do not have prior knowledge of $\bar{P}_T$ even after the horizon ends since the compared policies can be arbitrarily chosen in the feasible set. Thus, we cannot apply the standard adaptive step size tuning techniques such as doubling trick or self-confident tuning (Auer et al., 2002) to remove the dependence on $\bar{P}_T$. To address the issue, we employ a meta-base two-layer structure to handle the uncertainty (Zhang et al., 2018; Zhao et al., 2020b). Specifically, we first construct a step size pool $\mathcal{H} = \{\eta_1, \cdots, \eta_N\}$ (N is the number of candidate step sizes and is of order $\mathcal{O}(\log T)$ whose configuration will be specified later) to discretize value range of the optimal step size; and then initialize multiple base-learners simultaneously, denoted by $\mathcal{B}_1, \cdots, \mathcal{B}_N$, where $\mathcal{B}_i$ returns her prediction $q_{k,i}$ by performing O-REPS with step size $\eta_i \in \mathcal{H}$; finally a meta-algorithm is used to combine predictions of all base-learners and yield the final output $\{q_k\}_{k=1}^K$. Below, we specify the details.

At episode $k \in [K]$, the learner receives the decision $q_{k,i}$ from each base-learner $\mathcal{B}_i, \forall i \in [N]$ and the weight vector $p_k \in \Delta_N$ from meta-algorithm. Then the learner outputs the decisions by $q_k = \sum_{i=1}^N p_{k,i} q_{k,i}$, plays the policy $\pi(a|x) \propto q(x,a), \forall x, a$, and observes the loss function ℓ_k. After that, the base-learner $\mathcal{B}_i$ updates by performing O-REPS over the clipped space $\Delta(M,\alpha)$ with step size $\eta_i \in \mathcal{H}$, namely,

$$q_{k+1,i} = \arg\min_{q \in \Delta(M,\alpha)} \eta_i \langle q, \ell_k \rangle + D_\psi(q, q_{k,i}),$$

where $\eta_i \in \mathcal{H}$ is the step size of the base-learner $\mathcal{B}_i$. The meta-algorithm aims to track the unknown best base-learner. We employ Hedge algorithm (Freund & Schapire, 1997) that updates the weight $p_{k+1} \in \Delta_N$ by $p_{k+1,i} \propto \exp(-\varepsilon \sum_{s=1}^k h_{s,i})$ where $\varepsilon > 0$ is the learning rate, $h_k \in \mathbb{R}^N$ evaluates the performance of the base-learners and is set as $h_{k,i} = \langle q_{k,i}, \ell_k \rangle$ for $i \in [N]$.

Algorithm 1 DO-REPS

```
Input: step size pool $\mathcal{H}$, learning rate $\varepsilon$, clipping param $\alpha$.
Define $\psi(q) = \sum_{x,a} q(x,a) \log q(x,a)$.
Initialization: set $q_{1,i} = \arg\min_{q \in \Delta(M,\alpha)} \psi(q)$ and $p_{1,i} = 1/N, \forall i \in [N]$.
for $k = 1$ to $K$ do
    Receive $q_{k,i}$ from base-learner $\mathcal{B}_i$ for $i \in [N]$.
    Compute occupancy measure $q_k = \sum_{i=1}^N p_{k,i} q_{k,i}$.
    Play the induced policy $\pi_k(a|x) \propto q(x,a), \forall x, a$.
    Update the weight by $p_{k+1,i} \propto \exp(-\varepsilon \sum_{s=1}^k h_{s,i})$
        where $h_{k,i} = \langle q_{k,i}, \ell_k \rangle, \forall i \in [N]$.
    Each base-learner $\mathcal{B}_i$ updates prediction by
        $q_{k+1,i} = \arg\min_{q \in \Delta(M,\alpha)} \eta_i \langle q, \ell_k \rangle + D_\psi(q, q_{k,i})$.
end for
```

Algorithm 1 summarizes our proposed Dynamic O-REPS (DO-REPS) algorithm and the guarantee is as follows.

Theorem 1. *Set the clipping parameter $\alpha = 1/T^2$, the step size pool $\mathcal{H} = \{\eta_i = 2^{i-1}\sqrt{K^{-1}\log(|X||A|/H)} \mid i \in [N]\}$, where $N = \lceil \frac{1}{2}\log(1 + \frac{4K \log T}{\log(|X||A|/H)}) \rceil + 1$, and the learning rate of meta-algorithm as $\varepsilon = \sqrt{(\log N)/(HT)}$. DO-REPS (Algorithm 1) satisfies*

$$\begin{aligned} \mathbb{E}[\text{D-Reg}_K(\pi^{\mathsf{c}}_{1:K})] &\leq \mathcal{O}\big(\sqrt{T(H \log |X||A| + \bar{P}_T \log T)}\big) \\ &\leq \mathcal{O}\big(H\sqrt{K(\log |X||A| + P_T \log T)}\big), \end{aligned}$$

where $\bar{P}_T = \sum_{k=2}^K \|q^{\pi^{\mathsf{c}}_k} - q^{\pi^{\mathsf{c}}_{k-1}}\|_1$ is the path-length of occupancy measures and $P_T = \sum_{k=2}^K \sum_{l=0}^{H-1} \|\pi^{\mathsf{c}}_{k,l} - \pi^{\mathsf{c}}_{k-1,l}\|_{1,\infty}$ is the path-length of the compared policies.

Remark 1. *Setting compared policies $\pi^{\mathsf{c}}_{1:K} = \pi^*$ (then $P_T = 0$), Theorem 1 recovers the $\mathcal{O}(H\sqrt{K \log |X||A|})$ minimax optimal static regret of Zimin & Neu (2013).*

The proof can be found in Appendix C.3. Note that Theorem 1 presents two dynamic regret bounds in terms of either the path-length of occupancy measures $\bar{P}_T$ or the path-length of compared policies P_T (see definition at the end of Section 2.1). To achieve the latter one, we establish the relationship of path-length variations between compared policies and their induced occupancy measures. Indeed, we prove that $\bar{P}_T \leq HP_T$ in Lemma 6 of Appendix C.1.

We finally establish the lower bound in Theorem 2, which indicates the minimax optimality of our attained upper bound in terms of T and $\bar{P}_T$ (up to logarithmic factors).

Theorem 2. *For any online algorithm and any $\gamma \in [0, 2T]$, there exists an episode loop-free SSP instance with H layers, $|X|$ states and $|A|$ actions and a sequence of compared policies $\pi^{\mathsf{c}}_1, \ldots, \pi^{\mathsf{c}}_K$ such that*

$$\bar{P}_T \leq \gamma \text{ and } \mathbb{E}[\text{D-Reg}_K] \geq \Omega(\sqrt{T(H+\gamma)\log |X||A|})$$

under the full-information and known transition setting.

3. Episodic SSP

In this section, we consider the episodic SSP, which does not necessarily satisfy the loop-free structure and is thus more general and difficult than the loop-free SSP studied in Section 2. For this model, we first introduce the formal problem setup and then establish minimax dynamic regret.

3.1. Problem Setup

An episodic SSP instance is defined by a tuple $M = (X, g, A, P, \{\ell_k\}_{k=1}^K)$, as the same as introduced in Section 2.1, $x_0 \in X$ is the initial state and $g \notin X$ is the goal state. The learning protocol proceeds in K episodes. In each episode $k \in [K]$, environments decide a loss $\ell_k : X \times A \to [0, 1]$, and simultaneously the learner starts from state x_0 and moves to the next state until reaching the goal state g. Thus, the horizon in each episode depends on the learner's policy and is unfixed and can be even infinite, leading to inherent difficulties compared with episodic loop-free SSP. The goal of the learner is to reach the goal with the smallest cumulative loss. Again, we focus on the full-information setting, namely, the entire loss is revealed to the learner after the episode ends. Below we introduce several key concepts and we refer the reader to the recent work (Chen et al., 2021a) for more details.

Proper policy. A policy is called *proper* if playing it ensures that the goal state is reached within a finite number of steps with probability 1 starting from any state, otherwise it is called *improper*. The set of all proper policies is denoted by Π_{proper}. Following earlier studies (Rosenberg & Mansour, 2021; Chen et al., 2021a), we assume $\Pi_{\text{proper}} \neq \emptyset$.

Hitting time. Denote by $H^\pi(x)$ the expected hitting time of g when executing policy π and starting from state x. If π is proper, $H^\pi(x)$ is finite for any $x \in X$. Let $H^\pi \triangleq H^\pi(x_0)$ be the hitting time of policy π from the initial state x_0 to simplify notation. Another useful concept in SSP is the *fast policy* π^f, defined as the (deterministic) policy that achieves the minimum expected hitting time starting from any state. The diameter of the SSP is defined as $D = \max_{x \in X} \min_{\pi \in \Pi_{\text{proper}}} H^\pi(x) = \max_{x \in X} H^{\pi^f}(x)$. Note both π^f and D can be computed ahead of time as the transition kernel is known (Bertsekas & Tsitsiklis, 1991).

Cost-to-go function. Given a loss function ℓ and a policy π, the induced *cost-to-go function* $J^\pi : X \to [0, \infty)$ is defined as $J^\pi(x) = \mathbb{E}[\sum_{i=1}^I \ell(x_i, a_i) \big| P, \pi]$, where I denotes the number of steps before reaching g of policy π and the expectation is over the randomness of the stochastic policy and transition kernel. Denote by J_k^π the cost-to-go function for policy π with respect to loss ℓ_k from the initial state x_0.

Occupancy measure. For the episodic SSP, the occupancy measure $q^\pi \in \mathbb{R}^{|X||A|}$ is defined as the expected number of visits to (x, a) from x_0 to g when executing π, i.e., $q^\pi(s, a) = \mathbb{E}[\sum_{t=1}^I \mathbb{1}\{x_t = x, a_t = a\} \mid P, \pi, x_1 = x_0]$. Similar to the case in loop-free SSP, the inducted policy of a given occupancy measure $q : X \times A \to [0, \infty)$ can be calculated by $\pi(a|x) \propto q(x, a), \forall x, a$. It holds that $H^\pi = \sum_{x,a} q^\pi(x, a)$. Based on the occupancy measure, we can rewrite the cost-to-go function J_k^π as $J_k^\pi = \mathbb{E}[\sum_{i=1}^{I_k} \ell_k(x_i, a_i) \mid P, \pi_k] = \sum_{x,a} q^\pi(x, a)\ell_k(x, a) = \langle q^\pi, \ell_k \rangle$, where I_k denotes the number of steps before reaching g of policy π in episode k. Then the expected static regret in Eq. (1) for episodic SSP can be written as $\mathbb{E}\left[\textsc{Reg}_K\right] = \mathbb{E}\big[\sum_{k=1}^K (J_k^{\pi_k} - J_k^{\pi^*})\big] = \mathbb{E}\big[\sum_{k=1}^K \langle q^{\pi_k} - q^{\pi^*}, \ell_k \rangle\big]$, where $\pi^* = \arg\min_{\pi \in \Pi_{\text{proper}}} \sum_{k=1}^K J_k^\pi$. Two important quantities related to π^* are commonly used in the analysis: (i) its hitting time H^{π^*} from initial state x_0; and (ii) the cumulative loss $\sum_{k=1}^K J_k^{\pi^*}$ during K episodes. The cumulative loss of the best policy is smaller than the fast policy, i.e., $\sum_{k=1}^K J_k^{\pi^*} \leq \sum_{k=1}^K J_k^{\pi^f} \leq DK$, where the last inequality holds due to the definition of the fast policy and the boundedness of the loss range in $[0, 1]$.

Dynamic regret. Similar to the derivation for static regret, we can also rewrite the expected dynamic regret in Eq. (2) into a form with respect to the occupancy measure as

$$\mathbb{E}[\text{D-}\textsc{Reg}_K] \triangleq \mathbb{E}\left[\sum_{k=1}^K (J_k^{\pi_k} - J_k^{\pi_k^{\mathsf{c}}})\right] = \mathbb{E}\left[\sum_{k=1}^K \langle q^{\pi_k} - q^{\pi_k^{\mathsf{c}}}, \ell_k \rangle\right].$$

Similarly, we generalize the two crucial quantities to accommodate changing comparators: the largest hitting time starting from the initial state $H_* = \max_{k \in [K]} H^{\pi_k^{\mathsf{c}}}$ and the cumulative loss of compared policies $B_K = \sum_{k=1}^K J_k^{\pi_k^{\mathsf{c}}} = \sum_{k=1}^K \langle q^{\pi_k^{\mathsf{c}}}, \ell_k \rangle$. It is clear that $B_K \leq H_* K$. Notably, both quantities H_* and B_K are *unknown* to the learner due to involving the unknown compared policies. For the episodic (non-loop-free) SSP, the non-stationarity measure is naturally defined as $P_K = \sum_{k=2}^K \|\pi_k^{\mathsf{c}} - \pi_{k-1}^{\mathsf{c}}\|_{1,\infty}$.

3.2. Dynamic Regret

Before introducing our approach, we first review existing works studying static regret and then show that several crucial ingredients are required to achieve dynamic regret.

To resolve episodic (non-loop-free) SSP, Rosenberg & Mansour (2021) propose to deploy Online Mirror Descent (OMD) over the *parametrized* occupancy measure space. For an MDP instance M and a given horizon length H, the parameterized space is defined as $\Delta(M, H) = \{q \in \mathbb{R}_{\geq 0}^{|X||A|} \mid \sum_{x,a} q(x, a) \leq H \text{ and } \sum_a q(x, a) = \sum_{x',a'} P(x|x', a')q(x', a'), \forall x \in X\}$. The authors prove that OMD enjoys an $\widetilde{\mathcal{O}}(H\sqrt{K})$ static regret as long as $q^{\pi^*} \in \Delta(M, H)$. Therefore, if the largest hitting time H^{π^*} were known ahead of time, a simple choice of $H = H^{\pi^*}$ would attain the favorable static regret. However, such information is in fact unavailable in advance, which motivates a two-layer approach deal with this uncertainty.

Specifically, Chen et al. (2021a) maintain multiple base-learners $\mathcal{B}_1, \ldots, \mathcal{B}_N$, where $\mathcal{B}_i$ works with an occupancy measure space $\Delta(M, H_i)$ and a step size η_i and returns her individual occupancy measure q_k^i; and then a certain meta-algorithm is employed to combine predictions of base-learners to produce final decisions q_k. Let $\mathcal{B}_{i^*}$ be the base-learner whose space size H_{i^*} well approximates the unknown H^{π^*}. Denote by $L_K = \sum_{k=1}^K \langle q_k, \ell_k \rangle, L_K^{i^*} = \sum_{k=1}^K \langle q_k^{i^*}, \ell_k \rangle, L_K^{\mathsf{c}} = \sum_{k=1}^K \langle q^{\pi_k^{\mathsf{c}}}, \ell_k \rangle$ [1] the cumulative loss of final decisions, base-learner $\mathcal{B}_{i^*}$ and the compared policy, respectively. The overall regret can be decomposed as

$$\mathbb{E}[\textsc{Reg}_K] = \mathbb{E}[(L_K - L_K^{i^*})] + \mathbb{E}[(L_K^{i^*} - L_K^{\mathsf{c}})], \quad (4)$$

where the two terms are called *meta-regret* (that captures the regret overhead due to the two-layer ensemble) and *base-regret* (that measures the regret of the unknown best base-learner). To achieve a favorable regret, they propose two mechanisms to control base-regret and meta-regret respectively. First, they pick the base-algorithm with an $\widetilde{\mathcal{O}}(H_{i^*}/\eta_{i^*} + \eta_{i^*} L_K^{\mathsf{c}})$ *small-loss* static regret, which ensures an $\widetilde{\mathcal{O}}(\sqrt{H^{\pi^*} DK})$ base-regret by setting $\eta_i = \mathcal{O}(\sqrt{H_i/DK})$ as the cumulative loss of the best policy in hindsight satisfies $L_K^{\mathsf{c}} \leq DK$. Second, they design a small-loss type *multi-scale* online algorithm (roughly, OMD with weighted entropy $\bar{\psi}(p) = \sum_{i=1}^N \frac{1}{\varepsilon_i} p_i \log p_i$) as the meta-algorithm to make meta-regret adaptive to the individual loss range of experts, so that meta-regret is at most $\widetilde{\mathcal{O}}(1/\varepsilon_{i^*} + \varepsilon_{i^*} H_{i^*} L_K^{i^*})$. Combining the base-regret we further have $L_K^{i^*} \leq L_K^{\mathsf{c}} + \widetilde{\mathcal{O}}(\sqrt{H^{\pi^*} DK}) \leq DK + \widetilde{\mathcal{O}}(\sqrt{H^{\pi^*} DK}) = \widetilde{\mathcal{O}}(DK)$ as $H^{\pi^*} \leq DK$. So an $\widetilde{\mathcal{O}}(\sqrt{H^{\pi^*} DK})$ meta-regret is achievable by setting $\varepsilon_i = \widetilde{\mathcal{O}}(1/\sqrt{H_i DK})$, which in conjunction with the base-regret yields an $\widetilde{\mathcal{O}}(\sqrt{H^{\pi^*} DK})$ static regret.

However, it becomes more involved for dynamic regret. First of all, in addition to the uncertainty of unknown horizon length H_*, the base level also needs to deal with the unknown environmental non-stationarity P_K. Conceptually, this can be handled by maintaining more base-learners, which will be specified later. Second and more importantly, it is challenging to design a compatible meta-algorithm. To see this, suppose we already have an $\widetilde{\mathcal{O}}(\sqrt{B_K(P_K + H_*)})$ *small-loss dynamic regret* for the base-algorithm, where B_K is the cumulative loss of compared policies as defined at the end of Section 3.1, we then continue the above recipe and see the issue in meta-regret. Indeed, the meta-regret is at most $\widetilde{\mathcal{O}}(1/\varepsilon_{i^*} + \varepsilon_{i^*} H_{i^*} L_K^{i^*})$, and by the base-regret bound we have $L_K^{i^*} \leq L_K^{\mathsf{c}} + \texttt{base-regret} \leq B_K + \widetilde{\mathcal{O}}(\sqrt{B_K(P_K + H_*)})$. The natural upper bound of B_K depends on H_* (recall that $B_K \leq H_* K$) due to the arbitrary choice of compared policies. An important technical caveat is that here we cannot simply assume the cost-to-go functions of the compared policies $\{J_k^{\pi_k^{\mathsf{c}}}\}_{1,\ldots,K}$ are bounded by that of fast policy $J_k^{\pi^f}$, in contrast to the static regret analysis where we have $\sum_{k=1}^K J_k^{\pi^*} \leq \sum_{k=1}^K J_k^{\pi^f}$ due to the optimality of the compared offline policy. Hence, even with a multi-scale meta-algorithm, meta-regret will be $\widetilde{\mathcal{O}}(H_* \sqrt{K})$ and become the dominating term, making final dynamic regret linear in H_* and thus suboptimal.

To address above issues in both base and meta levels, building upon the structure of Chen et al. (2021a), we propose a novel two-layer approach to deal with the dual uncertainty of unknown horizon length and unknown non-stationarity. To achieve this, we introduce three crucial ingredients: *groupwise scheduling* for base-learners, injecting *corrections* in feedback loss of both base- and meta-algorithm, and a new *multi-scale* meta-algorithm. Below, we first describe the base-algorithm, then introduce the scheduling method that instantiates a bunch of base-learners with different parameter configurations, and finally design the meta-algorithm to adaptively combine all the base-learners.

Base-algorithm. The base-algorithm performs OMD over a clipped occupancy measure space. At each episode $k \in [K]$, the base-algorithm receives the loss ℓ_k and performs

$$q_{k+1} = \mathop{\arg\min}_{q \in \Delta(M,H,\alpha)} \eta \langle q, \ell_k + a_k \rangle + D_\psi(q, q_k), \quad (5)$$

where $\eta > 0$ is the step size, $\Delta(M, H, \alpha) = \{q \in \Delta(M, H) \mid q(x, a) \geq \alpha, \forall x, a\}$ is the clipped space with $\alpha \in (0, 1)$, ψ is the standard negative-entropy regularizer. Notably, we inject a *correction term* $a_k \in \mathbb{R}^{|X||A|}$ to the loss, set as $a_k = 32\eta \ell_k^2, \forall k \in [K]$. The purpose is to ensure a small-loss dynamic regret and simultaneously introduce an *negative term* that will be crucial to address the difficulty occurred in controlling meta-regret (as mentioned earlier). The base-algorithm enjoys the following guarantee.

Lemma 2. *Set $q_1 = \arg\min_{q \in \Delta(M,H,\alpha)} \psi(q)$ and $\eta \leq \frac{1}{64}$, for any compared policies $\pi_{1:K}^{\mathsf{c}} \in \{\pi \mid q^\pi \in \Delta(M, H, \alpha)\}$, Eq. (5) ensures $\sum_{k=1}^K \langle q_k - q^{\pi_k^{\mathsf{c}}}, \ell_k \rangle$ is upper bounded by*

$$\frac{1}{\eta}\Big(\bar{P}_K \log \frac{H}{\alpha} + H\big(1 + \log(|X||A|H)\big)\Big) + 32\eta B_K - 16\eta S_K,$$

where $S_K = \sum_{k=1}^K \langle q_k, \ell_k^2 \rangle$ and $\bar{P}_K = \sum_{k=2}^K \|q^{\pi_k^{\mathsf{c}}} - q^{\pi_{k-1}^{\mathsf{c}}}\|_1$ is the path-length of occupancy measures.

Scheduling. Lemma 2 indicates that given a horizon length H, it is crucial to set step size properly to achieve tight dynamic regret. Since H affects the base-learner's feasible domain (i.e., the parametrized occupancy measure space), we propose a *groupwise scheduling* scheme to simultaneously

[1] Here we define L_K^{c} in a general way to accommodate changing comparators, which will be later used in the dynamic regret analysis. For static regret, it becomes $L_K^{\mathsf{c}} = \sum_{k=1}^K \langle q^{\pi^*}, \ell_k \rangle$.

adapt to unknown non-stationarity $\bar{P}_K$ and horizon length H_*. Specifically, due to $H^{\pi^f} \le H_* \le K$, we first construct a horizon length pool $\mathcal{H} = \{H_i = 2^{i-1} \cdot H^{\pi^f} \mid i \in [G]\}$ where $G = 1 + \lceil \log((K+1)/H^{\pi^f})\rceil$ to exponentially discretize the possible range; and for each H_i in the pool, we further design a step size grid $\mathcal{E}_i = \{\eta_{i,j} = 1/(32 \cdot 2^j) \mid j \in [N_i]\}$ where $N_i = \lceil \frac{1}{2} \log(\frac{4K}{1+\log(|X||A|H_i)})\rceil$ to search the optimal optimal step size associated with H_i. Overall, we maintain $N = \sum_{i=1}^{G} N_i$ base-learners, each of which associates with a specific space size and step size. More precisely, let $\mathcal{B}_{i,1:N_i}$ be a shorthand of the i-th group of base-learners $\mathcal{B}_{i,1}, \ldots, \mathcal{B}_{i,N_i}$, in which they use the same space size H_i yet different step sizes (see the configuration of $\mathcal{E}_i$). Thus, the set of all base-learners can be denoted as $\{\mathcal{B}_{1,1:N_1}, \ldots, \mathcal{B}_{i,1:N_i}, \ldots, \mathcal{B}_{G,1:N_G}\}$. The decision of base-learner $\mathcal{B}_{i,j}$ in episode k is denoted by $q_k^{i,j}$.

Meta-algorithm. The meta-algorithm requires a careful design to achieve a favorable regret. We propose a new meta-algorithm under the standard OMD framework, where additional designs are required including a novel weighted entropy regularizer and an appropriate correction term. Specifically, the meta-algorithm updates $p_{k+1} \in \Delta_N$ by

$$p_{k+1} = \operatorname*{arg\,min}_{p \in \Delta_N} \langle p, h_k + b_k \rangle + D_{\bar{\psi}}(p, p_k), \tag{6}$$

where $h_k \in \mathbb{R}^N$ is the loss of meta-algorithm, defined as $h_k^{i,j} = \langle q_k^{i,j}, \ell_k \rangle, \forall i \in [G], j \in [N_i]$. Moreover, there are two important features in the design: (i) an injected correction term $b_k \in \mathbb{R}^N$; and (ii) a weighted entropy regularizer $\bar{\psi}(p) = \sum_{i=1}^{N} \frac{1}{\varepsilon_i} p_i \log p_i$ to realize the multi-scale online learning, where $\varepsilon_i > 0$ is a multi-scale learning rate for $i \in [N]$. Below we specify the details and motivation behind such designs.

First, in meta level we inject a correction term $b_k \in \mathbb{R}^N$ as

$$b_k^{i,j} = 32\varepsilon_{i,j}(h_k^{i,j})^2, \quad \forall i \in [G], j \in [N_i]. \tag{7}$$

Let $\mathcal{B}_{i^*,j^*}$ be the base-learner whose space size H_{i^*} well approximates the unknown H_* and step size η_{i^*,j^*} well approximates the unknown optimal step size. Although injecting a correction term for the meta-algorithm was also used in (Chen et al., 2021a) to ensure a small-loss type meta-regret of the form $\widetilde{\mathcal{O}}(1/\varepsilon_{i^*,j^*} + \varepsilon_{i^*,j^*} H_{i^*} L_K^{i^*,j^*})$, as aforementioned, this will not lead to an optimal meta-regret in our case due to the undesired upper bound of $L_K^{i^*,j^*}$. Asides from that, our key novelty is to *simultaneously* exploit the correction term in the base level, which contributes an additional negative term in the base-regret $\widetilde{\mathcal{O}}((\bar{P}_K + H_{i^*})/\eta_{i^*,j^*} + \eta_{i^*,j^*} B_K - \eta_{i^*,j^*} \sum_{k=1}^{K} \langle q_k^{i^*,j^*}, \ell_k^2 \rangle)$. By a careful design of step size $\eta_{i,j}$ and learning rate $\varepsilon_{i,j}$, we can successfully cancel the positive term $\varepsilon_{i^*,j^*} H_{i^*} L_K^{i^*,j^*}$ in the meta-regret by the negative term in the base-regret.

Algorithm 2 CODO-REPS

```
Input: horizon length pool $\mathcal{H}$, step size grid $\mathcal{E}_i, \forall i \in [G]$
    and clipping parameter $\alpha$.
Define the weighted entropy $\bar{\psi}(p)$ as in Eq. (8).
Initialize $q_1^{i,j} = \arg\min_{q \in \Delta(M,H,\alpha)} \psi(q)$ and $p_1^{i,j} \propto \varepsilon_{i,j}^2, \forall i \in [G], j \in [N_i]$.
for $k = 1, \ldots, K$ do
    Receive $q_k^{i,j}$ from base-learner $\mathcal{B}_{i,j}$ by Eq. (5).
    Sample $(i_k, j_k) \sim p_k$, play the induced policy $\pi_k(a|x) \propto q_k^{i_k,j_k}(x,a), \forall x, a$.
    Define $h_k^{i,j} = \langle q_k^{i,j}, \ell_k \rangle, b_k^{i,j} = 32\varepsilon_{i,j}(h_k^{i,j})^2, \forall i, j$.
    Update weight $p_{k+1} \in \Delta_N$ by Eq. (6).
end for
```

Second, it is known that OMD with a weighted entropy regularizer leads to a multi-scale expert-tracking algorithm (Bubeck et al., 2019). In our case, we set the weighted entropy regularizer $\bar{\psi} : \Delta_N \to \mathbb{R}$ as

$$\bar{\psi}(p) = \sum_{i=1}^{G} \sum_{j=1}^{N_i} \frac{1}{\varepsilon_{i,j}} p_{i,j} \log p_{i,j}, \text{ with } \varepsilon_{i,j} = \frac{\eta_{i,j}}{2H_i}. \tag{8}$$

In above, $\eta_{i,j}$ is the step size employed by the base-learner $\mathcal{B}_{i,j}$ as specified earlier. Note that the weighted entropy regularizer depends on both space size and step size such that the final meta-algorithm can successfully handle the groupwise scheduling over the base-learners.

Combining all above ingredients yields our COrrected DO-REPS (CODO-REPS) algorithm, as summarized in Algorithm 2. We have the following dynamic regret guarantee.

Theorem 3. *Set the clipping parameter $\alpha = 1/K^3$, the horizon length pool $\mathcal{H} = \{H_i = 2^{i-1} \cdot H^{\pi^f} \mid i \in [G]\}$ where $G = 1 + \lceil \log((K+1)/H^{\pi^f})\rceil$ and the step size grid $\mathcal{E}_i = \{\eta_{i,j} = 1/(32 \cdot 2^j) \mid j \in [N_i]\}$ where $N_i = \lceil \frac{1}{2} \log(\frac{4K}{1+\log(|X||A|H_i)})\rceil$. CODO-REPS ensures*

$$\mathbb{E}[\text{D-REG}_K] \le \widetilde{\mathcal{O}}\Big(\sqrt{(H_* + \bar{P}_K)(H_* + \bar{P}_K + B_K)}\Big).$$

Remark 2. *Setting compared policies $\pi_{1:K}^{\mathsf{c}} = \pi^*$ (then $P_T = 0$ and $B_K = \sum_{k=1}^{K} J_k^{\pi^*}$), Theorem 3 implies an $\widetilde{\mathcal{O}}(\sqrt{H_* B_K})$ static regret, which gives a small-loss type bound for the episodic SSP and is new to the literature to the best of our knowledge. The bound is no worse the minimax rate $\widetilde{\mathcal{O}}(\sqrt{H_* DK})$ of Chen et al. (2021a) as $B_K = \sum_{k=1}^{K} J_k^{\pi^*} \le DK$ in the static case, and can be much better than theirs when best policy behaves well.*

We remark that the upper bound in Theorem 3 depends on the path-length of occupancy measures $\bar{P}_K$ rather than that of compared policies P_K. A natural question is how to upper bound $\bar{P}_K$ by P_K (up to multiplicative dependence on H_*). However, we show this is generally impossible for episode (non-loop-free) SSP in Theorem 7 of Appendix D.1.

Finally we show that the result in Theorem 3 is actually minimax in terms of B_K and $\bar{P}_K$ up to logarithmic factors.

Theorem 4. *For any online algorithm and any $\gamma \in [0, 2T]$, there exists an episodic SSP instance with diameter D and a sequence of compared policies $\pi_1^{\mathsf{c}}, \dots, \pi_K^{\mathsf{c}}$ with the largest hitting time H_* such that*

$$\bar{P}_K \le \gamma \text{ and } \mathbb{E}[\text{D-REG}_K] \ge \Omega(\sqrt{DH_*K(1+\gamma/H_*)})$$

under the full-information and known transition setting.

4. Infinite-horizon MDPs

This section studies infinite-horizon MDPs. We begin with the problem setup, then present a reduction to the switching-cost expert problem and establish the dynamic regret bound.

4.1. Problem Setup

An infinite-horizon MDP instance is specified by a tuple $M = (X, A, P, \{\ell_t\}_{t=1}^{\infty})$, where X, A, P are the same as introduced in Section 2, $\ell_t \in \mathbb{R}_{[0,1]}^{|X||A|}$ is the loss function at time $t \in [T]$. Unlike episodic MDPs studied in previous two sections, infinite-horizon MDPs have no goal state. The learner aims to minimize the cumulative loss over a T-step horizon in the MDP. We investigate the uniform mixing MDPs (Even-Dar et al., 2009; Neu et al., 2010b).

Definition 1 (Uniform Mixing). *There exists a constant $\tau \ge 0$ such that for any policy π and any pair of distributions μ and μ' over X, we have $\|(\mu - \mu')P^{\pi}\|_1 \le e^{-1/\tau}\|\mu - \mu'\|_1$. The smallest τ is called the mixing time.*

The uniform mixing assumption is standard and widely adopted in online MDPs studies (Even-Dar et al., 2009; Neu et al., 2010b; 2014). Nevertheless, the assumption could be strong in some sense, and recent study trying to relax the assumption by considering a larger class of communicating MDPs (Chandrasekaran & Tewari, 2021). It would be interesting to see whether our results can be extended to the communicating MDPs, and we leave this as the future work.

Occupancy measure. For an infinite-horizon MDP, the occupancy measure $q^{\pi} \in \mathbb{R}_{[0,1]}^{|X||A|}$ is defined as the stationary distribution when executing policy π, i.e., $q^{\pi}(x, a) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{x_t = x, a_t = a\}$. For an infinite-horizon MDP instance M, its occupancy measure space is defined as $\Delta(M) = \{q \in \mathbb{R}_{[0,1]}^{|X||A|} \mid \sum_{x,a} q(x,a) = 1 \text{ and } \sum_a q(x,a) = \sum_{x',a'} P(x \mid x', a')q(x', a'), \forall x \in X\}$. For any occupancy measure $q \in \Delta(M)$, its induced policy π can be obtained by $\pi(a|x) \propto q(x,a), \forall x, a$.

Dynamic regret. As defined in Eq. (2), the dynamic regret benchmarks the learner's performance against a sequence of compared policies $\pi_{1:T}^{\mathsf{c}}$, namely,

$$\mathbb{E}[\text{D-REG}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(x_t, \pi_t(x_t)) - \sum_{t=1}^{T} \ell_t(x_t, \pi_t^{\mathsf{c}}(x_t))\right].$$

The non-stationarity measure for infinite-horizon MDPs is naturally defined as $P_T = \sum_{t=2}^{T}\|\pi_t^{\mathsf{c}} - \pi_{t-1}^{\mathsf{c}}\|_{1,\infty}$.

4.2. Reduction to Switching-cost Expert Problem

In this part, we present a reduction to the switching-cost expert problem for infinite-horizon MDPs. In fact, we have the following theorem.

Theorem 5. *For infinite-horizon MDPs, the expected dynamic regret against any compared policies $\pi_{1:T}^{\mathsf{c}}$ satisfies*

$$\begin{aligned} &\mathbb{E}[\text{D-REG}_T(\pi_{1:T}^{\mathsf{c}})] &(9)\\ &\le \sum_{t=1}^{T}\langle q_t - q^{\pi_t^{\mathsf{c}}}, \ell_t\rangle + \tau'\sum_{t=2}^{T}\|q_t - q_{t-1}\|_1 + \tau'^2 P_T + 4\tau'. \end{aligned}$$

where $\tau' = \tau + 1$ is introduced to simplify the notation.

Therefore, it suffices to design an algorithm to minimize the first two terms on the right-hand side of (9), as the last two terms $\tau'^2 P_T + 4\tau'$ are not related to the algorithm. This essentially provides a generic regret reduction from infinite-horizon MDPs to the *switching-cost expert problem* (Merhav et al., 2002). Specifically, for the expert problem, at each round $t \in [T]$, the learner chooses a decision $q_t \in \Delta_N$ as a weight over all N experts, receives the loss $\ell_t \in \mathbb{R}^N$ and suffers loss $\langle q_t, \ell_t\rangle$. In addition to the cumulative loss $\sum_{t=1}^{T}\langle q_t, \ell_t\rangle$, the switching-cost expert problem further takes the actions' switch into account by adding $\lambda\sum_{t=2}^{T}\|q_t - q_{t-1}\|_1$ as penalty, $\lambda > 0$ is the coefficient.

Our reduction also holds for the static regret (simply choosing all compared policies as a fixed one), perhaps surprisingly, there is no explicit reduction in the literature to the best of our knowledge, though proof of Theorem 5 is simple and all the ingredients are already in the pioneering work (Even-Dar et al., 2009). As another note, Agarwal et al. (2019) study online non-stochastic control and give a reduction to the switching-cost online learning problem (or called online convex optimization with memory), while their reduction does not apply to infinite-horizon MDPs.

4.3. Dynamic Regret

With the reduction on hand, we now consider the design of a two-layer approach to optimize the dynamic regret of the switching-cost expert problem. It turns out that a recent result (Zhao et al., 2022) has resolved that expert problem, building upon which we propose REgularized DO-REPS (REDO-REPS) algorithm for infinite-horizon MDPs.

As discussed before, it suffices to design an algorithm to minimize the first two terms in (9), namely, the dynamic regret in terms of the occupancy measure and a switching cost term. Notice that the first term also appears in optimizing dynamic regret of the episodic loop-free SSP (see Eq. (3)). Thus, a natural idea is to run (Algorithm 1) over the occupancy measure space $\Delta(M, \alpha) = \{q \in \Delta(M) \mid q(x, a) \geq \alpha, \forall x, a\}$. Specifically, we maintain N base-learners denoted by $\mathcal{B}_1, \ldots, \mathcal{B}_N$, where $\mathcal{B}_i$ generates the prediction $q_{t,i}$ by performing O-REPS with a particular step size η_i in the step size pool $\mathcal{H}$; then the meta-algorithm combines all the predictions to produce the final decision $q_t = \sum_{i=1}^N p_{t,i} q_{t,i}$ and updates the weight p_t. However, DO-REPS does not take the switching cost into account, leading to undesired behavior in this problem. To see this, it can be verified that the one-step switching cost can be decomposed as

$$\|q_t - q_{t-1}\|_1 \leq \sum_{i=1}^N p_{t,i} \|q_{t,i} - q_{t-1,i}\|_1 + \|p_t - p_{t-1}\|_1.$$

Summing over T, the second term in the right-hand side is the meta-algorithm's switching cost, which can be easily bounded by $\mathcal{O}(\sqrt{T})$ for common expert-tracking algorithms. However, the first term is the weighted switching cost of all base-learners, which could be very large and even grow linearly with iterations due to the base-learners with large step sizes. For example, when employing OMD as the base-algorithm, the switching cost of $\mathcal{B}_i$ is of order $\mathcal{O}(\eta_i T)$. Then, the construction of step size pool requires that $\eta_N = \mathcal{O}(1)$, leading to an $\mathcal{O}(T)$ switching cost of the base-learner $\mathcal{B}_N$, which ruins the overall regret bound. To address this, inspired by the recent progress on OCO with memory (Zhao et al., 2022), we add a switching-cost regularization in evaluating each base-learner, i.e., the feedback loss of the meta-algorithm $h_t \in \mathbb{R}^N$ is constructed as

$$h_{t,i} = \langle q_{t,i}, \ell_t \rangle + \lambda \|q_{t,i} - q_{t-1,i}\|_1. \tag{10}$$

Set $\lambda = \tau'$, it can be verified the first two terms $\sum_{t=1}^T \langle q_t - q^{\pi_t^{\mathsf{c}}}, \ell_t \rangle + \tau' \sum_{t=2}^T \|q_t - q_{t-1}\|_1$ in (9) can be written as

$$\sum_{t=1}^T (\langle p_t, h_t \rangle - h_{t,i}) + \tau' \sum_{t=2}^T \|p_t - p_{t-1}\|_1 \\ + \sum_{t=1}^T \langle q_{t,i} - q^{\pi_t^{\mathsf{c}}}, \ell_t \rangle + \tau' \sum_{t=2}^T \|q_{t,i} - q_{t-1,i}\|_1.$$

We have decomposed the switching-cost dynamic regret into two parts — the first part is the meta-regret over the regularized loss h_t that measures the regret overhead of the meta-algorithm penalized by the switching cost, and the second part is the base-regret of a specific base-learner $\mathcal{B}_i$ taking her switching cost into account. By slightly modifying DO-REPS (Algorithm 1), we get REgularized DO-REPS (REDO-REPS) algorithm as shown in Algorithm 3. The key difference is the designed switching-cost-regularized loss for meta-algorithm's updates in Lines 7–8, such that the overall two-layer approach enjoys nice following guarantee.

Algorithm 3 REDO-REPS

```
Input: step size pool $\mathcal{H}$, learning rate $\varepsilon$, clipping param $\alpha$.
1: Define: $\psi(q) = \sum_{x,a} q(x,a) \log q(x,a)$.
2: Initialization: set $q_{1,i} = \arg\min_{q \in \Delta(M,\alpha)} \psi(q)$ and
   $p_{1,i} = 1/N$ for $\forall i \in [N]$.
3: for $t = 1$ to $T$ do
4:     Receive $q_{t,i}$ from base-learner $\mathcal{B}_i, \forall i \in [N]$.
5:     Compute $q_t = \sum_{i=1}^N p_{t,i} q_{t,i}$, play the induced policy
       $\pi_t(a|x_t) \propto q_t(x_t, a), \forall a \in A$.
6:     Suffer loss $\ell_t(x_t, a_t)$ and observe loss function $\ell_t$.
7:     Construct switching-cost-regularized loss by Eq. (10)
8:     Update weight by $p_{t+1,i} \propto \exp(-\varepsilon \sum_{s=1}^t h_{s,i})$.
9:     Each base-learner $\mathcal{B}_i$ updates prediction by
       $q_{t+1,i} = \arg\min_{q \in \Delta(M,\alpha)} \eta_i \langle q, \ell_t \rangle + D_\psi(q, q_{t,i})$.
10: end for
```

Theorem 6. *Set the clipping parameter $\alpha = 1/T^2$, the step size pool $\mathcal{H} = \{2^{i-1}\sqrt{T^{-1}\log|X||A|} \mid i \in [N]\}$ where $N = \lceil \frac{1}{2}\log(1 + \frac{4T\log T}{\log|X||A|}) \rceil + 1$ and the learning rate $\varepsilon = (2\tau + 3)^{-1}\sqrt{(\log N)/2T}$. REDO-REPS ensures*

$$\mathbb{E}[\text{D-REG}_T] \leq \mathcal{O}\big(\sqrt{\tau T(\log|X||A| + \tau P_T \log T)} + \tau^2 P_T\big).$$

Remark 3. *Set $\pi_{1:T}^{\mathsf{c}} = \pi^* \in \arg\min_\pi \sum_{t=1}^T \ell_t(x_t, \pi(x_t))$ (then $P_T = 0$), we recover the best known static regret $\mathcal{O}(\sqrt{\tau T \log|X||A|})$ (Zimin & Neu, 2013). Our dynamic regret scales with the path-length of compared policies rather than that of occupancy measures. We achieve so by establishing their relationships in Lemma 11 of Appendix E.1.*

5. Conclusion

In this paper we investigate learning in three foundational online MDPs with adversarially changing loss functions and known transition kernel. We propose novel online ensemble algorithms and establish their dynamic regret guarantees for the first time. In particular, the results for episodic (loop-free) SSP are provably minimax optimal in terms of time horizon and certain non-stationarity measure.

Our results present an initial resolution for dynamic regret of online MDPs. There are plenty of future works to investigate. For example, this paper focuses on the full-information feedback and known transition, and how to extend the results to the bandit feedback and unknown transition setting is important and challenging. Moreover, it is interesting to further consider function approximation in those models.

Acknowledgements

This research was supported by NSFC (61921006) and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

Agarwal, N., Bullins, B., Hazan, E., Kakade, S. M., and Singh, K. Online control with adversarial disturbances. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 111–119, 2019.

Auer, P., Cesa-Bianchi, N., and Gentile, C. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, pp. 48–75, 2002.

Baby, D. and Wang, Y.-X. Online forecasting of total-variation-bounded sequences. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pp. 11071–11081, 2019.

Baby, D. and Wang, Y.-X. Optimal dynamic regret in exp-concave online learning. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, pp. 359–409, 2021.

Baby, D. and Wang, Y.-X. Optimal dynamic regret in proper online learning with strongly convex losses and beyond. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1805–1845, 2022.

Bertsekas, D. P. and Tsitsiklis, J. N. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

Besbes, O., Gur, Y., and Zeevi, A. J. Non-stationary stochastic optimization. *Operations Research*, pp. 1227–1244, 2015.

Bubeck, S., Devanur, N. R., Huang, Z., and Niazadeh, R. Multi-scale online learning: Theory and applications to online auctions and pricing. *Journal of Machine Learning Research*, 20:62:1–62:37, 2019.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1283–1294, 2020.

Cesa-Bianchi, N., Gaillard, P., Lugosi, G., and Stoltz, G. Mirror descent meets fixed share (and feels no regret). In *Advances in Neural Information Processing Systems 25 (NIPS)*, pp. 989–997, 2012.

Chandrasekaran, G. and Tewari, A. Online learning in adversarial MDPs: Is the communicating case harder than ergodic? *ArXiv preprint*, 2111.02024, 2021.

Chen, G. and Teboulle, M. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, pp. 538–543, 1993.

Chen, L., Luo, H., and Wei, C.-Y. Minimax regret for stochastic shortest path with adversarial costs and known transition. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, pp. 1180–1215, 2021a.

Chen, L., Luo, H., and Wei, C.-Y. Impossible tuning made possible: A new expert algorithm and its applications. In *Proceedings of 34th Conference on Learning Theory (COLT)*, pp. 1216–1259, 2021b.

Chen, X., Wang, Y., and Wang, Y.-X. Non-stationary stochastic optimization under $L_{p,q}$-variation measures. *Operations Research*, 67(6):1752–1765, 2019.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1843–1854, 2020.

Chiang, C.-K., Yang, T., Lee, C.-J., Mahdavi, M., Lu, C.-J., Jin, R., and Zhu, S. Online optimization with gradual variations. In *Proceedings of the 25th Conference On Learning Theory (COLT)*, pp. 6.1–6.20, 2012.

Daniely, A., Gonen, A., and Shalev-Shwartz, S. Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1405–1411, 2015.

Domingues, O. D., Ménard, P., Pirotta, M., Kaufmann, E., and Valko, M. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3538–3546, 2021.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Online Markov decision processes. *Mathematics of Operations Research*, pp. 726–736, 2009.

Fei, Y., Yang, Z., Wang, Z., and Xie, Q. Dynamic regret of policy optimization in non-stationary environments. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Gajane, P., Ortner, R., and Auer, P. A sliding-window algorithm for Markov decision processes with arbitrarily changing rewards and transitions. *ArXiv preprint*, arXiv:1805.10066, 2018.

György, A. and Szepesvári, C. Shifting regret, mirror descent, and matrices. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pp. 2943–2951, 2016.

Hazan, E. Introduction to Online Convex Optimization. *Foundations and Trends in Optimization*, pp. 157–325, 2016.

Hazan, E. and Seshadhri, C. Efficient learning algorithms for changing environments. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 393–400, 2009.

Jadbabaie, A., Rakhlin, A., Shahrampour, S., and Sridharan, K. Online optimization : Competing with dynamic comparators. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 398–406, 2015.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, pp. 1563–1600, 2010.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 4860–4869, 2020a.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of the 33rd Conference on Learning Theory (COLT)*, pp. 2137–2143, 2020b.

Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, D., Allen, J., Lam, V., Bewley, A., and Shah, A. Learning to drive in a day. In *Proceedings of the 2019 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8248–8254, 2019.

Luo, H., Zhang, M., Zhao, P., and Zhou, Z.-H. 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)*, 2022.

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., and Basar, T. Near-optimal model-free reinforcement learning in non-stationary episodic MDPs. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 7447–7458, 2021.

Merhav, N., Ordentlich, E., Seroussi, G., and Weinberger, M. J. On sequential strategies for loss functions with memory. *IEEE Transactions on Information Theory*, 48 (7):1947–1958, 2002.

Neu, G., György, A., and Szepesvári, C. The online loop-free stochastic shortest-path problem. In *Proceedings of 23rd Conference on Learning Theory (COLT)*, pp. 231–243, 2010a.

Neu, G., György, A., Szepesvári, C., and Antos, A. Online Markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pp. 1804–1812, 2010b.

Neu, G., György, A., and Szepesvári, C. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 805–813, 2012.

Neu, G., György, A., Szepesvári, C., and Antos, A. Online Markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, pp. 676–691, 2014.

Ortner, R., Gajane, P., and Auer, P. Variational regret bounds for reinforcement learning. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 81–90, 2019.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 5478–5486, 2019a.

Rosenberg, A. and Mansour, Y. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems 32 (NIPS)*, pp. 2209–2218, 2019b.

Rosenberg, A. and Mansour, Y. Stochastic shortest path with adversarially changing costs. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2936–2942, 2021.

Schulman, J., Levine, S., Abbeel, P., Jordan, M. I., and Moritz, P. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1889–1897, 2015.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T. P., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. Mastering the game of go with deep neural networks and tree search. *Nature*, pp. 484–489, 2016.

Touati, A. and Vincent, P. Efficient learning in non-stationary linear Markov decision processes. *ArXiv preprint*, arXiv:2010.12870, 2020.

Wei, C.-Y. and Luo, H. Non-stationary reinforcement learning without prior knowledge: an optimal black-box approach. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, pp. 4300–4354, 2021.

Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 10746–10756, 2020.

Yang, T., Zhang, L., Jin, R., and Yi, J. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 449–457, 2016.

Yu, J. Y., Mannor, S., and Shimkin, N. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, pp. 737–757, 2009.

Zhang, L., Lu, S., and Zhou, Z.-H. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 1330–1340, 2018.

Zhang, Y.-J., Zhao, P., and Zhou, Z.-H. A simple online algorithm for competing with dynamic comparators. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 390–399, 2020.

Zhao, P. and Zhang, L. Improved analysis for dynamic regret of strongly convex and smooth functions. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control (L4DC)*, pp. 48–59, 2021.

Zhao, P., Zhang, L., Jiang, Y., and Zhou, Z.-H. A simple approach for non-stationary linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 746–755, 2020a.

Zhao, P., Zhang, Y.-J., Zhang, L., and Zhou, Z.-H. Dynamic regret of convex and smooth functions. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 12510–12520, 2020b.

Zhao, P., Wang, G., Zhang, L., and Zhou, Z.-H. Bandit convex optimization in non-stationary environments. *Journal of Machine Learning Research*, 22(125):1–45, 2021a.

Zhao, P., Zhang, Y.-J., Zhang, L., and Zhou, Z.-H. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *ArXiv preprint*, arXiv:2112.14368, 2021b.

Zhao, P., Wang, Y.-X., and Zhou, Z.-H. Non-stationary online learning with memory and non-stochastic control. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. to appear, 2022.

Zhou, H., Chen, J., Varshney, L. R., and Jagmohan, A. Nonstationary reinforcement learning with linear function approximation. *ArXiv preprint*, arXiv:2010.04244, 2020.

Zimin, A. and Neu, G. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pp. 1583–1591, 2013.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pp. 928–936, 2003.

A. Related Work

We presents discussions on several topics related to this work. The first part is about the development of static regret for online adversarial MDPs, and the second part reviews related advance of dynamic regret in non-stationary online learning.

A.1. Online Adversarial MDPs

Learning adversarial MDPs has attracted much attention in recent years. We briefly discuss related works on three models of online MDPs studied in this paper, including episodic loop-free SSP, episodic SSP, and infinite-horizon MDPs.

Episodic loop-free SSP. Neu et al. (2010a) first study learning in the episodic SSP with a loop-free structure and known transition, where an $\widetilde{\mathcal{O}}(H^2\sqrt{K})$ regret is achieved in the full information setting and K is the number of the episodes and H is the horizon length in each episode. Later Zimin & Neu (2013) propose the O-REPS algorithm which applies mirror descent over occupancy measure space and achieves the optimal regret of order $\widetilde{\mathcal{O}}(H\sqrt{K})$. Neu et al. (2010a); Zimin & Neu (2013) also consider the bandit feedback setting. Neu et al. (2012); Rosenberg & Mansour (2019a) investigate the unknown transition kernel and full-information setting. Rosenberg & Mansour (2019b) and Jin et al. (2020a) further consider the harder unknown transition kernel and bandit-feedback setting. The linear function approximation setting is also studied (Cai et al., 2020). Notably, our results for episodic loop-free SSP (see Section 2) focus on known transition and full-information feedback setting. Different from all mentioned results minimizing static regret, our proposed algorithm is equipped with dynamic regret guarantee, which can recover the $\widetilde{\mathcal{O}}(H\sqrt{K})$ minimax optimal static regret when choosing compared policies as the best fixed policy in hindsight. Furthermore, when the environments are predictable, we enhance the algorithm to capture such adaptivity and hence enjoy better dynamic regret guarantees than the minimax rate.

Episodic SSP. Rosenberg & Mansour (2021) first consider learning in episodic (non-loop-free) SSP with full-information loss feedback. Their algorithm achieves an $\widetilde{\mathcal{O}}(\frac{D}{c_{\min}}\sqrt{K})$ regret for the known transition setting, where $c_{\min} \in (0, 1]$ is the lower bound of the loss function and D is the diameter of the MDP. They also study the zero costs case and unknown transition setting. Chen et al. (2021a) develop algorithms that significantly improve the results and achieve minimax regret $\widetilde{\mathcal{O}}(\sqrt{H^{\pi^*}DK})$ for the full information with known transition setting, where H^{π^*} is the hitting time of the optimal policy. They also investigate the unknown transition setting. Our results for episodic SSP (see Section 3) focus on the known transition and full-information setting. We develop an algorithm with optimal dynamic regret guarantees. Our result immediately recovers the optimal $\widetilde{\mathcal{O}}(\sqrt{H^{\pi^*}DK})$ static regret when setting comparators as the best fixed policy in hindsight. We further enhance our algorithm to achieve a more adaptive bound when the environments are predictable.

Infinite-horizon MDPs. Even-Dar et al. (2009) consider learning in unichain MDPs with known transition and full-information feedback, they propose the algorithm MDP-E that enjoys $\widetilde{\mathcal{O}}(\sqrt{\tau^3 T})$ regret, where τ is the mixing time. Another work (Yu et al., 2009) achieves $\widetilde{\mathcal{O}}(T^{2/3})$ regret in a similar setting. The O-REPS algorithm of Zimin & Neu (2013) achieves an $\widetilde{\mathcal{O}}(\sqrt{\tau T})$ regret. Neu et al. (2010b; 2014) consider the known transition kernel and bandit feedback setting. These studies focus on the MDPs with uniform mixing properties, which could be strong. Recent study tries to relax the assumption by considering the larger class of communicating MDPs (Chandrasekaran & Tewari, 2021). Our results for infinite-horizon MDPs (see Section 4) focus on the known transition and full-information feedback setting and propose an algorithm that enjoys dynamic regret which can recover the best-known $\widetilde{\mathcal{O}}(\sqrt{\tau T})$ static regret.

Discussion. We note that all those works focus on the static regret minimization, and our work establishes the dynamic regret for all the three online MDPs models. In a setting most similar to ours, Fei et al. (2020) investigate the dynamic regret of episodic loop-free SSP (with function approximation). They propose two model-free algorithms and prove the dynamic regret bound scaling with non-stationarity of environments. However, we note that their algorithms require the prior knowledge of non-stationarity measure P_T as input, which is generally unavailable to the learner in practice. By contrast, our proposed algorithms are *parameter-free* to those unknown quantities related to the underlying environments (including non-stationarity measure P_T and adaptivity quantity V_T). More importantly, we also consider dynamic regret of two more challenging settings of online MDPs — episodic (non-loop-free) SSP and infinite-horizon MDPs.

A.2. Non-stationary Online Learning

In this part, we first discuss related works of online non-stationary MDPs (whose loss functions are stochastic, whereas we study the adversarial setting) then discuss dynamic regret of online convex optimization whose techniques are related to us.

Online Non-stationary MDPs. Another related line of research is on the online non-stationary MDPs. More specifically, in contrast to learning with adversarial MDPs where the online loss functions are generated in an adversarial way, online non-stationary MDPs consider the setting where reward (loss) functions are generated in a *stochastic* way according to a certain reward distribution that might be non-stationary over the time. For infinite-horizon MDPs, Jaksch et al. (2010) consider the piecewise-stationary setting where the losses and transition kernels are allowed to change a fixed number and then propose UCRL2 with restarting mechanism to handle the non-stationarity. Later, Gajane et al. (2018) propose an alternative approach based on the sliding-window update for the same setting, and is later generalized to more general non-stationary setting with gradual drift (Ortner et al., 2019). However, all above approaches require the prior knowledge on the degree of non-stationarity, either the number of piecewise changes or the tensity of gradual drift. Recently, Cheung et al. (2020) propose the Bandit-over-RL algorithm to remove the requirement of unknown non-stationarity measure, but nevertheless can only obtain suboptimal result. Other results for non-stationary MDPs includes episodic non-stationary MDPs (Mao et al., 2021; Domingues et al., 2021) and episodic non-stationary linear MDPs (Touati & Vincent, 2020; Zhou et al., 2020). The techniques in those studies are related to the thread of stochastic linear bandits (Jin et al., 2020b; Yang & Wang, 2020; Zhao et al., 2020a). A recent breakthrough is made by Wei & Luo (2021), who propose a black-box approach that can turn a certain algorithm with optimal static regret in a stationary environment into another algorithm with optimal dynamic regret in a non-stationary environment, and more importantly, the overall approach does not require any prior knowledge on the degree of non-stationarity. They achieve optimal dynamic regret for episodic tabular MDPs (Mao et al., 2021; Zhou et al., 2020; Touati & Vincent, 2020). For infinite-horizon MDPs, they can achieve optimal dynamic regret when the maximum diameter of MDP is known or the degree of non-stationarity is known (Gajane et al., 2018; Cheung et al., 2020); when none of them is know, they attain suboptimal regret but is still the best-known result.

Non-stationary Online Convex Optimization. Online convex optimization (OCO) is a fundamental and versatile framework for modeling online prediction problems (Hazan, 2016). Dynamic regret of OCO has drawn increasing attention in recent years, and techniques are highly related to ours. We here briefly review some related results and refer the reader to the latest paper (Zhao et al., 2021b) for a more thorough treatment. Dynamic regret ensures the online learner to be competitive with a sequence of changing comparators, and is sometimes called tracking regret or switching regret in the study of prediction with expert advice setting (Cesa-Bianchi et al., 2012). As mentioned in Section 1, this paper focuses on the general dynamic regret that allows the any feasible comparators in the decision set, which is also called *universal* dynamic regret. A special variant is called *worst-case* dynamic regret, which only competes with the sequence of minimizers of online functions and has gained much attention in the literature (Besbes et al., 2015; Jadbabaie et al., 2015; Yang et al., 2016; György & Szepesvári, 2016; Chen et al., 2019; Baby & Wang, 2019; Zhang et al., 2020; Zhao & Zhang, 2021). However, the worst-case dynamic regret would be problematic or even misleading in many cases, for example, approaching the minimizer of each-round online function would lead to overfitting when the environments admit some noise (Zhang et al., 2018). Thus, the universal dynamic regret is generally more desired to be performance measure for algorithm design in non-stationary online learning. We now introduce the results in this regard. Zinkevich (2003) first considers the universal dynamic regret of OCO and shows that Online Gradient Descent (OGD) enjoys $\mathcal{O}(\sqrt{T}(1+P_T))$ dynamic regret, where P_T is the path-length of the comparators reflecting the non-stationarity of the environments. Later, Zhang et al. (2018) propose a novel algorithm and prove a minimax optimal $\mathcal{O}(\sqrt{T(1+P_T)})$ dynamic regret guarantee without requiring the knowledge of unknown P_T. Their proposed algorithm employs the meta-base structure, which turns out to be a key component to handle unknown non-stationarity measure P_T. When the environments are predictable and the loss functions are convex and smooth, Zhao et al. (2020b; 2021b) develop an algorithm, achieving problem-dependent dynamic regret which could be much smaller than the minimax rate. Baby & Wang (2021; 2022) consider OCO with exp-concave or strongly convex loss functions. Dynamic regret of bandit online learning is studied for adversarial linear bandits (Luo et al., 2022) and bandit convex optimization (Zhao et al., 2021a). More discussions can be found in the latest advance (Zhao et al., 2021b).

B. Useful Lemmas Related to Online Mirror Descent

In this section, we present several important lemmas used frequently in the analysis of (optimistic) online mirror descent.

Lemma 3 (Lemma 3.2 of Chen & Teboulle (1993))**.** *Define* $q^* = \arg\min_{q\in\mathcal{K}} \eta\langle q, \ell\rangle + D_\psi(q, \widehat{q})$ *for some compact set* $\mathcal{K} \subseteq \mathbb{R}^d$*, convex function* ψ*, an arbitrary point* $\ell \in \mathbb{R}^d$*, and a point* $\widehat{q} \in \mathcal{K}$*. Then for any* $u \in \mathcal{K}$*,*

$$\langle q^* - u, \ell\rangle \leq \frac{1}{\eta}(D_\psi(u, \widehat{q}) - D_\psi(u, q^*) - D_\psi(q^*, \widehat{q})).$$

Lemma 4 (Lemma 5 and Proposition 7 of Chiang et al. (2012))**.** *Let* $q_t = \arg\min_{q \in \mathcal{K}} \eta\langle q, m_t\rangle + D_\psi(q, \widehat{q}_t)$ *and* $\widehat{q}_{t+1} = \arg\min_{q\in\mathcal{K}} \eta\langle q, \ell_t\rangle + D_\psi(q, \widehat{q}_t)$ *for some compact convex set* $\mathcal{K} \subseteq \mathbb{R}^d$*, convex function* ψ*, arbitrary points* $\ell_t, m_t \in \mathbb{R}^d$*, and a point* $\widehat{q}_1 \in \mathcal{K}$*. Then, for any* $u \in \mathcal{K}$*,*

$$\langle q_t - u, \ell_t\rangle \le \langle q_t - \widehat{q}_{t+1}, \ell_t - m_t\rangle + \frac{1}{\eta}(D_\psi(u, \widehat{q}_t) - D_\psi(u, \widehat{q}_{t+1}) - D_\psi(\widehat{q}_{t+1}, q_t) - D_\psi(q_t, \widehat{q}_t)),$$

and when ψ *is* λ*-strongly convex function w.r.t. the norm* $\|\cdot\|$*, we have*

$$\|q_t - \widehat{q}_{t+1}\| \le \frac{1}{\lambda}\|\ell_t - m_t\|_*.$$

Lemma 5 (Lemma 1 of Chen et al. (2021b))**.** *Define* $\psi(q) = \sum_{i=1}^d \frac{1}{\eta_i} q_i \log q_i$*, where* d *is the dimension of* q*. Let* $a_t \in \mathbb{R}^d$ *with* $a_{t,i} = 32\eta_i(\ell_{t,i} - m_{t,i})^2$*,* $q_t = \arg\min_{q\in\mathcal{K}}\langle q, m_t\rangle + D_\psi(q, \widehat{q}_t)$ *and* $\widehat{q}_{t+1} = \arg\min_{q\in\mathcal{K}}\langle q, \ell_t + a_t\rangle + D_\psi(q, \widehat{q}_t)$ *for some compact convex set* $\mathcal{K} \subseteq \mathbb{R}^d$*, arbitrary points* $\ell_t, m_t \in \mathbb{R}^d$*, and a point* $\widehat{q}_t \in \mathcal{K}$*. Suppose* $32\eta_i|\ell_{t,i} - m_{t,i}| \le 1$ *holds for all* $i \in [d]$*. Then, for any* $u \in \mathcal{K}$*,*

$$\langle q_t - u, \ell_t\rangle \le D_\psi(u, \widehat{q}_t) - D_\psi(u, \widehat{q}_{t+1}) + 32\sum_{i=1}^d \eta_i u_i(\ell_{t,i} - m_{t,i})^2 - 16\sum_{i=1}^d \eta_i q_{t,i}(\ell_{t,i} - m_{t,i})^2.$$

Proof. This lemma is originally proven by Chen et al. (2021b). For self-containedness, we present their proof and adapt to our notations. By Lemma 4, we have

$$\langle q_t - u, \ell_t + a_t\rangle \le D_\psi(u, \widehat{q}_t) - D_\psi(u, \widehat{q}_{t+1}) + \langle q_t - \widehat{q}_{t+1}, \ell_t - m_t + a_t\rangle - D_\psi(\widehat{q}_{t+1}, q_t).$$

For the last two terms, define $q^* = \arg\max_{q\in\mathbb{R}^d_{>0}}\langle q_t - q, \ell_t - m_t + a_t\rangle + D_\psi(q, q_t)$, by the optimality of q^*, we have: $\ell_t - m_t + a_t = \nabla\psi(q_t) - \nabla\psi(q^*)$ and thus $q^*_i = q_{t,i}e^{-\eta_i(\ell_{t,i} - m_{t,i} + a_{t,i})}$. Therefore, we have

$$\begin{aligned}
&\langle q_t - \widehat{q}_{t+1}, \ell_t - m_t + a_t\rangle - D_\psi(\widehat{q}_{t+1}, q_t) \\
&\le \langle q_t - q^*, \ell_t - m_t + a_t\rangle - D_\psi(q^*, q_t) \\
&= \langle q_t - q^*, \nabla\psi(q_t) - \nabla\psi(q^*)\rangle - D_\psi(q^*, q_t) \\
&= D_\psi(q_t, q^*) = \sum_{i=1}^d \frac{1}{\eta_i}\left(q_{t,i}\log\frac{q_{t,i}}{q^*_i} - q_{t,i} + q^*_i\right) \\
&= \sum_{i=1}^d \frac{q_{t,i}}{\eta_i}\left(\eta_i(\ell_{t,i} - m_{t,i} + a_{t,i}) - 1 + e^{-\eta_i(\ell_{t,i} - m_{t,i} + a_{t,i})}\right) \\
&\le \sum_{i=1}^d \eta_i q_{t,i}(\ell_{t,i} - m_{t,i} + a_{t,i})^2,
\end{aligned}$$

where the last inequality holds due to the fact $e^{-x} - 1 + x \le x^2$ for $x \ge -1$ and the condition that $\eta_i|\ell_{t,i} - m_{t,i}| \le \frac{1}{32}$ such that $\eta_i|\ell_{t,i} - m_{t,i} + a_{t,i}| \le \eta_i|\ell_{t,i} - m_{t,i}| + 32\eta_i^2(\ell_{t,i} - m_{t,i})^2 \le \frac{1}{16}$. Using the definition of a_t and the condition $\eta_i|\ell_{t,i} - m_{t,i}| \le \frac{1}{32}$, we have

$$\langle q_t - \widehat{q}_{t+1}, \ell_t - m_t + a_t\rangle - D_\psi(\widehat{q}_{t+1}, q_t) \le \sum_{i=1}^d \eta_i q_{t,i}(\ell_{t,i} - m_{t,i} + 32\eta_i(\ell_{t,i} - m_{t,i})^2)^2 \le 4\sum_{i=1}^d \eta_i q_{t,i}(\ell_{t,i} - m_{t,i})^2.$$

To sum up, we have

$$\langle q_t - u, \ell_t + a_t\rangle \le D_\psi(u, \widehat{q}_t) - D_\psi(u, \widehat{q}_{t+1}) + 4\sum_{i=1}^d \eta_i q_{t,i}(\ell_{t,i} - m_{t,i})^2.$$

Finally, moving $\langle q_t - u, a_t\rangle$ to the right-hand side of the inequality and using the definition of a_t finishes the proof. □

C. Proofs for Section 2 (Episodic Loop-free SSP)

In this section, we provide the proofs for Section 2. First, we introduce the relationship between the path-length of policies and the path-length of occupancy measures, and then provide proof of the dynamic regret of DO-REPS algorithm in Section 2.2. Finally we present the proof of the dynamic regret lower bound.

C.1. Path-length of Policies and Occupancy Measures

This part introduces the relationship between the path-length of policies and path-length of occupancy measures.

Lemma 6. *For any occupancy measure sequence $q^{\pi_1},\dots,q^{\pi_K}$ induced by the policy sequence $\pi_1,\dots,\pi_K$, it holds that*

$$\sum_{k=2}^{K}\|q^{\pi_k}-q^{\pi_{k-1}}\|_1 \le H\sum_{k=2}^{K}\sum_{l=0}^{H-1}\|\pi_{k,l}-\pi_{k-1,l}\|_{1,\infty}.$$

Proof. Let $d_l^{\pi_k}(x)\triangleq\sum_a q^{\pi_k}(x,a),\forall x\in X_l, k\in[K]$, we have

$$\begin{aligned}
&\sum_{x,a}|q^{\pi_k}(x,a)-q^{\pi_{k-1}}(x,a)|\\
&=\sum_{l=0}^{H-1}\sum_{x\in X_l}\sum_{a}|q^{\pi_k}(x,a)-q^{\pi_{k-1}}(x,a)|\\
&=\sum_{l=0}^{H-1}\sum_{x\in X_l}\sum_{a}|d_l^{\pi_k}(x)\pi_k(a|x)-d_l^{\pi_{k-1}}(x)\pi_{k-1}(a|x)|\\
&\le\sum_{l=0}^{H-1}\sum_{x\in X_l}\sum_{a}|d_l^{\pi_k}(x)\pi_k(a|x)-d_l^{\pi_{k-1}}(x)\pi_k(a|x)|+|d_l^{\pi_{k-1}}(x)\pi_k(a|x)-d_l^{\pi_{k-1}}(x)\pi_{k-1}(a|x)|\\
&=\sum_{l=0}^{H-1}\sum_{x\in X_l}|d_l^{\pi_k}(x)-d_l^{\pi_{k-1}}(x)|\sum_a\pi_k(a|x)+\sum_{l=0}^{H-1}\sum_{x\in X_l}d_l^{\pi_{k-1}}(x)\sum_a|\pi_k(a|x)-\pi_{k-1}(a|x)|\\
&\le\sum_{l=0}^{H-1}\|d_l^{\pi_k}-d_l^{\pi_{k-1}}\|_1+\sum_{l=0}^{H-1}\|\pi_{k,l}-\pi_{k-1,l}\|_{1,\infty},
\end{aligned}\tag{11}$$

where the first inequality due to the triangle inequality. Next, we bound the term $\|d_l^{\pi_k}-d_l^{\pi_{k-1}}\|_1$. Since $X_0=\{x_0\}$, we have $\|d_0^{\pi_k}-d_0^{\pi_{k-1}}\|_1=0$. For $l\ge1$, we have

$$\begin{aligned}
\|d_l^{\pi_k}-d_l^{\pi_{k-1}}\|_1&=\|d_{l-1}^{\pi_k}P^{\pi_k}-d_{l-1}^{\pi_{k-1}}P^{\pi_{k-1}}\|_1\\
&\le\|d_{l-1}^{\pi_k}P^{\pi_k}-d_{l-1}^{\pi_k}P^{\pi_{k-1}}\|_1+\|d_{l-1}^{\pi_k}P^{\pi_{k-1}}-d_{l-1}^{\pi_{k-1}}P^{\pi_{k-1}}\|_1\\
&\le\|\pi_{k,l-1}-\pi_{k-1,l-1}\|_{1,\infty}+\|d_{l-1}^{\pi_k}-d_{l-1}^{\pi_{k-1}}\|_1,
\end{aligned}$$

where the last inequality holds due to Lemma 7 and Lemma 8. Summing the above inequality from 1 to l, we have

$$\|d_l^{\pi_k}-d_l^{\pi_{k-1}}\|_1\le\sum_{i=0}^{l-1}\|\pi_{k,i}-\pi_{k-1,i}\|_{1,\infty}.\tag{12}$$

Combining (11) and (12), we obtain

$$\begin{aligned}
\sum_{x,a}|q^{\pi_k}(x,a)-q^{\pi_{k-1}}(x,a)|&\le\sum_{l=0}^{H-1}\sum_{i=0}^{l-1}\|\pi_{k,i}-\pi_{k-1,i}\|_{1,\infty}+\sum_{l=0}^{H-1}\|\pi_{k,l}-\pi_{k-1,l}\|_{1,\infty}\\
&=\sum_{l=0}^{H-1}\sum_{i=0}^{l}\|\pi_{k,i}-\pi_{k-1,i}\|_{1,\infty}\le H\sum_{l=0}^{H-1}\|\pi_{k,l}-\pi_{k-1,l}\|_{1,\infty}.
\end{aligned}$$

We complete the proof by summing the inequality over all iterations. □

C.2. Proof of Lemma 1

Proof. Denote by $q'_{k+1} = \arg\min_q \eta\langle q, \ell_k\rangle + D_\psi(q, q_k)$, or equivalently, $q'_{k+1}(x,a) = q_k(x,a)\exp(-\eta\ell_k(x,a))$. By standard analysis of online mirror descent, we have

$$\begin{aligned}\sum_{k=1}^K \langle q_k - q^{\pi^c_k}, \ell_k\rangle &= \sum_{k=1}^K \langle q_k - q'_{k+1}, \ell_k\rangle + \langle q'_{k+1} - q^{\pi^c_k}, \ell_k\rangle \\ &\le \sum_{k=1}^K \langle q_k - q'_{k+1}, \ell_k\rangle + \frac{1}{\eta}\sum_{k=1}^K \left(D_\psi(q^{\pi^c_k}, q_k) - D_\psi(q^{\pi^c_k}, q'_{k+1})\right) \\ &\le \sum_{k=1}^K \langle q_k - q'_{k+1}, \ell_k\rangle + \frac{1}{\eta}\sum_{k=1}^K \left(D_\psi(q^{\pi^c_k}, q_k) - D_\psi(q^{\pi^c_k}, q_{k+1})\right),\end{aligned} \tag{13}$$

where the first inequality holds due to Lemma 3 and the last inequality holds due to Pythagoras theorem. For the first term, applying the inequality $1 - e^{-x} \le x$, we obtain

$$\sum_{k=1}^K \langle q_k - q'_{k+1}, \ell_k\rangle \le \eta \sum_{k=1}^K \sum_{x,a} q_k(x,a)\ell_k^2(x,a) \le \eta \sum_{k=1}^K \sum_{x,a} q_k(x,a) \le \eta HK = \eta T. \tag{14}$$

For the last term, we have

$$\begin{aligned}&\sum_{k=1}^K \left(D_\psi(q^{\pi^c_k}, q_k) - D_\psi(q^{\pi^c_k}, q_{k+1})\right) \\ &= D_\psi(q^{\pi^c_1}, q_1) + \sum_{k=2}^K \left(D_\psi(q^{\pi^c_k}, q_k) - D_\psi(q^{\pi^c_{k-1}}, q_k)\right) \\ &= D_\psi(q^{\pi^c_1}, q_1) + \sum_{k=2}^K \sum_{x,a} \left(q^{\pi^c_k}(x,a)\log\frac{q^{\pi^c_k}(x,a)}{q_k(x,a)} - q^{\pi^c_{k-1}}(x,a)\log\frac{q^{\pi^c_{k-1}}(x,a)}{q_k(x,a)}\right) \\ &= D_\psi(q^{\pi^c_1}, q_1) + \sum_{k=2}^K \sum_{x,a} \left(q^{\pi^c_k}(x,a) - q^{\pi^c_{k-1}}(x,a)\right)\log\frac{1}{q_k(x,a)} + \psi(q^{\pi^c_K}) - \psi(q^{\pi^c_1}) \\ &\le \sum_{k=2}^K \|q^{\pi^c_k} - q^{\pi^c_{k-1}}\|_1 \log\frac{1}{\alpha} + D_\psi(q^{\pi^c_1}, q_1) + \psi(q^{\pi^c_K}) - \psi(q^{\pi^c_1}),\end{aligned} \tag{15}$$

where the last inequality holds due to $q_k(x,a) \ge \alpha$, since $q_k \in \Delta(M,\alpha)$ for all k and x,a. It remains to bound the last two terms. Since q_1 minimize ψ over $\Delta(M,\alpha)$, we have $\langle \nabla\psi(q_1), q^{\pi^c_1} - q_1\rangle \ge 0$, and thus

$$D_\psi(q^{\pi^c_1}, q_1) + \psi(q^{\pi^c_K}) - \psi(q^{\pi^c_1}) \le \psi(q^{\pi^c_K}) - \psi(q_1) \le \sum_{x,a} q_1(x,a)\log\frac{1}{q_1(x,a)} \le H\log\frac{|X||A|}{H}. \tag{16}$$

Combining(14), (15) and (16), we obtain

$$\sum_{k=1}^K \langle q_k - q^{\pi^c_k}, \ell_k\rangle \le \eta T + \frac{1}{\eta}\left(H\log\frac{|X||A|}{H} + \bar{P}_T \log\frac{1}{\alpha}\right),$$

where $\bar{P}_T = \sum_{k=2}^K \|q^{\pi^c_k} - q^{\pi^c_{k-1}}\|_1$. This completes the proof. □

C.3. Proof of Theorem 1

Proof. Without loss of generality, we assume that all states are reachable with positive probability under the uniform policy $\pi^u(a|x) = 1/|A|, \forall x \in X, a \in A$ (otherwise remove the unreachable states since they are unreachable by any policy). Assume T is large enough such that the occupancy measure of π^u satisfies $q^{\pi^u} \in \Delta(M, \frac{1}{T})$, then define $u_k = (1 - \frac{1}{T})q^{\pi^{\mathsf{c}}_k} + \frac{1}{T}q^{\pi^u} \in \Delta(M, \frac{1}{T^2})$, we have

$$\begin{aligned}\sum_{k=1}^{K}\langle q_k - q^{\pi^{\mathsf{c}}_k}, \ell_k\rangle &= \sum_{k=1}^{K}\langle q_k - u_k, \ell_k\rangle + \frac{1}{T}\sum_{k=1}^{K}\langle q^{\pi^u} - q^{\pi^{\mathsf{c}}_k}, \ell_k\rangle \\ &\leq \sum_{k=1}^{K}\langle q_k - u_k, \ell_k\rangle + 2 \\ &\leq \underbrace{\sum_{k=1}^{K}\langle q_k - q_{k,i}, \ell_k\rangle}_{\texttt{meta-regret}} + \underbrace{\sum_{k=1}^{K}\langle q_{k,i} - u_k, \ell_k\rangle}_{\texttt{base-regret}} + 2, \end{aligned} \tag{17}$$

where the first inequality follows from the definition $u_k = (1 - \frac{1}{T})q^{\pi^{\mathsf{c}}_k} + \frac{1}{T}q^{\pi^u}$ and the last inequality holds for any index i.

Upper bound of base-regret. Since $u_k \in \Delta(M, \frac{1}{T^2}), \forall k \in [K]$, from Lemma 1 we have

$$\texttt{base-regret} \leq \eta T + \frac{H \log\frac{|X||A|}{H} + 2\sum_{k=2}^{K}\|u_k - u_{k-1}\|_1 \log T}{\eta} \leq \eta T + \frac{H \log\frac{|X||A|}{H} + 2\bar{P}_T \log T}{\eta},$$

where the last inequality holds due to $\sum_{k=2}^{K}\|u_k - u_{k-1}\|_1 \leq \sum_{k=2}^{K}\|q^{\pi^{\mathsf{c}}_k} - q^{\pi^{\mathsf{c}}_{k-1}}\|_1 = \bar{P}_T$. It is clear that the optimal step size is $\eta^* = \sqrt{(H\log\left(|X||A|/H\right) + 2\bar{P}_T\log T)/T}$. From the definition of $\bar{P}_T = \sum_{k=2}^{K}\|q^{\pi^{\mathsf{c}}_k} - q^{\pi^{\mathsf{c}}_{k-1}}\|_1$, we have $0 \leq \bar{P}_T \leq 2KH = 2T$. Thus, the possible range of the optimal step size is

$$\eta_{\min} = \sqrt{\frac{H\log(|X||A|/H)}{T}}, \text{ and } \eta_{\max} = \sqrt{\frac{H\log(|X||A|/H)}{T} + 4\log T}.$$

By the construction of the candidate step size pool $\mathcal{H} = \{\eta_i = 2^{i-1}\sqrt{K^{-1}\log(|X||A|/H)} \mid i \in [N]\}$, where $N = \lceil\frac{1}{2}\log(1 + \frac{4K\log T}{\log(|X||A|/H)})\rceil + 1$, we know that the step size therein is monotonically increasing, in particular

$$\eta_1 = \sqrt{\frac{H\log(|X||A|/H)}{T}} = \eta_{\min}, \text{ and } \eta_N \geq \sqrt{\frac{H\log(|X||A|/H)}{T} + 4\log T} = \eta_{\max}.$$

Thus, we ensure there exists a base-learner i^* whose step size satisfies $\eta_{i^*} \leq \eta^* \leq \eta_{i^*+1} = 2\eta_{i^*}$. Since the regret decomposition in (17) holds for any $i \in [N]$, we choose the base-learner i^* to analysis to obtain a sharp bound.

$$\begin{aligned}\texttt{base-regret} &\leq \eta_{i^*}T + \frac{H\log(|X||A|/H) + 2\bar{P}_T\log T}{\eta_{i^*}} \\ &\leq \eta^* T + \frac{2(H\log(|X||A|/H) + 2\bar{P}_T\log T)}{\eta^*} \\ &= 3\sqrt{T\left(H\log(|X||A|/H) + 2\bar{P}_T\log T\right)},\end{aligned} \tag{18}$$

where the second inequality holds due to $\eta_{i^*} \leq \eta^* \leq \eta_{i^*+1} = 2\eta_{i^*}$ and the last equality holds by substituting the optimal step size $\eta^* = \sqrt{(H\log(|X||A|/H) + 2\bar{P}_T\log T)/T}$.

Upper bound of meta-regret. From the construction that $h_{k,i} = \langle q_{k,i}, \ell_t\rangle, \forall k \in [K], i \in [N]$, the meta-regret can be written in the following way:

$$\texttt{meta-regret} = \sum_{k=1}^{K}\langle q_k - q_{k,i}, \ell_k\rangle = \sum_{k=1}^{K}\langle\sum_{i=1}^{N}p_{k,i}q_{k,i} - q_{k,i}, \ell_k\rangle = \sum_{k=1}^{K}\langle p_k - e_i, h_k\rangle$$

It is known that the update $p_{k+1,i} \propto \exp(-\varepsilon\sum_{s=1}^{k}h_{s,i}), \forall i \in [N]$ is equal to the update $p_{k+1} = \arg\min_{p\in\Delta_N}\varepsilon\langle p, h_k\rangle + D_\psi(p, p_k)$, where $\psi(p) = \sum_{i=1}^{N}p_i\log p_i$ is the standard negative entropy. This can be further reformulated solving the unconstrained optimization problem $p'_{k+1} = \arg\min_p \varepsilon\langle p, h_k\rangle + D_\psi(p, p_k)$ at first and then projecting p'_{k+1} to the simplex Δ_N as $p_{k+1} = \arg\min_{p\in\Delta_N} D_\psi(p, p'_{k+1})$. By standard analysis of OMD, we have

$$\begin{aligned}\sum_{k=1}^{K}\langle p_k - e_i, h_k\rangle &\le \sum_{k=1}^{K}\langle p_k - p'_{k+1}, h_k\rangle + \sum_{k=1}^{K}\langle p'_{k+1} - e_i, h_k\rangle\\ &\le \sum_{k=1}^{K}\langle p_k - p'_{k+1}, h_k\rangle + \frac{1}{\varepsilon}\sum_{k=1}^{K}(D_\psi(e_i, p_k) - D_\psi(e_i, p'_{k+1}))\\ &\le \sum_{k=1}^{K}\langle p_k - p'_{k+1}, h_k\rangle + \frac{1}{\varepsilon}\sum_{k=1}^{K}(D_\psi(e_i, p_k) - D_\psi(e_i, p_{k+1}))\\ &\le \sum_{k=1}^{K}\langle p_k - p'_{k+1}, h_k\rangle + \frac{1}{\varepsilon}D_\psi(e_i, p_1),\end{aligned}$$

where the second inequality holds due to Lemma 3 and the third inequality holds due to Pythagoras theorem. Using the fact that $1 - e^{-x} \le x$ and the definition that $p_{1,i} = 1/N, h_{k,i} \le H, \forall k \in [K], i \in [N]$, we have

$$\sum_{k=1}^{K}\langle p_k - p'_{k+1}, h_k\rangle + \frac{1}{\varepsilon}D_\psi(e_i, p_1) \le \varepsilon\sum_{k=1}^{K}\sum_{i=1}^{N} p_{k,i}h_{k,i}^2 + \frac{\ln N}{\varepsilon} \le \varepsilon HT + \frac{\ln N}{\varepsilon}.$$

Therefore, for any base-learner $i \in [N]$, we have

$$\sum_{k=1}^{K}\langle q_k - q_{k,i}, \ell_k\rangle = \sum_{k=1}^{K}\langle p_k - e_i, h_k\rangle \le \varepsilon HT + \frac{\log N}{\varepsilon}.$$

In particular, for the best base-learner $i^* \in [N]$, we have

$$\texttt{meta-regret} = \sum_{k=1}^{K}\langle q_k - q_{k,i^*}, \ell_k\rangle \le \varepsilon HT + \frac{\log N}{\varepsilon} = \sqrt{HT\log N}, \tag{19}$$

where the last equality holds due to the setting $\varepsilon = \sqrt{(\log N)/(HT)}$.

Upper bound of overall dynamic regret. Combining (17), (18) and (19), we obtain

$$\begin{aligned}\sum_{k=1}^{K}\langle q_k - q^{\pi_k^c}, \ell_k\rangle &\le \texttt{base-regret} + \texttt{meta-regret}\\ &\le 3\sqrt{T\left(H\log(|X||A|/H) + 2\bar{P}_T\log T\right)} + \sqrt{HT\log N} + 2\\ &\le \mathcal{O}\left(\sqrt{HT\left(\log(|X||A|/H) + P_T\log T\right)}\right),\end{aligned}$$

where the last equality is due to $\bar{P}_T \le HP_T$ by Lemma 6, $N = \lceil\frac{1}{2}\log(1 + \frac{4K\log T}{\log(|X||A|/H)})\rceil + 1$. This finishes the proof. □

C.4. Proof of Theorem 2

Proof. The proof is similar to the proof of the minimax lower bound of dynamic regret for online convex optimization (Zhang et al., 2018). For any $\gamma \in [0, 2T]$, we first construct a piecewise-stationary comparator sequence, whose path-length is smaller than γ, then we split the whole time horizon into several pieces, where the comparator is fixed in each piece. By this construction, we can apply the existed minimax static regret lower bound of episodic loop-free SSP (Zimin & Neu, 2013) in each piece, and finally sum over all pieces to obtain the lower bound for the dynamic regret.

Denote by $R_K(\Pi, \mathcal{F}, \gamma)$ the minimax dynamic regret, which is defined as

$$R_K(\Pi, \mathcal{F}, \gamma) = \inf_{\pi_1\in\Pi}\sup_{\ell_1\in\mathcal{F}}\ldots\inf_{\pi_K\in\Pi}\sup_{\ell_K\in\mathcal{F}}\left(\sum_{k=1}^{K}\langle q^{\pi_k}, \ell_k\rangle - \min_{(\pi_1^c,\ldots,\pi_K^c)\in\mathcal{U}(\gamma)}\sum_{k=1}^{K}\langle q^{\pi_k^c}, \ell_k\rangle\right)$$

where Π denotes the set of all policies, $\mathcal{F}$ denotes the set of loss functions $\ell \in \mathbb{R}_{[0,1]}^{|X||A|}$, and $\mathcal{U}(\gamma) = \{(\pi_1^{\mathsf{c}}, \ldots, \pi_K^{\mathsf{c}}) \mid \forall k \in [K], \pi_k^{\mathsf{c}} \in \Pi, \text{ and } \bar{P}_T = \sum_{k=2}^{K} \|q^{\pi_k^{\mathsf{c}}} - q^{\pi_{k-1}^{\mathsf{c}}}\|_1 \leq \gamma\}$ is the set of feasible policy sequences with the path-length $\bar{P}_T$ of the occupancy measures less than γ.

We first consider the case of $\gamma \leq 2H$. Then we can directly utilize the established lower bound of the static regret for learning in episodic loop-free SSP (Zimin & Neu, 2013) as a natural lower bound of dynamic regret,

$$R_K(\Pi, \mathcal{F}, \gamma) \geq C_1 H \sqrt{K \log(|X||A|)}, \tag{20}$$

where $C_1 = 0.03$ is the constant appeared in the lower bound of static regret.

We next deal with the case that $\gamma \geq 2H$. Without loss of generality, we assume $L = \lceil \gamma/2H \rceil$ divides K and split the whole time horizon into L pieces equally. Next, we construct a special policy sequence in $\mathcal{U}(\gamma)$ 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 variation of the occupancy measure at each change point is at most $2H$, the path-length $\bar{P}_T$ of the occupancy measures does not exceed γ. As a result, we have

$$R_K(\Pi, \mathcal{F}, \gamma) \geq LC_1 H \sqrt{\frac{K}{L} \log(|X||A|)} \geq \frac{\sqrt{2}C_1}{2} \sqrt{HK\gamma \log(|X||A|)}. \tag{21}$$

Combining (20) and (21), we obtain the final lower bound

$$R_K(\Pi, \mathcal{F}, \gamma) \geq \frac{\sqrt{2}C_1}{2} \sqrt{HK \log(|X||A|)} \max(\sqrt{2H}, \sqrt{\gamma}) \geq \Omega(\sqrt{HK(H+\gamma) \log(|X||A|)}),$$

which finishes the proof. □

C.5. Useful Lemmas

In this part, we present some basic lemmas in episodic loop-free SSP. For any policy $\pi(a|x)$, we define P^π to be the transition matrix induced by π, where the component $[P^\pi]_{x,x'}$ is the transition probability from x to x', i.e., $[P^\pi]_{x,x'} = \sum_a \pi(a|x) P_{x,x'}^a$. Then, we have the following useful lemmas.

Lemma 7 (Lemma 6.3 of Even-Dar et al. (2009))**.** *For any policies π and π and any state distribution d, we have*

$$\|dP^\pi - dP^{\pi'}\|_1 \leq \|\pi - \pi'\|_{1,\infty}.$$

Proof. Consider the case when d is a delta function on x. The difference in the next state distributions, $\|dP^\pi - dP^{\pi'}\|_1$, is

$$\begin{aligned}\sum_{x'} \left| [P^\pi]_{x,x'} - [P^{\pi'}]_{x,x'} \right| &= \sum_{x'} \sum_a |P(x'|x,a)\,(\pi(a|x) - \pi'(a|x))| \\ &\leq \sum_{x',a} P(x'|x,a) |\pi(a|x) - \pi'(a|x)| = \sum_a |\pi(a|x) - \pi'(a|x)|.\end{aligned}$$

Linearity of expectation leads to the result for arbitrary d. □

Lemma 8. *For any state distribution d and d', and any policy π, we have*

$$\|dP^\pi - d'P^\pi\|_1 \leq \|d - d'\|_1. \tag{22}$$

Proof. Note that the relationship that $d(x') = \sum_x d(x) P_{x,x'}^\pi$, therefore, we have

$$\begin{aligned}\|dP^\pi - d'P^\pi\|_1 &= \sum_{x'} |\sum_x d(x) P_{x,x'}^\pi - d'(x) P_{x,x'}^\pi| \leq \sum_{x'} \sum_x |d(x) P_{x,x'}^\pi - d'(x) P_{x,x'}^\pi| \\ &= \sum_{x'} \sum_x |d(x) - d'(x)| P_{x,x'}^\pi = \sum_x |d(x) - d'(x)| \sum_{x'} P_{x,x'}^\pi \\ &= \sum_x |d(x) - d'(x)| = \|d - d'\|_1.\end{aligned}$$

This finishes the proof. □

Finally, we introduce the following lemma, which shows the strongly convexity of the regularizer.

Lemma 9. *$\psi(w) = \sum_{i=1}^d w_i \log w_i$ is $\frac{1}{H}$-strongly convex w.r.t. $\|\cdot\|_1$ for $\{w \in \mathbb{R}^d_{\geq 0} \mid \sum_{i=1}^d w_i = H\}$.*

Proof. For any $y, z \in \{w \in \mathbb{R}^d_{\geq 0} \mid \sum_{i=1}^N w_i = H\}$, we have $\frac{y}{H}, \frac{z}{H} \in \{w \in \mathbb{R}^d_{\geq 0} \mid \sum_{i=1}^d w_i = 1\}$. Then, it holds that

$$\psi(y) - \psi(z) - \langle \nabla\psi(y), y - z\rangle = \sum_{i=1}^d y_i \log\frac{y_i}{z_i} = H\sum_{i=1}^d \frac{y_i}{H}\log\frac{y_i/H}{z_i/H} \geq \frac{1}{2H}\|y-z\|_1^2,$$

where the last inequality holds due to Pinsker's Inequality. This finishes the proof. □

D. Proofs for Section 3 (Episodic SSP)

In this section, we first give the impossible result to bound the path-length of occupancy measures by the path-length of policies. Next we provide proofs of the dynamic regret of CODO-REPS algorithm and the lower bound in Section 3.2.

D.1. Path-length of Policies and Occupancy Measures

In the following, we give the impossible result to bound the path-length of the occupancy measures by the path-length of the corresponding policies.

Theorem 7. *For any $H_* > 1$ and any positive integer $c > 0$, there exists an SSP instance with $|X| = 2c+1$ states, $|A| = 2$ actions and a policy sequence $\pi_1^{\mathsf{c}}, \ldots, \pi_K^{\mathsf{c}}$ with largest expected hitting time H_* such that $\bar{P}_K \geq cP_K$.*

Proof. For any $H_* > 1$ and any positive integer $c > 0$, we construct an episodic SSP with $n+1$ states $X = \{x_0, \ldots, x_n\}$ with $n = 2c$ and two actions $A = \{a_1, a_2\}$ as in Figure 1. Let the transition kernel be deterministic and the corresponding transitions are shown in Figure 1. Specifically, taking a_1 and a_2 in initial state x_0 leads to the state g and x_1 respectively.

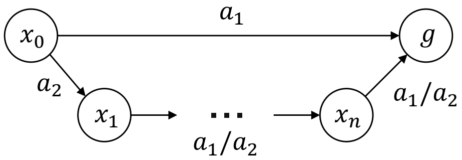

Figure 1: State transitions for Theorem 7.

Taking any action in state x_i leads to state $x_{i+1}, \forall i \in [n-1]$ and taking any action in state x_n leads to the goal state state g. Then, we consider two policies π and π' with $\pi(a_1|x_i) = 1, \forall i \in \{0\} \cup [n]$ and $\pi'(a_1|x_0) = 1-\varepsilon, \pi'(a_1|x_i) = 1, \forall i \in [n]$. It is clear that $\|\pi - \pi'\|_{1,\infty} = 2\varepsilon$ and $H^\pi(x_0) = 1, H^{\pi'}(x_0) = 1 + \varepsilon n$. For any $H_* > 1$ and $c > 0$, let $\varepsilon = (H_* - 1)/n$, we have $H^{\pi'} = 1 + \varepsilon n = H_*$, i.e., the largest hitting time of π and π' is H_*. Then we consider the occupancy measure discrepancy of π and π'. It is easy to verify

$$\sum_{x,a} |q^\pi(x,a) - q^{\pi'}(x,a)| = \varepsilon + \varepsilon(n+1) = \varepsilon(n+2) = 2\varepsilon(c+1) = (c+1)\|\pi - \pi'\|_{1,\infty}.$$

Therefore, we have $\|q^\pi - q^{\pi'}\|_1 \geq c\|\pi - \pi'\|_{1,\infty}$. Thus, the policy sequence $\pi, \pi', \pi, \pi', \ldots$ satisfies $\bar{P}_K = K\|q^\pi - q^{\pi'}\|_1 \geq cK\|\pi - \pi'\|_{1,\infty} = cP_K$, which completes the proof. □

D.2. Proof of Lemma 2

Proof. Since $\eta \leq \frac{1}{64}$, we ensure that $32\eta|\ell_{k,i}| \leq 1, \forall k \in [K], i \in [|X||A|]$. Taking $m_k = 0$ in Lemma 5, we obtain

$$\sum_{k=1}^K \langle q_k - q^{\pi_k^{\mathsf{c}}}, \ell_k\rangle \leq \sum_{k=1}^K \left(D_\psi(q^{\pi_k^{\mathsf{c}}}, \widehat{q}_k) - D_\psi(q^{\pi_k^{\mathsf{c}}}, \widehat{q}_{k+1})\right) + 32\eta\sum_{k=1}^K \langle q_k, \ell_k^2\rangle - 16\eta\sum_{k=1}^K \langle q^{\pi_k^{\mathsf{c}}}, \ell_k^2\rangle. \tag{23}$$

For the first term, from the definition that $D_\psi(q,q') = \sum_{x,a} q(x,a)\log\frac{q(x,a)}{q'(x,a)} - \sum_{x,a}(q(x,a) - q'(x,a))$, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\Big(D_\psi(q^{\pi^{\mathsf{c}}_k}, \widehat{q}_k) - D_\psi(q^{\pi^{\mathsf{c}}_k}, \widehat{q}_{k+1})\Big) \qquad (24)\\
&= D_\psi(q^{\pi^{\mathsf{c}}_1}, \widehat{q}_1) + \sum_{k=2}^{K}\Big(D_\psi(q^{\pi^{\mathsf{c}}_k}, \widehat{q}_k) - D_\psi(q^{\pi^{\mathsf{c}}_{k-1}}, \widehat{q}_k)\Big)\\
&= D_\psi(q^{\pi^{\mathsf{c}}_1}, \widehat{q}_1) + \frac{1}{\eta}\sum_{k=2}^{K}\sum_{x,a}\left(q^{\pi^{\mathsf{c}}_k}(x,a)\log\frac{q^{\pi^{\mathsf{c}}_k}(x,a)}{\widehat{q}_k(x,a)} - q^{\pi^{\mathsf{c}}_{k-1}}(x,a)\log\frac{q^{\pi^{\mathsf{c}}_{k-1}}(x,a)}{\widehat{q}_k(x,a)}\right) + \frac{1}{\eta}\sum_{k=2}^{K}\sum_{x,a}\Big(q^{\pi^{\mathsf{c}}_{k-1}}(x,a) - q^{\pi^{\mathsf{c}}_k}(x,a)\Big)\\
&= D_\psi(q^{\pi^{\mathsf{c}}_1}, \widehat{q}_1) + \frac{1}{\eta}\sum_{k=2}^{K}\sum_{x,a}\Big(q^{\pi^{\mathsf{c}}_k}(x,a) - q^{\pi^{\mathsf{c}}_{k-1}}(x,a)\Big)\log\frac{1}{\widehat{q}_k(x,a)} + \frac{\psi(q^{\pi^{\mathsf{c}}_K}) - \psi(q^{\pi^{\mathsf{c}}_1}) - \sum_{x,a}(q^{\pi^{\mathsf{c}}_K}(x,a) - q^{\pi^{\mathsf{c}}_1}(x,a))}{\eta}\\
&\le \frac{1}{\eta}\log\frac{H}{\alpha}\sum_{k=2}^{K}\|q^{\pi^{\mathsf{c}}_k} - q^{\pi^{\mathsf{c}}_{k-1}}\|_1 + D_\psi(q^{\pi^{\mathsf{c}}_1}, \widehat{q}_1) + \frac{\psi(q^{\pi^{\mathsf{c}}_K}) - \psi(q^{\pi^{\mathsf{c}}_1}) - \sum_{x,a}(q^{\pi^{\mathsf{c}}_K}(x,a) - q^{\pi^{\mathsf{c}}_1}(x,a))}{\eta},
\end{aligned}$$

where the last inequality holds due to $|\log\widehat{q}_k(x,a)| \le \log\frac{H}{\alpha}$ since $\alpha \le \widehat{q}_k(x,a) \le H$ for $\widehat{q}_k \in \Delta(M,H,\alpha)$. For the last two term, since $\widehat{q}_1$ minimize ψ over $\Delta(M,H,\alpha)$, we have $\langle\nabla\psi(\widehat{q}_1), q^{\pi^{\mathsf{c}}_1} - \widehat{q}_1\rangle \le 0$, thus

$$\begin{aligned}
&D_\psi(q^{\pi^{\mathsf{c}}_1}, \widehat{q}_1) + \frac{\psi(q^{\pi^{\mathsf{c}}_K}) - \psi(q^{\pi^{\mathsf{c}}_1}) - \sum_{x,a}(q^{\pi^{\mathsf{c}}_K}(x,a) - q^{\pi^{\mathsf{c}}_1}(x,a))}{\eta}\\
&\le \frac{\psi(q^{\pi^{\mathsf{c}}_1}) - \psi(\widehat{q}_1)}{\eta} + \frac{\psi(q^{\pi^{\mathsf{c}}_K}) - \psi(q^{\pi^{\mathsf{c}}_1}) - \sum_{x,a}\big(q^{\pi^{\mathsf{c}}_K}(x,a) - q^{\pi^{\mathsf{c}}_1}(x,a)\big)}{\eta}\\
&\le \frac{\psi(q^{\pi^{\mathsf{c}}_K}) - \psi(\widehat{q}_1) - \sum_{x,a}\big(q^{\pi^{\mathsf{c}}_K}(x,a) - q^{\pi^{\mathsf{c}}_1}(x,a)\big)}{\eta} \qquad (25)\\
&\le \frac{H\log(|X||A|) + H\log H + H}{\eta} = \frac{H(1+\log(|X||A|H))}{\eta},
\end{aligned}$$

where the last inequality holds due to $-H\log(|X||A|) \le \psi(q) \le H\log H$ and $0 \le \sum_{x,a} q(x,a) \le H$ for any $q \in \Delta(M,H,\alpha)$ from Lemma 10. Combining (23), (24) and (25), we have

$$\sum_{k=1}^{K}\langle q_k - q^{\pi^{\mathsf{c}}_k}, \ell_k\rangle \le \frac{H(1+\log(|X||A|H)) + \bar{P}_K\log(H/\alpha)}{\eta} + 32\eta\sum_{k=1}^{K}\langle q_k, \ell_k^2\rangle - 16\eta\sum_{k=1}^{K}\langle q^{\pi^{\mathsf{c}}_k}, \ell_k^2\rangle,$$

where $\bar{P}_K = \sum_{k=2}^{K}\|q^{\pi^{\mathsf{c}}_k} - q^{\pi^{\mathsf{c}}_{k-1}}\|_1$. This finishes the proof. □

D.3. Proof of Theorem 3

Proof. We only need to consider the case $H_* \le K$ (otherwise the claimed regret bound is vacuous). Since all the compared policies are proper, they will not visit the states from which the goal state g is not accessible (otherwise the hitting time will be infinite) and the states which are not accessible from initial state x_0. We can remove them from the SSP since we consider the known transition setting. Then, suppose K is large enough such that these exists at least a policy π^u whose occupancy measure q^{π^u} satisfies $q^{\pi^u} \in \Delta(M, K, \frac{1}{K})$. Then, we define $u_k = (1-\frac{1}{K^2})q^{\pi^{\mathsf{c}}_k} + \frac{1}{K^2}q^{\pi^u}$ and the corresponding policy π^{u_k}. For any $k \in [K]$, we ensure that the hitting time $H^{\pi^{u_k}} \le (1-\frac{1}{K^2})H_* + \frac{K}{K^2} \le H_* + 1$ and the occupancy measure $u_k(x,a) \ge \frac{1}{K^3}, \forall x, a$, i.e., $u_k \in \Delta(M, H_*+1, \frac{1}{K^3})$. Thus, we have

$$\begin{aligned}
\mathbb{E}[\text{D-R}\textsc{eg}_K(\pi^{\mathsf{c}}_{1:K})] &= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{i,j} p_k^{i,j}\langle q_k^{i,j}, \ell_k\rangle - \sum_{k=1}^{K}\langle q^{\pi^{\mathsf{c}}_k}, \ell_k\rangle\right]\\
&= \mathbb{E}\left[\sum_{k=1}^{K}\sum_{i,j} p_k^{i,j}\langle q_k^{i,j}, \ell_k\rangle - \sum_{k=1}^{K}\langle u_k, \ell_k\rangle\right] + \frac{1}{K^2}\mathbb{E}\left[\sum_{k=1}^{K}\langle q^{\pi^u} - q^{\pi^{\mathsf{c}}_k}, \ell_k\rangle\right]
\end{aligned}$$

$$\begin{aligned}
&\le \mathbb{E}\left[\sum_{k=1}^{K}\sum_{i,j} p_k^{i,j}\langle q_k^{i,j}, \ell_k\rangle - \sum_{k=1}^{K}\langle u_k, \ell_k\rangle\right] + 2,\\
&\le \underbrace{\mathbb{E}\left[\sum_{k=1}^{K}\langle p_k - e_{i,j}, h_k\rangle\right]}_{\texttt{meta-regret}} + \underbrace{\mathbb{E}\left[\sum_{k=1}^{K}\langle q_k^{i,j} - u_k, \ell_k\rangle\right]}_{\texttt{base-regret}} + 2 \qquad (26)
\end{aligned}$$

where the first inequality holds due to $\sum_{x,a} q^u(x,a) \le K$ and $\sum_{x,a} q^{\pi_k^{\mathrm{c}}}(x,a) \le H_* \le K$, the last inequality holds due to the definition that $h_k^{i,j} = \langle q_k^{i,j}, \ell_k\rangle, \forall i \in [G], j \in [N_i]$ and the decomposition holds for any index $i \in [G], j \in [N_i]$.

Upper bound of base-regret. Since the possible range of H_* is $H^{\pi^f} \le H_* \le K$. From the construction of horizon length pool $\mathcal{H} = \{H_i = 2^{i-1} \cdot H^{\pi^f} | i \in [G]\}$ where $G = 1 + \lceil \log((K+1)/H^{\pi^f})\rceil$, we ensure

$$H_1 = H^{\pi^f} \le H_* + 1 \text{ and } H_G = K + 1 \ge H_* + 1.$$

So for any unknown H_*, there exist an index i for the space pool that satisfies $H_{i^*-1} = \frac{H_{i^*}}{2} \le H_* + 1 \le H_{i^*}$. Then, we analysis the base-regret of the base learners in group i^*. From the construction of each step size pool, we ensure $\eta_{i,j} \le \frac{1}{64}$, i.e., $32\eta_{i,j}|\ell_{k,r}| \le 1, \forall i \in [G], j \in [N_i], k \in [K], r \in [|X||A|]$. Since $q_k^{i^*,j}, u_k \in \Delta(M, H_{i^*}, 1/K^3), \forall j \in [N_i^*], k \in [K]$, from Lemma 2, we have

$$\begin{aligned}
\texttt{base-regret} &\le \frac{4\sum_{k=2}^{K}\|u_k - u_{k-1}\|_1 \log K + H_{i^*}(1 + \log(|X||A|H_{i^*}))}{\eta_{i^*,j}} + 16\eta_{i^*,j}\sum_{k=1}^{K}(2\langle u_k, \ell_k\rangle - \langle q_k^{i^*,j}, \ell_k^2\rangle)\\
&\le \frac{4\bar{P}_K \log K + H_{i^*}(1 + \log(|X||A|H_{i^*}))}{\eta_{i^*,j}} + 32\eta_{i^*,j}B_K - 16\eta_{i^*,j}\sum_{k=1}^{K}\langle q_k^{i^*,j}, \ell_k^2\rangle + 1, \qquad (27)
\end{aligned}$$

where the last inequality holds due to $\sum_{k=2}^{K}\|u_k - u_{k-1}\|_1 \le \sum_{k=2}^{K}\|q^{\pi_k^{\mathrm{c}}} - q^{\pi_{k-1}^{\mathrm{c}}}\|_1 = \bar{P}_K$ and $B_K = \sum_{k=1}^{K}\langle q^{\pi_k^{\mathrm{c}}}, \ell_k\rangle$.

Upper bound of meta-regret. Then, we consider the meta-regret with respect to base-learner $\mathcal{B}_{i^*,j}, \forall j \in N_{i^*}$. From the construction of the regularizer $\bar{\psi}(p)$ in meta-algorithm, we have $32\varepsilon_{i,j}|h_k^{i,j}| = 32\frac{\eta_{i,j}}{2H_i}|\langle q_k^{i,j}, \ell_k\rangle| \le 1, \forall i \in [G], j \in [N_i], k \in [K]$. From the analysis of OMD in Lemma 5, we have

$$\begin{aligned}
\texttt{meta-regret} &\le D_{\bar{\psi}}(e_{i^*,j}, p_1) + 32\varepsilon_{i^*,j}\sum_{k=1}^{K}(h_k^{i^*,j})^2\\
&= \frac{1}{\varepsilon_{i^*,j}}\log\frac{1}{p_1^{i^*,j}} + \sum_{r=1}^{G}\sum_{s=1}^{N_i}\frac{p_1^{r,s}}{\varepsilon_{r,s}} + 32\varepsilon_{i^*,j}\sum_{k=1}^{K}(h_k^{i^*,j})^2\\
&= \frac{1}{\varepsilon_{i^*,j}}\log\frac{\sum_{r=1}^{G}\sum_{s=1}^{N_i}\varepsilon_{r,s}^2}{\varepsilon_{i^*,j}^2} + \frac{\sum_{r=1}^{G}\sum_{s=1}^{N_i}\varepsilon_{r,s}}{\sum_{r=1}^{G}\sum_{s=1}^{N_i}\varepsilon_{r,s}^2} + 32\varepsilon_{i^*,j}\sum_{k=1}^{K}\langle q_k^{i^*,j}, \ell_k\rangle^2,
\end{aligned}$$

where the first equality holds due to $D_{\bar{\psi}}(p, p') = \sum_{i,j}\frac{1}{\varepsilon_{i,j}}(p_{i,j}\log\frac{p_{i,j}}{p'_{i,j}} - p_{i,j} + p'_{i,j})$ and the last equality is due to $p_1^{i,j} \propto \varepsilon_{i,j}^2, h_k^{i,j} = \langle q_k^{i,j}, \ell_k\rangle, \forall i \in [G], j \in [N_i]$. From the definition of the horizon length pool $\mathcal{H} = \{H_i = 2^{i-1} \cdot H^{\pi^f} \mid i \in [G]\}$ where $G = 1 + \lceil \log((K+1)/H^{\pi^f})\rceil$, the step size pools $\mathcal{E}_i = \left\{\frac{1}{32 \cdot 2^j} \mid j \in [N_i]\right\}, i \in [G]$, where $N_i = \lceil \frac{1}{2}\log\left(\frac{4K}{1 + \log(|X||A|H_i)}\right)\rceil$ and learning rate $\varepsilon_{i,j} = \frac{\eta_{i,j}}{2H_i}, \forall i \in [G], j \in [N_i]$, we ensure that $\sum_{r=1}^{G}\sum_{s=1}^{N_i}\varepsilon_{r,s} = \Theta(1/H_1)$ and $\sum_{r=1}^{G}\sum_{s=1}^{N_i}\varepsilon_{r,s}^2 = \Theta(1/H_1^2)$. Thus,

$$\begin{aligned}
\texttt{meta-regret} &\le \Theta\left(\frac{H_{i^*}}{\eta_{i^*,j}}\log\frac{H_{i^*}}{H_1\eta_{i^*,j}}\right) + 16\frac{\eta_{i^*,j}}{H_{i^*}}\sum_{k=1}^{K}\langle q_k^{i^*,j}, \ell_k\rangle^2 + \Theta(H_1)\\
&\le \Theta\left(\frac{H_{i^*}}{\eta_{i^*,j}}\log\frac{H_{i^*}}{H_1\eta_{i^*,j}}\right) + 16\eta_{i^*,j}\sum_{k=1}^{K}\langle q_k^{i^*,j}, \ell_k^2\rangle + \Theta(H_1),
\end{aligned} \qquad (28)$$

where the last inequality holds due to Cauchy–Schwarz inequality.

Upper bound of over all dynamic regret. Combining (26), (27) and (28), we obtain

$$\mathbb{E}[\text{D-Reg}_K] \leq \frac{4\bar{P}_T \log K + H_{i^*}(1+\log(|X||A|H_{i^*}))}{\eta_{i^*,j}} + 32\eta_{i^*,j}B_K + \Theta(\frac{H_{i^*}}{\eta_{i^*,j}} \log \frac{H_{i^*}}{H_1\eta_{i^*,j}}), \tag{29}$$

holds for any index $j \in [N_{i^*}]$. Omit the last term, it is clear that the optimal step size is $\eta^* = \sqrt{(H_{i^*}(1+\log(|X||A|H_{i^*})) + 4\bar{P}_K \log K)/(32B_K)}$. Meanwhile, since $\sum_{x,a} q_k(x,a) \leq H_*, \forall k \in [K]$, we have $0 \leq \bar{P}_K = \sum_{k=2}^K \|q^{\pi_k^c} - q^{\pi_{k-1}^c}\|_1 \leq 2H_*K \leq 2H_{i^*}K$ and $B_K \leq H_*K \leq H_{i^*}K$. Therefore, we ensure that

$$\eta^* \geq \sqrt{\frac{1+\log(|X||A|H_{i^*})}{32K}}.$$

From the construction of the candidate step size pool H_{i^*}, we know that the step size therein is monotonically decreasing with respect to the index, in particular,

$$\eta_1 = \frac{1}{64}, \text{ and } \eta_N = \sqrt{\frac{1+\log(|X||A|H_{i^*})}{128K}} \leq \eta^*$$

Let j^* be the index of base learner in group i^* with step size closest to the η^*. Then, we consider the base regret of the base learner B_{i^*,j^*}. We consider the following two cases:

- when $\eta^* \leq \frac{1}{64}$, then $\eta_{i^*,j^*} \leq \eta^* \leq 2\eta_{i^*,j^*} = \eta_{i^*,j^*-1}$, we have

$$\begin{aligned}\text{R.H.S of (29)} &\leq \frac{8\bar{P}_K \log K + 2H_{i^*}(1+\log(|X||A|H_{i^*}))}{\eta^*} + 32\eta^* B_K + \Theta\left(\frac{H_{i^*}}{\eta^*}\log\frac{H_{i^*}}{H_1\eta^*}\right)\\ &\leq \widetilde{\mathcal{O}}\left(\sqrt{\left(\bar{P}_K + H_*\right)B_K}\right).\end{aligned}$$

- when $\eta^* > \frac{1}{64}$, then $\eta_{i^*,j^*} = \frac{1}{64}$, we have

$$\text{R.H.S of (29)} \leq 256\left(\bar{P}_K \log K + H_{i^*}(1+\log(|X||A|H_{i^*}))\right) + \frac{1}{2}B_K + \Theta(H_*) \leq \widetilde{\mathcal{O}}\left(\bar{P}_K + H_*\right),$$

  where the last inequality holds due to $\sqrt{(H_{i^*}(1+\log(|X||A|H_{i^*})) + 4\bar{P}_K \log K)/(32B_K)} \geq \frac{1}{64}$.

As a result, taking both cases into account yields

$$\sum_{k=1}^K \langle q_k - q^{\pi_k^c}, \ell_k\rangle \leq \widetilde{\mathcal{O}}\left(\sqrt{\left(H_* + \bar{P}_K\right)\left(H_* + \bar{P}_K + B_K\right)}\right).$$

This finishes the proof. □

D.4. Proof of Theorem 4

Proof. The proof is similar to that of Theorem 2. For any $\gamma \in [0, 2T]$, we first construct a piecewise-stationary comparator sequence, whose path-length is smaller than γ, then we split the whole time horizon into several pieces, where the comparator is fixed in each piece. By this construction, we can apply the existed minimax static regret lower bound of episodic SSP (Chen et al., 2021a) in each piece, and finally sum over all pieces to obtain the lower bound for the dynamic regret.

Denote by $R_K(\Pi, \mathcal{F}, \gamma)$ the minimax dynamic regret, which is defined as

$$R_K(\Pi, \mathcal{F}, \gamma) = \inf_{\pi_1 \in \Pi} \sup_{\ell_1 \in \mathcal{F}} \ldots \inf_{\pi_K \in \Pi} \sup_{\ell_K \in \mathcal{F}} \left(\sum_{k=1}^K \langle q^{\pi_k}, \ell_k\rangle - \min_{(\pi_1^c,\ldots,\pi_K^c)\in\mathcal{U}(\gamma)} \sum_{k=1}^K \langle q^{\pi_k^c}, \ell_k\rangle\right)$$

where Π denotes the set of all policies, $\mathcal{F}$ denotes the set of loss functions $\ell \in R_{[0,1]}^{|X||A|}$ and $\mathcal{U}(\gamma) = \{(\pi_1^c, \ldots, \pi_K^c) \mid \forall k \in [K], \pi_k^c \in \Pi, \text{ and } \bar{P}_K = \sum_{k=2}^K \|q^{\pi_k^c} - q^{\pi_{k-1}^c}\|_1 \leq \gamma\}$ is the set of feasible policy sequences with path-length $\bar{P}_K$ of the occupancy measures less than γ.

We first consider the case of $\gamma \le 2(H_* + 1)$. From Theorem 3 of Chen et al. (2021a), we ensure for any D, H_*, K with $K \ge D + 1$, there exists an SSP instance such that its diameter is $D + 2$, the hitting time of the best fixed policy is $H_* + 1$ and the expected regret of any policy after K episodes is at least $\Omega(\sqrt{DH_*K})$. Then we can set all compared policies as the best fixed policy and directly utilize this lower bound of the static regret as a natural lower bound of dynamic regret,

$$R_K(\Pi, \mathcal{F}, \gamma) \ge \Omega(\sqrt{DH_*K}). \tag{30}$$

We next deal with the case that $\gamma \ge 2(H_* + 1)$. Without loss of generality, we assume $L = \lceil \gamma/2(H_* + 1) \rceil$ devides K and split the whole time horizon into L pieces equally. Next, we construct an SSP instance such that its diameter is $D + 2$, the hitting time of the best fixed policy is $H_* + 1$ and the expected regret of any policy after K episodes is at least $\Omega(\sqrt{DH_*K})$ in each piece. Then, we choose the best fixed policy in each piece as the comparator sequence, whose hitting time are all $H_* + 1$. Since the sequence changes at most $L - 1 \le \gamma/2(H_* + 1)$ times and the variation of the policy sequence at each change point is at most $2(H_* + 1)$ (Note that $\|q^{\pi^{\mathsf{c}}_k} - q^{\pi^{\mathsf{c}}_{k-1}}\| \le \|q^{\pi^{\mathsf{c}}_k}\|_1 + \|q^{\pi^{\mathsf{c}}_{k-1}}\|_1 = 2(H_* + 1), \forall \pi^{\mathsf{c}}_k \neq \pi^{\mathsf{c}}_{k-1}$), the path-$\bar{P}_K$ does not exceed γ. As a result,

$$R_K(\Pi, \mathcal{F}, \gamma) \ge L\Omega(\sqrt{DH_*K/L}) \ge \Omega(\sqrt{DK\gamma}). \tag{31}$$

Combining (30) and (31), we have

$$R_K(\Pi, \mathcal{F}, \gamma) \ge \Omega(\sqrt{DH_*K}) + \Omega(\sqrt{DK\gamma}) \ge \Omega(\sqrt{DH_*K(1 + \gamma/H_*)}),$$

which finishes the proof. □

D.5. Useful Lemmas

we introduce the following lemma which shows the boundedness of the regularizer.

Lemma 10. *Let $H \ge 1$, it holds that $-H\log(|X||A|) \le \sum_{x,a} q(x,a)\log q(x,a) \le H\log H$ for every $q \in \Delta(M, H)$.*

Proof. First, we prove the right-hand side of the inequality.

$$\sum_{x,a} q(x,a)\log q(x,a) = \sum_{x,a} q(x,a)\log\frac{q(x,a)}{H} + \sum_{x,a} q(x,a)\log H \le \sum_{s,a} q(x,a)\log H \le H\log H.$$

Then, we prove the left-hand side of the inequality.

$$-\sum_{x,a} q(x,a)\log q(x,a) = -\sum_{x,a} q(x,a)\log\frac{q(x,a)}{H} - \sum_{x,a} q(x,a)\log H \le -H\sum_{s,a}\frac{q(x,a)}{H}\log\frac{q(x,a)}{H} \le H\log|X||A|.$$

This finishes the proof. □

E. Proofs for Section 4 (Infinite-horizon MDPs)

In this section, we first show the relationship between the path-length of policies and the path-length of occupancy measures. Next, we show the proofs of the reduction to switching-cost expert problem in Section 4.2. Finally, we give the proofs of the dynamic regret of our algorithm in Section 4.3.

E.1. Path-length of Policies and Occupancy Measures

We introduce the relationship between the path-length of policies and the path-length of occupancy measures as follows.

Lemma 11. *For any occupancy measure sequence $q_1, \ldots, q_T$ induced by the policy sequence $\pi_1, \ldots, \pi_T$, it holds that*

$$\sum_{t=2}^{T} \|q^{\pi_t} - q^{\pi_{t-1}}\|_1 \le (\tau + 2)\sum_{t=2}^{T} \|\pi_t - \pi_{t-1}\|_{1,\infty}.$$

Proof. Consider any two policies π and π' with occupancy measure q^{π} and $q^{\pi'}$, let $d^{\pi}(x) \triangleq \sum_{x,a} q^{\pi}(x,a), d^{\pi'}(x) \triangleq \sum_{x,a} q^{\pi'}(x,a), \forall x \in X$, we have

$$\begin{aligned}
\|q^{\pi} - q^{\pi'}\|_1 &= \sum_{x,a} |q^{\pi}(x,a) - q^{\pi'}(x,a)| \\
&= \sum_{x,a} |d^{\pi}(x)\pi(a|x) - d^{\pi'}(x)\pi'(a|x)| \\
&\leq \sum_{x,a} |d^{\pi}(x)\pi(a|x) - d^{\pi}(x)\pi'(a|x)| + \sum_{x,a} |d^{\pi}(x)\pi'(a|x) - d^{\pi'}(x)\pi'(a|x)| \\
&= \sum_{x} d^{\pi}(x) \sum_{a} |\pi(a|x) - \pi'(a|x)| + \sum_{x} |d^{\pi}(x) - d'(x)| \sum_{a} \pi'(a|x) \\
&\leq \|\pi - \pi'\|_{1,\infty} + \|d^{\pi} - d^{\pi'}\|_1 \\
&\leq (\tau + 2)\|\pi - \pi'\|_{1,\infty},
\end{aligned}$$

where the first inequality holds due to the triangle inequality and the last inequality holds due to Lemma 14. We finish the proof by summing the inequality over T. □

E.2. Proof of Theorem 5

To prove Theorem 5, we first introduce two lemmas which measure the difference between the sum of average losses and the actual losses of the learner and compared policies. Denote by ρ_t^{π} the *average loss per step* corresponding π: $\rho_t^{\pi} \triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\ell_t(x_t, a_t)|P, \pi] = \langle q^{\pi}, \ell_t \rangle$ and the actual cumulative loss suffered by the learner $L_T \triangleq \mathbb{E}[\ell_t(x_t, \pi_t(x_t))|P, \pi]$, where the randomness is over the transition kernel and policy sequence $\pi_{1:T}$. Similarly, the actual cumulative loss suffered by the compared policy sequence $\pi_{1:T}^{\mathsf{c}}$ is $L_T^{\mathsf{c}} \triangleq \mathbb{E}[\ell_t(x_t, \pi_t^{\mathsf{c}}(x_t))|P, \pi]$. Let d^{π} be the stationary state distribution, i.e., $d^{\pi}(x) \triangleq \sum_a q^{\pi}(x,a), \forall x \in X$. Denote by $\mu_t = \mu_1 P^{\pi_1} \cdots P^{\pi_{t-1}}$ the state distribution after executing $\pi_1, \ldots, \pi_{t-1}$, where μ_1 is the initial distribution, similarly, $\mu_t^{\mathsf{c}} = \mu_1 P^{\pi_1^{\mathsf{c}}} \cdots P^{\pi_{t-1}^{\mathsf{c}}}$.

Lemma 12. *For any compared policy sequence $\pi_1^{\mathsf{c}}, \ldots, \pi_T^{\mathsf{c}}$, it holds that $\sum_{t=1}^{T} \rho_t^{\pi_t^{\mathsf{c}}} - L_T^{\mathsf{c}} \leq (\tau+1)^2 P_T + 2(\tau+1)$.*

Proof. From the definition that $\mu_t^{\mathsf{c}} = \mu_1 P^{\pi_1^{\mathsf{c}}} \cdots P^{\pi_{t-1}^{\mathsf{c}}}$, we have

$$\begin{aligned}
\sum_{t=1}^{T} \rho_t^{\pi_t^{\mathsf{c}}} - L_T^{\mathsf{c}} &= \sum_{t=1}^{T} \sum_{x} \left(d^{\pi_t^{\mathsf{c}}}(x) - \mu_t^{\mathsf{c}}(x) \right) \sum_{a} \pi_t^{\mathsf{c}}(a|x) \ell_t(x,a) \\
&\leq \sum_{t=1}^{T} \|d^{\pi_t^{\mathsf{c}}} - \mu_t^{\mathsf{c}}\|_1 \\
&\leq 2(\tau+1) + (\tau+1) \sum_{t=2}^{T} \|d^{\pi_t^{\mathsf{c}}} - d^{\pi_{t-1}^{\mathsf{c}}}\|_1 \\
&\leq 2(\tau+1) + (\tau+1)^2 \sum_{t=2}^{T} \|\pi_t^{\mathsf{c}} - \pi_{t-1}^{\mathsf{c}}\|_{1,\infty},
\end{aligned}$$

where the second inequality holds due to Lemma 15 and the last inequality holds due to Lemma 14. □

Lemma 13. *For any occupancy measure sequence $q^{\pi_1}, \ldots, q^{\pi_T}$ returned by the learner, it holds that $L_T - \sum_{t=1}^{T} \rho_t^{\pi_t} \leq (\tau+1) \sum_{t=2}^{T} \|q^{\pi_t} - q^{\pi_{t-1}}\|_1 + 2(\tau+1)$.*

Proof. From the definition that $\mu_t = \mu_1 P^{\pi_1} \cdots P^{\pi_{t-1}}$, we have

$$L_T - \sum_{t=1}^{T} \rho_t^{\pi_t} = \sum_{t=1}^{T} \sum_{x} (\mu_t(x) - d^{\pi_t}(x)) \sum_{a} \pi_t(a|x) \ell_t(x,a)$$

$$
\begin{aligned}
&\le \sum_{t=1}^{T} \|\mu_t - d^{\pi_t}\|_1 \\
&\le 2(\tau+1) + (\tau+1)\sum_{t=2}^{T} \|d^{\pi_t} - d^{\pi_{t-1}}\|_1 \\
&= 2(\tau+1) + (\tau+1)\sum_{t=2}^{T}\sum_{x} \Big|\sum_{a} q^{\pi_t}(x,a) - q^{\pi_{t-1}}(x,a)\Big| \\
&\le 2(\tau+1) + (\tau+1)\sum_{t=2}^{T} \|q^{\pi_t} - q^{\pi_{t-1}}\|_1,
\end{aligned}
$$

where the second inequality holds due to Lemma 15. □

Then, we present the proof of Theorem 5.

Proof of Theorem 5. Note that the dynamic regret for infinite-horizon MDPs is defined as $\mathbb{E}[\text{D-}\textsc{Reg}(\pi^{\mathsf{c}}_{1:T})] = \mathbb{E}[\sum_{t=1}^{T} \ell_t(x_t, \pi_t(x_t)) - \ell_t(x_t, \pi^{\mathsf{c}}_t(x_t))]$. Then it can be written as

$$
\begin{aligned}
\mathbb{E}[\text{D-}\textsc{Reg}_T(\pi^{\mathsf{c}}_{1:T})] &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(x_t, \pi_t(x_t)) - \ell_t(x_t, \pi^{\mathsf{c}}_t(x_t))\right] \\
&= L_T - \sum_{t=1}^{T} \rho_t^{\pi_t} + \sum_{t=1}^{T} (\rho_t^{\pi_t} - \rho_t^{\pi^{\mathsf{c}}_t}) + \sum_{t=1}^{T} \rho_t^{\pi^{\mathsf{c}}_t} - L_T^{\mathsf{c}} \\
&\le \sum_{t=1}^{T} \langle q_t - q^{\pi^{\mathsf{c}}_t}, \ell_t\rangle + \sum_{t=2}^{T} \|q_t - q_{t-1}\|_1 + (\tau+1)^2 P_T + 4(\tau+1),
\end{aligned}
$$

where the last inequality holds due to Lemma 12 and Lemma 13 and the definition that $P_T = \sum_{t=2}^{T} \|\pi_t - \pi^{\mathsf{c}}_t\|_{1,\infty}$. □

E.3. Proof of Theorem 6

Proof. Similar to the proof in Appendix C.3, since the MDP is ergodic according to Definition 1, we assume T is large enough such that there at least exists a policy π^u whose occupancy measure q^u satisfies $q^{\pi^u} \in \Delta(M, \frac{1}{T})$, then define $u_t = (1 - \frac{1}{T}) q^{\pi^{\mathsf{c}}_t} + \frac{1}{T} q^{\pi^u} \in \Delta(M, \frac{1}{T^2})$, from the dynamic regret decomposition in (9), we have

$$
\begin{aligned}
&\mathbb{E}[\text{D-}\textsc{Reg}_T(\pi^{\mathsf{c}}_{1:T})] \qquad (32) \\
\le& \sum_{t=1}^{T} \langle q_t - q^{\pi^{\mathsf{c}}_t}, \ell_t\rangle + (\tau+1)\sum_{t=2}^{T} \|q_t - q_{t-1}\|_1 + (\tau+1)^2 P_T + 4(\tau+1) \\
=& \sum_{t=1}^{T} \langle q_t - u_t, \ell_t\rangle + \frac{1}{T}\sum_{t=1}^{T} \langle q^{\pi^u} - q^{\pi^{\mathsf{c}}_t}, \ell_t\rangle + (\tau+1)\sum_{t=2}^{T} \|q_t - q_{t-1}\|_1 + (\tau+1)^2 P_T + 4(\tau+1) \\
=& \underbrace{\sum_{t=1}^{T} \langle q_t - u_t, \ell_t\rangle + (\tau+1)\sum_{t=2}^{T} \|q_t - q_{t-1}\|_1}_{\texttt{term (a)}} + (\tau+1)^2 P_T + 4(\tau+1) + 2, \qquad (33)
\end{aligned}
$$

where the first equality follows from the definition that $u_t = (1-\frac{1}{T}) q^{\pi^{\mathsf{c}}_t} + \frac{1}{T} q^{\pi^u}$. We only need to consider $\texttt{term (a)}$ since the remaining terms are not related to the algorithm. From the definition that $h_{t,i} = \langle q_{t,i}, \ell_t\rangle + (\tau+1)\|q_{t,i} - q_{t-1,i}\|_1, \forall i \in [N]$, it can be verified that $\texttt{term (a)}$ can be written as

$$
\underbrace{\sum_{t=1}^{T} \langle p_t, h_t\rangle - \sum_{t=1}^{T} h_{t,i}}_{\texttt{meta-regret}} + \underbrace{(\tau+1)\sum_{t=2}^{T} \|p_t - p_{t-1}\|_1}_{\texttt{meta-switching-cost}} + \underbrace{\sum_{t=1}^{T} \langle q_{t,i} - u_t, \ell_t\rangle}_{\texttt{base-regret}} + \underbrace{(\tau+1)\sum_{t=2}^{T} \|q_{t,i} - q_{t-1,i}\|_1}_{\texttt{base-switching-cost}},
$$

which hold for any index i. Next, we bound these terms separately.

Upper bound of base-regret. From the standard analysis of OMD similar to that in (13) and (15), we have

$$\begin{aligned}\sum_{t=1}^{T}\langle q_{t,i}-u_t,\ell_t\rangle &\le \eta_i \sum_{t=1}^{T}\sum_{x,a} q_{t,i}(x,a)\ell_t^2(x,a) + \frac{1}{\eta_i}\sum_{t=1}^{T}\left(D_\psi(u_t,q_{t,i}) - D_\psi(u_t,q_{t+1,i})\right)\\ &\le \eta_i T + \frac{\log|X||A|}{\eta_i} + \frac{2\log T}{\eta_i}\sum_{t=2}^{T}\|u_t-u_{t-1}\|_1 \\ &\le \eta_i T + \frac{\log|X||A| + 2\bar{P}_T\log T}{\eta_i},\end{aligned} \tag{34}$$

which the last inequality holds due to $\sum_{t=2}^{T}\|u_t-u_{t-1}\|_1 \le \sum_{t=2}^{T}\|q^{\pi_t^{\mathsf{c}}} - q^{\pi_{t-1}^{\mathsf{c}}}\|_1 = \bar{P}_T$.

Upper bound of meta-regret. From the definition that $h_{t,i} = \langle q_{t,i},\ell_t\rangle + (\tau+1)\|q_{t,i}-q_{t-1,i}\|_1, \forall i\in[N]$, we have $0 \le h_{t,i} \le 1 + 2(\tau+1) = 2\tau+3, \forall i \in [N]$. By the standard analysis of Hedge similar to the analysis of meta-regret in Appendix C.3, we have

$$\sum_{t=1}^{T}\langle p_t,h_t\rangle - \sum_{t=1}^{T}h_{t,i} \le \varepsilon\sum_{t=1}^{T}\sum_{i=1}^{N}p_{t,i}h_{t,i}^2 + \frac{\log N}{\varepsilon} \le \varepsilon(2\tau+3)^2T + \frac{\log N}{\varepsilon}. \tag{35}$$

Upper bound of switching-cost. From Lemma 3, we have

$$\|q_{t,i}-q_{t-1,i}\|_1 \le \eta_i\|\ell_t\|_\infty \le \eta_i, \text{ and } \|p_t-p_{t-1}\|_1 \le \varepsilon\|h_t\|_\infty \le \varepsilon(2\tau+3), \forall t\ge 2. \tag{36}$$

Upper bound of overall dynamic regret. Combining (34), (35) and (36), we obtain

$$\begin{aligned}\texttt{term (a)} &\le \eta_i(\tau+2)T + \frac{\log(|X||A|)+2\bar{P}_T\log T}{\eta_i} + \varepsilon(2\tau+3)^2T + \frac{\log N}{\varepsilon} + \varepsilon(2\tau+3)(\tau+1)T\\ &\le \eta_i(\tau+2)T + \frac{\log(|X||A|)+2\bar{P}_T\log T}{\eta_i} + 2\varepsilon(2\tau+3)^2T + \frac{\log N}{\varepsilon}.\end{aligned}$$

From the configuration that $\varepsilon = \sqrt{\frac{\log N}{2T(2\tau+3)^2}}$, we obtain

$$\texttt{term (a)} \le \eta_i(\tau+2)T + \frac{\log(|X||A|)+2\bar{P}_T\log T}{\eta_i} + (4\tau+6)\sqrt{2T\log N}.$$

It is clear that the the optimal step size is $\eta^* = \sqrt{\frac{\log|X||A|+2\bar{P}_T\log T}{(\tau+2)T}}$. From the definition of $\bar{P}_T$, we have $0 \le \bar{P}_T = \sum_{t=2}^{T}\|q^{\pi_t^{\mathsf{c}}} - q^{\pi_{t-1}^{\mathsf{c}}}\|_1 \le 2T$, we ensure the possible range of η^* is

$$\sqrt{\frac{\log|X||A|}{(\tau+2)T}} \le \eta^* \le \sqrt{\frac{\log(|X||A|)+4T\log T}{(\tau+2)T}}.$$

Set the step size pool as $\mathcal{H} = \left\{2^{i-1}\sqrt{\frac{\log|X||A|}{T}} \mid i\in[N]\right\}$ where $N = \lceil\frac{1}{2}\log(1+\frac{4T\log T}{\log|X||A|})\rceil + 1$. We can verify that

$$\eta_1 = \sqrt{\frac{\log|X||A|}{(\tau+2)T}} \le \eta^*, \text{ and } \eta_N \ge \sqrt{\frac{\log(|X||A|)+4T\log T}{(\tau+2)T}} = \eta^*.$$

Thus, we confirm that there exists a base-learner whose step size satisfies $\eta_{i^*} \le \eta^* \le \eta_{i^*+1} = 2\eta_{i^*}$. Then, we choose i^* to analysis to obtain a sharp bound. Thus $\texttt{term (a)}$ is bounded by

$$\texttt{term (a)} \le \eta_{i^*}(\tau+2)T + \frac{\log(|X||A|)+2\bar{P}_T\log T}{\eta_{i^*}} + (4\tau+6)\sqrt{2T\log N}$$

$$\leq \eta^*(\tau+2)T + \frac{2(\log(|X||A|) + 2\bar{P}_T \log T)}{\eta^*} + (4\tau+6)\sqrt{2T\log N}$$
$$\leq 3\sqrt{(\tau+2)T(\log|X||A| + 2\bar{P}_T \log T)} + (4\tau+6)\sqrt{2T\log N}, \tag{37}$$

where the second inequality holds due to the condition $\eta_{i^*} \leq \eta^* \leq \eta_{i^*+1} = 2\eta_{i^*}$ and the last inequality holds by substituting the optimal step size $\eta^* = \sqrt{\frac{\log|X||A|+2\bar{P}_T \log T}{(\tau+2)T}}$. Therefore, combining (33) and (37), we obtain

$$\begin{aligned}
&\mathbb{E}[\text{D-REG}_T(\pi^{\mathtt{c}}_{1:T})]\\
&\leq \mathtt{term\ (a)} + (\tau+1)^2 P_T + 4\tau + 6\\
&\leq 3\sqrt{(\tau+2)T(\log|X||A| + 2\bar{P}_T\log T)} + (4\tau+6)\sqrt{2T\log N} + (\tau+1)^2 P_T + 4\tau + 6\\
&\leq 3\sqrt{(\tau+2)T(\log|X||A| + 2(\tau+2)P_T\log T)} + (4\tau+6)\sqrt{2T\log N} + (\tau+1)^2 P_T + 4\tau + 6\\
&\leq \mathcal{O}(\sqrt{\tau T(\log|X||A| + \tau P_T \log T)} + \tau^2 P_T),
\end{aligned}$$

where the third inequality uses $\bar{P}_T \leq (\tau+2)P_T$ from Lemma 11. This finishes the proof. □

E.4. Useful Lemmas

In this part, we present some useful lemmas in infinite-horizon MDPs. Denote by d^π the stationary state distribution induced by policy π under transition kernel P, i.e., $d^\pi(x) \triangleq \sum_a q^\pi(x,a), \forall x \in X$. Then we have the following useful lemmas which show the relationships between policy discrepancy and distribution discrepancy.

Lemma 14 (Lemma 4 of Neu et al. (2014))**.** *For any two policies π and π', it holds that*

$$\|d^\pi - d^{\pi'}\|_1 \leq (\tau+1)\|\pi - \pi'\|_{1,\infty}.$$

Proof. The statement follows from solving

$$\|d^\pi - d^{\pi'}\|_1 \leq \|d^\pi P^\pi - d^{\pi'} P^\pi\|_1 + \|d^{\pi'} P^\pi - d^{\pi'} P^{\pi'}\|_1 \leq e^{-1/\tau}\|d^\pi - d^{\pi'}\|_1 + \|\pi - \pi'\|_{1,\infty}$$

for $\|d^\pi - d^{\pi'}\|_1$ and using $\frac{1}{1-e^{-1/\tau}} \leq \tau + 1$. □

Lemma 15. *Consider the distribution $\mu_t = \mu_1 P^{\pi_1} \cdot P^{\pi_{t-1}}$, where μ_1 is any distribution over X and $\pi_1, \ldots, \pi_t$ is any policy sequence, it holds that*

$$\sum_{t=1}^T \|\mu_t - d^{\pi_t}\|_1 \leq 2(\tau+1) + (\tau+1)\sum_{t=2}^T \|d^{\pi_t} - d^{\pi_{t-1}}\|_1.$$

Proof. It is trivial for $t = 1$ since $\|\mu_1 - d^{\pi_1}\|_1 \leq 2$. Thus, in what follows we only consider the case that $t \geq 2$.By the triangle inequality, we have

$$\begin{aligned}
\|\mu_t - d^{\pi_t}\|_1 &\leq \|\mu_t - d^{\pi_{t-1}}\|_1 + \|d^{\pi_{t-1}} - d^{\pi_t}\|_1\\
&= \|\mu_{t-1}P^{\pi_{t-1}} - d^{\pi_{t-1}}P^{\pi_{t-1}}\|_1 + \|d^{\pi_{t-1}} - d^{\pi_t}\|_1\\
&\leq e^{-1/\tau}\|\mu_{t-1} - d^{\pi_{t-1}}\|_1 + \|d^{\pi_{t-1}} - d^{\pi_t}\|_1\\
&\leq e^{-1/\tau}\left(e^{-1/\tau}\|\mu_{t-2} - d^{\pi_{t-2}}\|_1 + \|d^{\pi_{t-2}} - d^{\pi_{t-1}}\|_1\right) + \|d^{\pi_{t-1}} - d^{\pi_t}\|_1\\
&\leq \cdots \leq e^{-(t-1)/\tau}\|\mu_1 - d^{\pi_1}\|_1 + \sum_{n=0}^{t-2} e^{-n/\tau}\|d^{\pi_{t-n}} - d^{\pi_{t-n-1}}\|_1\\
&\leq 2e^{-(t-1)/\tau} + \sum_{n=0}^{t-2} e^{-n/\tau}\|d^{\pi_{t-n}} - d^{\pi_{t-n-1}}\|_1.
\end{aligned}$$

where the first equality holds since $d^{\pi_{t-1}}$ is the stationary distribution of π_{t-1}, i.e., $d^{\pi_{t-1}} = d^{\pi_{t-1}} P^{\pi_{t-1}}$ and the second inequality holds due to Definition 1. Summing above inequality over t, we have

$$\begin{aligned}\sum_{t=1}^{T} \|\mu_t - d^{\pi_t}\|_1 &\leq 2 + 2\sum_{t=2}^{T} e^{-(t-1)/\tau} + \sum_{t=2}^{T}\sum_{n=0}^{t-2} e^{-n/\tau} \|d^{\pi_{t-n}} - d^{\pi_{t-n-1}}\|_1 \\ &\leq 2(\tau+1) + \sum_{t=2}^{T}(\sum_{n=0}^{T-t} e^{-n/\tau}) \|d^{\pi_t} - d^{\pi_{t-1}}\|_1 \\ &\leq 2(\tau+1) + (\tau+1)\sum_{t=2}^{T} \|d^{\pi_t} - d^{\pi_{t-1}}\|_1.\end{aligned}$$

This finishes the proof. □