# Slowly Changing Adversarial Bandit Algorithms are Efficient for Discounted MDPs

**Ian A. Kash** IANKASH@UIC.EDU
*Computer Science, University of Illinois at Chicago, Chicago, Illinois, USA*

**Lev Reyzin** LREYZIN@UIC.EDU
*Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, Illinois, USA*

**Zishun Yu** ZYU32@UIC.EDU
*Computer Science, University of Illinois at Chicago, Chicago, Illinois, USA*

**Editors:** Claire Vernade and Daniel Hsu

## Abstract

Reinforcement learning generalizes multi-armed bandit problems with additional difficulties of a longer planning horizon and unknown transition kernel. We explore a black-box reduction from discounted infinite-horizon tabular reinforcement learning to multi-armed bandits, where, specifically, an *independent* bandit learner is placed in each state. We show that, under ergodicity and fast mixing assumptions, any *slowly changing* adversarial bandit algorithm achieving optimal regret in the adversarial bandit setting can also attain optimal expected regret in infinite-horizon discounted Markov decision processes, with respect to the number of rounds $T$. Furthermore, we examine our reduction using a specific instance of the exponential-weight algorithm.

**Keywords:** Multi-armed bandits, reinforcement learning, discounted Markov decision processes, black-box reduction.

## 1. Introduction

Reinforcement learning (RL) and multi-armed bandits (MAB) are long-standing models for decision-making problems. RL generalizes bandits with a long-term planning horizon and unknown transition dynamics. Due to these additional complexities, RL is typically viewed as a more challenging problem compared to MAB. However, there is a large literature (Kearns and Singh, 2002; Osband et al., 2013; Dann et al., 2017; Osband and Van Roy, 2017; Agrawal and Jia, 2017; Fruit et al., 2018a; Jin et al., 2018; Dann et al., 2019; Simchowitz and Jamieson, 2019; Russo, 2019; Zhang and Ji, 2019; Zhang et al., 2020; Cai et al., 2020; Zhang et al., 2020; Neu and Pike-Burke, 2020; Pacchiano et al., 2021; Ménard et al., 2021; Li et al., 2021; Zhang et al., 2021) that guarantees RL can achieve the optimal regret $\Omega(\sqrt{T})$ in the dependency of the number of rounds $T$, and is often optimal in terms of the cardinalities, $S$ and $A$, of state and action spaces. Recent episodic horizon-free works (Wang et al., 2020a; Zhang et al., 2021, 2022; Li and Yang, 2023) further show the potential to close the formal complexity gap between RL and bandits, with RL's regret approaching the lower bound of the (contextual) MAB problem $\Omega(\sqrt{SAT})$ (Bubeck et al., 2013; Auer et al., 1995; Gerchinovitz and Lattimore, 2016). These findings imply that the longer planning horizon and unknown transition kernels in RL may not introduce additional difficulties compared to bandits.

We therefore ask: *Is there a reduction from tabular reinforcement learning to multi-armed bandits?* Specifically, in a *decentralized* setting, could one place an *independent* bandit learner in each state (referred as local learners), such that this set of local learners achieves sub-linear regret in

MDPs collectively, without needing to acquire information (for example value estimations) from their co-learners, except for the shared global rewards?

We answer this question positively for discounted infinite-horizon MDPs. We prove that, under ergodicity and fast mixing assumptions, one could trivially place $\tilde{\mathcal{O}}(S)$[1] arbitrary *slowly changing* bandit algorithms to achieve a regret bound of $\tilde{\mathcal{O}}(\text{poly}(S, A, H, \tau, \frac{1}{\beta}, \frac{1}{1-\gamma}) \cdot (\sqrt{T} + c_T T))$ (which depends on various problem parameters specified in later sections), if the bandit learners are optimal in the adversarial bandit setting. Here, $c_T$ represents the changing rate for the chosen bandit algorithm. The regret bound is optimal with respect to $T$ (up to polylogarithmic factors) when $c_T$ is $\tilde{\mathcal{O}}(1/\sqrt{T})$, which is a mild requirement as discussed in later sections.

Despite the decentralized framework where each state is managed by an independent learner being a compelling problem in itself, the decoupling from the temporal difference framework makes it straightforward to leverage techniques from the bandit toolbox. For instance, in Section 5 we show how our reduction framework effectively handles delayed feedback, benefiting from the robustness of adversarial bandits to such feedback. This also opens up possibilities for straightforward translation of existing bandit results, such as delayed or aggregated feedback (Joulani et al., 2013; Pike-Burke et al., 2018), to MDPs, especially since these settings are gaining traction in RL as well (Howson et al., 2021; Jin et al., 2022b; Mondal and Aggarwal, 2023). In addition, understanding the reduction to independent learners can be connected to multi-agent RL, where such decentralization allows mitigating the curse of multiagency (Jin et al., 2022a; Cravic et al., 2023), and can be also bridged to Monte Carlo methods, as detailed in Section 2.

## 2. Related Work

The work most closely related to ours is perhaps that of Cheng et al. (2020a), who propose a reduction from RL to continuous online learning (Cheng et al., 2020b) under a generative oracle setting. This setting allows algorithms to query transitions from the true dynamics without interacting with the environment. In addition to the generative model requirement, our work is significantly different from theirs in the sense that their work considers centralized no-regret learners, communicating through the value function estimations while ours is decentralized.

On the other hand, diverging from the canonical temporal difference scheme makes the Monte Carlo evaluation a natural choice for our reduction, as detailed in Section 4. This positions our work within the realm of Monte Carlo methods: for example, Monte Carlo Exploring Starts (MCES) (Sutton and Barto, 2018). Similar to MCES, our reduction associates each state with an independent decision-maker using Monte Carlo estimations. The primary difference lies in the exploration technique: MCES uses exploring starts[2], whereas in our reduction, exploration is partly delegated to the bandit learners. Despite being considered as "one of the most fundamental open theoretical questions in reinforcement learning" (Sutton and Barto, 2018), MCES had relatively few guarantees until recent works on its convergence (Wang et al., 2021; Liu, 2021; Dong et al., 2022; Winnicki and Srikant, 2023), while an earlier result by Tsitsiklis (2002) requires more restrictive assumptions.

In terms of implementations and technical tools, our reduction aligns more closely with research in the area of online MDPs (Even-Dar et al., 2009; Neu et al., 2010; Rosenberg and Mansour, 2019; Jin et al., 2020). Specifically, Even-Dar et al. (2009) implemented a framework where each state is managed by an expert algorithm, while Neu et al. (2010) proposed a model with a bandit learner

1. $\tilde{\mathcal{O}}(\cdot)$ compresses polylog dependencies.
2. Exploring starts sample an initial $(s_0, a_0)$ randomly for each episode, ensuring all $(s, a)$ are visited infinitely often.

assigned to each state. Yet, in the work on online MDPs, policy evaluation is still done in a temporal difference fashion, which differs from our reduction. The "slowly changing" property required by our reduction, is also an important insight from these works. One can analyze slowly changing policies with their stationary distributions which are generally easier to handle, see Lemma 11 for details. But in general our analysis is still very different because of our decentralized setup. We further leveraged the slowly changing property to give our results in Section 5.2 and Section 5.3 to address the corresponding difficulties raised by such decentralization.

**Additional Related Works.** We consider infinite-horizon discounted MDPs, akin to the setting considered by the reduction in Cheng et al. (2020a). We would like to note that one could often translate the results from infinite horizon setting to episodic setting but not vice versa (Ortner, 2020; Wang et al., 2020b), because of infinite planning horizon and lack of restarting mechanism. In contrast to episodic MDPs listed in Section 1, the study in the area of infinite-horizon discounted MDPs (Wang et al., 2020b; Liu and Su, 2020; He et al., 2021; Zhou et al., 2021; Yang et al., 2021; Yan et al., 2023, etc.) is relatively limited. In terms of planning horizon, in addition we have the category of infinite horizon average reward setting (Auer et al., 2008; Ouyang et al., 2017; Talebi and Maillard, 2018; Fruit et al., 2018b; Ortner, 2020; Dewanto et al., 2020; Wei et al., 2021; Zhang and Xie, 2023, etc.). While the majority of works discussed above are measured by regret, the line of works (Kakade, 2003; Strehl et al., 2006; Strehl and Littman, 2008; Kolter and Ng, 2009; Bartlett and Tewari, 2009; Szita and Szepesvári, 2010; Lattimore and Hutter, 2012; Lattimore et al., 2013; Dann and Brunskill, 2015; Modi et al., 2020; Xu et al., 2020, etc.) that established with *sample complexity of exploration* (Kakade, 2003) is also a major direction.

## 3. Preliminaries

### 3.1. Discounted Infinite Horizon MDPs

A tabular MDP $\mathcal{M}$ is often described by a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, respectively. We denote their cardinality by $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$. Let $\Delta(X)$ be all probability distributions over space $X$, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the *unknown* stochastic transition function, $r : \mathcal{S} \times \mathcal{A} \to \Delta([0, 1])$ is the *unknown* reward function, and $\gamma \in [0, 1)$ is a discount factor.
**Policy.** A policy is a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. In our case, a policy $\pi_t$, at time $t$, is collectively determined by the set of bandit learners, as each learner determines the strategy for a state $\pi_t(\cdot|s)$, see Algorithm 1 for details.
**Value Functions.** Given a policy $\pi$, the state value function $V^\pi(s)$ and state-action value function $Q^\pi(s, a)$ are defined as,

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t | \pi, S_0 = s\right], \quad Q^\pi(s, a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t | \pi, S_0 = s, A_0 = a\right].$$

**Optimality.** The optimal policy $\pi^* := \arg\max_\pi V^\pi(s)$, for all $s \in \mathcal{S}$. $V^*, Q^*$ denote value functions corresponding to $\pi^*$.
**State Distributions.** The state distribution $\nu$ at $t + 1$ is recursively characterized by $\nu_{t+1} := \nu_t \mathbb{P}^{\pi_t}$, where $\mathbb{P}^\pi$ is the transition kernel induced by $\pi$ and we denote $\nu_1$ as the initial distribution. The stationary distribution $\mu$ of a policy $\pi$ is the left eigenvector of $\mathbb{P}^\pi$, i.e. $\mu^\pi \mathbb{P}^\pi = \mu^\pi$.

For brevity, we use $Q_t$, $V_t$, $\mu_t$ to denote $Q^{\pi_t}$, $V^{\pi_t}$, $\mu^{\pi_t}$, respectively.

### 3.2. Regret

An obstacle to address is the different languages used in bandits and infinite-horizon discounted RL literature. While bandits community often measures algorithms' performance by regret, the community of infinite-horizon RL often uses the *sample complexity of exploration* (Kakade, 2003) (sample complexity in short). These two notions are often not translatable to each other, as regret measures the quantity of cumulative sub-optimailities but sample complexity counts the number of sub-optimailities that violate a threshold $\epsilon$. In addition to the difference between cumulative sub-optimality value vs. number of sub-optimailities, the sample complexity is not a function of the total number of rounds $T$. Two $T$-step optimal MDP learners, in the sense that they reach the optimal policy in $T$ steps, could be considered equally "good" in terms of sample complexity during those initial $T$ steps, but they could show significant differences in terms of regret measures.

To the end of a black-box reduction, we align the performance measures by leveraging a recent regret definition for discounted infinite-horizon MDPs, used by Liu and Su (2020); He et al. (2021); Zhou et al. (2021), which measures the cumulative sub-optimality $V^*(s_t) - V_t(s_t)$ that defined by the state value function.

**Definition 1** *Regret for infinite-horizon discounted MDPs*

$$\mathfrak{Regret}(T) := \sum_{t=1}^{T} \Big[ V^*(s_t) - V_t(s_t) \Big].$$

While this regret and sample complexity are not directly comparable (for example, a policy with fewer, yet larger suboptimalities may have worse regret but better sample complexity, or vice versa), bounds on sample complexity can however imply upper bounds on regret. He et al. (2021) shows that a sample complexity bound of $\mathcal{O}(M\epsilon^{-\alpha})$ implies a maximum regret of $\mathcal{O}(M^{1/(\alpha+1)}(1-\gamma)^{-1/(\alpha+1)}T^{\alpha/\alpha+1})$. This suggests that, for instance, a $\mathcal{O}(\epsilon^{-2})$ sample complexity implies a worst-case regret of $\mathcal{O}(T^{2/3})$. Although quantifying the tightness of this approximate translation is challenging, it offers a general sense of the regret notion's strength. Further insights into the comparison between sample complexity and this regret notion are discussed in Liu and Su (2020).

### 3.3. Assumptions

We make two additional assumptions.

**Assumption 1** *The stationary distributions are uniformly bounded away from zero.*

$$\inf_{\pi,s} \mu^{\pi}(s) \geq \beta \ \text{ for some } \beta > 0.$$

**Assumption 2** *There exists some fixed positive $\tau$ such that for any two arbitrary distributions $d$ and $d'$ over $\mathcal{S}$,*

$$\sup_{\pi} ||(d - d')\mathbb{P}^{\pi}||_1 \leq e^{-1/\tau}||d - d'||_1,$$

*where $\tau$ is the mixing time, we further assume $\tau \geq 1$ without loss of generality.*

Assumption 2 bounds the mixing time, of Markov chain induced by some policy $\pi$, by $\tau$. It also implies the existence and uniqueness of stationary distribution $\mu^{\pi}$. These assumptions combined

guarantee the MDP is "well behaved" in the sense that all states are likely to be visited often, thus ensuring frequent updates for each bandit learner, regardless of the policy and starting point. This is essential as an "out-dated" bandit would potentially hurt the overall performance. In addition, our assumptions play a similar role to the exploring starts in MCES, as it ensures exploration over $\mathcal{S}$, akin to the exploration over $\mathcal{S} \times \mathcal{A}$ provided by exploring starts. These assumptions have been used in prior work on online learning in MDPs such as that of Neu et al. (2010); Rosenberg and Mansour (2019), and the latter is also made in literature of stochastic games such as Etesami (2022).[3] For scenarios without these assumptions, a counter-example is provided in Appendix A.

### 3.4. Slowly Changing Algorithms

Our main result requires that the bandits placed in each state are slowly changing, for which we now provide a formal definition. To measure the change rate of an algorithm, we first introduce the 1-$\infty$ norm. For a "conditional matrix" $\mathbf{M}(y|x)$, it is defined as $||\mathbf{M}||_{1,\infty} := \max_x \sum_y |\mathbf{M}(y|x)|$, which can be used to measure the difference between two policies $||\pi - \pi'||_{1,\infty} = \max_s \sum_a [\pi(a|s) - \pi'(a|s)]$.

**Definition 2 (Slowly Changing)** *An algorithm $\mathfrak{A}$ is slowly changing with a (non-increasing) rate of $c_T$ if, for all $t$, $||\pi_{t+1} - \pi_t||_{1,\infty} \leq c_T$, where $\pi_t$ is the policy produced by $\mathfrak{A}$ at time $t$.*

Note that, throughout this paper, we assume the number of rounds $T$ is known. When $T$ is unknown, it can be managed using the standard doubling trick, see Shalev-Shwartz et al. (2012) for example.

Our analysis relies on using bandits in our algorithm that themselves are slowly changing. This slowly changing definition also applies to bandit algorithms, as one could consider the state space of bandit learners as a singleton $\{s\}$. The assumption of slowly changing bandits is mild and has been used in prior works on online learning in MDPs, such as Even-Dar et al. (2009); Neu et al. (2010). For completeness, we prove in Section 6 that EXP3 (Auer et al., 2002) is slowly changing in this respect, a fact also observed and indirectly used by Neu et al. (2010).

## 4. A Black-Box Algorithm

We now present our framework in Algorithm 1, which is based on a slowly changing bandit algorithm, referred to as LOCAL. Accordingly, Algorithm 1 is named MAIN. The key idea of our reduction is to deploy an instance of LOCAL in each state, thereby determining the strategy for that particular state.

Furthermore, we require that this bandit algorithm can accommodate delayed feedback. Robustness to delays allows us to wait to provide feedback to the algorithm, until a time such that the difference between the return at that time and the return of the full trajectory is sufficiently small, ensuring the return estimation is sufficiently accurate for the corresponding action pulled. We discuss how delayed feedback can be addressed in a black-box fashion in Section 5.4. In addition, as bandits may be updated over the course of the trajectory, the slowly changing property guarantees these changes have only a "small" effect on the expected return. Combined, these properties ensure that error in the feedback used to update the bandit, relative to the true value, is manageable.

3. Our setting is akin to cooperative games to certain extend, in the sense that local learners aimed to maximize shared global payoff without knowing the strategy of its co-learners.

**Algorithm 1:** (MAIN) Bandits for MDPs

**Require:** $\gamma \in [0,1), T, H := \left\lceil \log_\gamma \frac{1-\gamma}{\sqrt{T}} \right\rceil$, LOCAL
**Initialize:** $\{\text{LOCAL}_s : s = 1, 2, \cdots, |\mathcal{S}|\}$ ▷ Initialize one instance for each state.
**for** $t = 1, 2, \ldots, T$ **do**
  Observe state $S_t$
  Obtain action distr. $\pi_t(\cdot|S_t)$ (from $\text{LOCAL}_{S_t}$)
  Draw $A_t \sim \pi_t(\cdot|S_t)$
  Observe reward $R_t$
  **if** $t > H$ **then**
    Cumulative gain $\bar{G}_{t-H} = \sum_{i=t-H}^{t} \gamma^i R_i$
    Return $\bar{G}_{t-H}$ to $\text{LOCAL}_{S_{\{t-H\}}}$ as feedback ▷ Delayed feedback and local update.
  **end**
**end**

### 4.1. Monte Carlo Estimator

To estimate the value of a policy, one could straightforwardly use a Monte Carlo estimator $G_t := \sum_{i=t}^{\infty} \gamma^{i-t} R_i$, as implemented in methods such as REINFORCE and MCES (Sutton and Barto, 2018). In our setting of infinite horizon MDPs, we practically use its finite horizon counterpart $\bar{G}_t := \sum_{i=t}^{t+H} \gamma^{i-t} R_i$, with the *effective horizon* $H = \mathcal{O}(\log \sqrt{T} / \log(1/\gamma)) = \tilde{O}(1/\log(1/\gamma))$, as defined in Algorithm 1. However, given that our policy changes due to local bandit updates during the period of collecting $G_t$, $G_t$ is not an unbiased estimator of $Q_t(S_t, A_t)$. This issue also applies to $\bar{G}_t$. Instead, $G_t$ and $\bar{G}_t$ are unbiased to the conditional expectations below,

$$U_t(s,a) := \mathbb{E}[G_t | S_t = s, A_t = a, \mathcal{F}_{t-1}], \quad \bar{U}_t(s,a) := \mathbb{E}[\bar{G}_t | S_t = s, A_t = a, \mathcal{F}_{t-1}].$$

Note that $U_t$ is a non-stationary analogue of action-value function $Q_t$. The difference is that $Q_t(s,a)$ depends only on the stationary policy $\pi_t$ while $U_t$ depends on the past histories $\mathcal{F}_{t-1} := \{(S_i, A_i, R_i) : 1 \le i \le t-1\}$, in additional to the MDP. As with $Q_t$, $U_t$ is well defined even at states and actions other than those visited at time $t$.

## 5. Regret Analysis

In developing the proof for our main theorem, Theorem 20, we (1) begin by decomposing the global regret into local regrets; (2) then address the challenges posed by our algorithm designs and the regret decomposition; (3) and conclude the final theorem with prior results.

### 5.1. Global to Local

We begin by defining *local regret with oracle feedback* (referred to as *local regret* when no confusion arises) as follows,

**Definition 3** *For $s \in \mathcal{S}$, the local regret with oracle feedback $Q_t$ is defined as:*

$$\mathfrak{R}_s(T) := \sum_{t=1}^{T} \left[ \mathbb{E}_{a \sim \pi^*(s)} Q_t(s,a) - \mathbb{E}_{a \sim \pi_t(s)} Q_t(s,a) \right] = \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \left( \pi^*(a|s) - \pi_t(a|s) \right) Q_t(s,a).$$

We adapt the idea by Even-Dar et al. (2009) that the global regret can be decomposed into local ones, to our discounted setting along with the new (global) regret definition. In Lemma 4, we show that the expected regret of learning in MDPs can be bounded by the cumulative regret of the set of local bandit problems, assuming the feedback $Q$-functions are given by an oracle. This can be done with the help of *performance difference lemma* (Kakade and Langford, 2002; Kakade, 2003), which is deferred to Appendix D along with the proof of Lemma 4.

**Lemma 4** *The global* $\mathfrak{Regret}(T)$ *can be bounded by the cumulative local regret, scaled by* $\frac{1}{1-\gamma}$

$$\mathfrak{Regret}(T) = \sum_{t=1}^{T} \left(V^*(s_t) - V^{\pi_t}(s_t)\right) \leq \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} \mathfrak{R}_s(T).$$

Now we decompose our problem into smaller pieces, where each state is in fact corresponding to a LOCAL bandit learner. This decomposition allows us to conduct analysis at the bandit level.

### 5.2. Objective Mismatch

While Lemma 4 helps us to break down our problem into sub-problems, it also introduces some challenges. The first major challenge is the discrepancy between the oracle feedback $Q_t$ and our approximation target $\bar{U}_t$. As discussed in Section 4.1, $\bar{G}_t$ is an unbiased estimator of $\bar{U}_t$ but is biased to $Q_t$, while the local regret $\mathfrak{R}_s$ is unfortunately measured using $Q_t$. Therefore, we refer to this issue as objective mismatch.

To address the mismatch between objectives, we rely on the slowly changing property. Intuitively speaking, the deviation of expected return, $\bar{U}_t$ versus $Q_t$, should be relatively small if the policy changes sufficiently slow. Thanks to the slowly changing guarantee, we show that one could bound the gap between $\bar{U}_t$ and $Q_t$ in Lemma 5.

**Lemma 5** *If* MAIN *is slowly changing with a rate of* $c_T$*, then*

$$\left|\bar{U}_t(s,a) - Q_t(s,a)\right| \leq \frac{H(S+HA)}{1-\gamma} c_T + \frac{1}{\sqrt{T}}.$$

This gap shows that the additional error introduced by the non-stationarity during the effective horizon $H$ can be controlled. We defer its proof to Appendix E as it is quite technical.

**Corollary 6** *Let* $\mathfrak{R}_s^{\bar{U}}(T) \coloneqq \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} (\pi^*(a|s) - \pi_t(a|s)) \bar{U}_t(s,a)$*, we have*

$$\left|\mathfrak{R}_s^{\bar{U}}(T) - \mathfrak{R}_s(T)\right| \leq \frac{2H(S+HA)}{1-\gamma} c_T T + 2\sqrt{T}.$$

Corollary 6 shows that the difference between local regret measured with $\bar{U}_t$ and local regret with oracle feedback $Q_t$ is manageable, if $c_T$ is sufficiently small. It hence allows us to analyze the local problems using the oracle feedback $Q_t$, instead of $\bar{U}_t$ that the actual feedback $\bar{G}_t$ approximates. This largely simplifies our subsequent analysis as $Q_t$ is much easier to handle.

### 5.3. Sticky Bandits

Another challenge is that at each time we are only in a single state so only a single bandit is updated, while local regret $\mathfrak{R}_s(T)$ is measured over the entire time span $T$. We term this the *sticky* bandit setting, in the spirit of sticky actions in the Arcade learning environment (Machado et al., 2018), because from the perspective of a bandit it is given feedback and the opportunity to change its policy only occasionally.

#### 5.3.1. GENERAL DEFINITIONS

We start with a general definition to isolate the issue of sticky bandits.

**Definition 7 (Sticky Bandit)** *Let $T$ be the total number of rounds, and $X_i$ be the time $t$ at which the bandit is allowed to act for the $i$-th time. The action is sticky in the sense that $p_t(a) = p_{X_i}(a)$ for $X_i \leq t < X_{i+1}$, where $p_t$ is the distribution over $\mathcal{A}$ at time $t$.*

As the decomposition lemma requires the regret of a local bandit during the full time span $T$, we thereby define three regret notions, full (time) span regret, observed regret and unobserved regret.

**Definition 8** *Full-span regret $R^{fs}(T)$, observed regret $R^{ob}(T)$ and unobserved regret $R^{un}(T)$*

$$
\begin{aligned}
R^{fs}(T) &:= \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \left(p_t^*(a) - p_t(a)\right) r_t(a) \\
R^{ob}(T) &:= \mathbb{E}_{\{X_i\}} \sum_{t \in \{X_i\}} \sum_{a \in \mathcal{A}} \left(p_t^*(a) - p_t(a)\right) r_t(a) \\
R^{un}(T) &:= R^{fs}(T) - R^{ob}(T).
\end{aligned}
$$

It is note-worthy that $R^{\text{fs}}(T)$ degenerates to local regret $\mathfrak{R}_s(T)$, if one apply $p_t(a) = \pi_t(a|s)$, $p_t^* = \pi^*(a|s)$ and $r_t(a) = Q_t(s,a)$. Therefore, if one could prove that sub-linear observed regret implies sub-linear full-span regret, then we could translate observed regret of local bandits to global regret in MDPs. Assumptions made in Section 3.3 ensure that each state will be visited sufficiently often, meaning each local bandit will be updated often. However, it is also generally impossible for an arbitrary bandit algorithm to be no-regret, for the full time span, with these assumptions alone.

**A Hard Instance for Sticky Bandits.** Consider a sticky and adversary setting with two actions $a_1$ and $a_2$. We assume the bandit learner is only able to pull every 10 rounds and the first pull is at $t = 1$ without loss generality. And the adversary choose the reward function below

$$
r_{a_1}(t) = \begin{cases} 1 & t\%10 \in [1,5], \\ 0 & \text{otherwise.} \end{cases} \qquad r_{a_2}(t) = \begin{cases} 0 & t\%10 \in [1,5], \\ 1 & \text{otherwise.} \end{cases}
$$

Then a bandit learner is likely leaning to pull $a_1$, as it is never able to observe that $a_2$ achieves a reward of 1. It in turn implies that the bandit player will have an $\mathcal{O}(T)$ full-span regret. This challenge is caused by the possibility of dramatic reward changes. Therefore one could not predict what is the regret while the bandit player cannot pull and observe, even if it pulls frequently enough. However, in Section 5.3.2 we prove that the reward/feedback function of local bandits are also in the family of slowly changing functions. Therefore one could estimate the regret occurred, when bandits are not able to react, by its latest regret seen.

### 5.3.2. LEARNING IN MDPS

We now connect these regret definitions to learning in MDPs, by applying $\pi^*$ as the comparator and $Q_t$ as the feedback function.

**Definition 9** *Full-span regret, observed and unobserved regret in MDPs are defined as follows*

$$R^{\textit{fs-mdp}} := \sum_{s\in\mathcal{S}}\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}(\pi^*(a|s)-\pi_t(a|s))Q_t(s,a) =: \sum_{s\in\mathcal{S}}\mathfrak{R}_s(T) \geq (1-\gamma)\mathfrak{Regret}(T)$$

$$R^{\textit{ob-mdp}} := \sum_{s\in\mathcal{S}}\sum_{t=1}^{T}\nu_t(s)\sum_{a\in\mathcal{A}}(\pi^*(a|s)-\pi_t(a|s))Q_t(s,a) \quad R^{\textit{un-mdp}} := R^{\textit{fs-mdp}} - R^{\textit{ob-mdp}}$$

*where the inequality follows from Lemma 4, and $\nu_t := \nu_{t-1}\mathbb{P}^{\pi_{t-1}}$ denotes the state distribution at $t$.*

In MDPs, the full-span regret $R^{\text{fs-mdp}}$ is simply defined by accumulating all local regret $\mathfrak{R}_s(T)$, given the aforementioned choices of comparator and feedback function. The observed regret $R^{\text{ob-mdp}}$ similarly accumulates the observed local ones, based on the state visitation distribution $\nu_t$.

It is clear that the observed regret is sub-linear if the bandit learners are no-regret. However, this conclusion is not sufficient to help us infer anything about full-span regret. As discussed in Section 5.3.1, the first challenge is the potential dramatic change of feedback, which in turn leads to difficulty to measure the unobserved regret. We show, in Lemma 10, that $Q_t$ is indeed slowly changing because $\pi_t$ is, with its proof deferred to Appendix F.

**Lemma 10** *If* MAIN *is slowly changing with a non-increasing rate of $c_T$, we have*

$$|Q_{t+n}(s,a) - Q_t(s,a)| \leq \frac{(S+HA)n}{1-\gamma}c_T + \frac{2}{\sqrt{T}}.$$

The second difficulty is raised by the state distribution $\nu_t$. It is generally difficult to analyze $\nu_t$ because it is a product of a sequence of prior policies. We therefore leverage the insight from the online MDPs literature (Even-Dar et al., 2009; Neu et al., 2010) that $\nu_t$ is close to its stationary distribution $\mu_t$ if the algorithm is slowly changing, as shown in Lemma 11 whose proof can be found in Appendix G. It is much easier to conduct analysis with the stationary distributions.

**Lemma 11** *If the sequence of policies $\{\pi_t\}$ is slowly changing with rate $c_T$, then*

$$||\nu_t - \mu_t||_1 \leq \tau(\tau+1)c_T + 2e^{-(t-1)/\tau}.$$

**Corollary 12** *As a result of Lemma 11, one could bound the observed/unobserved regret as follows,*

$$R^{\textit{ob-mdp}} \leq \kappa T + \underbrace{\sum_{s\in\mathcal{S}}\sum_{t=1}^{T}\mu_t(s)\sum_{a\in\mathcal{A}}(\pi^*(a|s)-\pi_t(a|s))Q_t(s,a)}_{=:\tilde{R}^{\textit{ob-mdp}}}$$

$$R^{\textit{un-mdp}} \leq \kappa T + \underbrace{\sum_{s\in\mathcal{S}}\sum_{t=1}^{T}(1-\mu_t(s))\sum_{a\in\mathcal{A}}(\pi^*(a|s)-\pi_t(a|s))Q_t(s,a)}_{=:\tilde{R}^{\textit{un-mdp}}}$$

$$\kappa = \left(\tau(\tau+1)c_T + 2e^{-(t-1)/\tau}\right)/(1-\gamma).$$

These bounds are useful as $\mu_t$ is uniformly bounded below given Assumption 1, which in turn implies uniformly sufficient visitation. Combined with the slowly changing feedback as established in Lemma 10, these conditions together are adequate to address the challenges posed by the sticky bandit issue. We have now converted the original problem associated with $\nu_t$, to a surrogate problem with stationary distributions $\mu_t$.

The observed and unobserved regret, $\tilde{R}^{\text{ob-mdp}}$ and $\tilde{R}^{\text{un-mdp}}$, for this surrogate problem are defined in Corollary 12. Bounding the surrogate unobserved regret $\tilde{R}^{\text{un-mdp}}$ leads to a bound of the original regret $R^{\text{un-mdp}}$. Now we are ready to show, in Lemma 13, that $\tilde{R}^{\text{un-mdp}}$ can be bounded by $\tilde{R}^{\text{ob-mdp}}$ up to a factor $\beta$ as well as additional terms that are sub-linear in $T$, with proper choice of $c_T$.

**Lemma 13** *If Assumption 1 and 2 are satisfied, and the* LOCAL *learner is slowly changing with a rate $c_T$, we have*

$$\tilde{R}^{\textit{un-mdp}}(T) \leq \frac{\tilde{R}^{\textit{ob-mdp}}(T)}{\beta} + \frac{2(S+HA)}{(1-\gamma)\beta^3} c_T T + 4S\sqrt{T}.$$

It in turn leads to our second key result regarding the full-span regret $R^{\text{fs-mdp}}(T)$ in Corollary 14, following from Corollary 12 and Lemma 13.

**Corollary 14** *Suppose assumption 1 and 2 are satisfied. If a slowly changing* LOCAL *bandit, with a rate of $c_T$, enjoys $R^{\textit{ob-mdp}}(T) = \tilde{\mathcal{O}}(g(\cdot)S\sqrt{T})$ observed regret, then the full-span regret is*

$$R^{\textit{fs-mdp}}(T) = \tilde{\mathcal{O}}\left(\frac{g(\cdot)S}{\beta}\sqrt{T} + \left(\frac{S+HA}{(1-\gamma)\beta^3} + \frac{\tau^2}{1-\gamma}\right) c_T T\right)$$

*where $g(\cdot)$ is a function of other problem parameters, such as $A$ and $H$, specified in later sections.*

### 5.4. Delayed Feedback

Due to our construction, we introduced *constant* feedback delays into our Algorithm 1. For the purpose of black-box reduction, one need to address the delays in a black-box fashion. We leverage the result from Joulani et al. (2013), which bounds the regret of delayed problems for arbitrary bandit algorithm with its non-delayed guarantees. They provide a black-box algorithm for (arbitrary) delay. The algorithm is presented in Algorithm 2, in the context of constant delay. However, this step may not be necessary in practice, as many adversary bandit algorithms have been shown robust to constant delay (Neu et al., 2010; Joulani et al., 2013; Cesa-Bianchi et al., 2016; Pike-Burke et al., 2018; Bistritz et al., 2019; Thune et al., 2019), etc. See further discussion in Section 6.2.

The essence of the construction is using $H+1$ BASE instances, so that each instance can update after receiving the feedback of its last decision. Therefore, a delayed problem is reduced to $H+1$ non-delayed problems. Now it is possible to handle the delayed feedback in a black-box fashion.

**Lemma 15** *[Joulani et al. (2013)] Suppose that the* BASE *used in* LOCAL *enjoys an expected regret bound $R^{\text{BASE}}(T)$ in non-delayed setting. Assume, furthermore, that the delays are constant $H$. Then the expected regret of* LOCAL *after $T$ time steps satisfies*

$$R^{\text{LOCAL}} \leq (H+1) R^{\text{BASE}}\left(T/(H+1)\right).$$

**Corollary 16** *Suppose* BASE *has $\tilde{\mathcal{O}}(f(A)\sqrt{T})$ regret, Algorithm 2 then enjoys $\tilde{\mathcal{O}}(f(A)\sqrt{HT})$ regret, where $f(A)$ denotes the dependency on $A$.*

**Algorithm 2:** (LOCAL$_s$) black-box online learning under (constant) delayed feedback

**Require:** constant delay $H$
**Initialize:** BASE$_s^h : h = 1, 2, \ldots, H+1$
**for** $t = 1, 2, \ldots, T$ **do**
  Set $h_t = [t \bmod (H+1)] + 1$
  Choose BASE$_{s_t}^{h_t}$ to make prediction
  **if** $t > H$ **then**
    Receive feedback $\bar{G}_{t-H}$
    Update BASE$_{s_{t-H}}^{h_{t-H}}$ with $\bar{G}_{t-H}$
  **end**
**end**

### 5.5. MAIN is Slowly Changing

To summarize, the black-box reduction flow is now

$$\text{MAIN} \xrightarrow{\text{Algo. 1}} \text{LOCAL}_s \xrightarrow{\text{Algo. 2}} \text{BASE}_s^h$$

with $\mathcal{O}(H)$ bandit learners per state, and $\mathcal{O}(HS)$ in total.

While Section 5.2 and Section 5.3 rely on the slowly changing property of MAIN, we have not yet show that MAIN is slowly changing if BASE is slowly changing. It is not difficult to see that MAIN is slowly changing if LOCAL is. The corresponding lemma and its proof can be found in Appendix I. Unfortunately, even if BASE is slowly changing, LOCAL is not necessarily slowly changing due to the switching mechanism - alternating among various BASE instances - designed by Algorithm 2, as each of the $H+1$ BASE bandits could have arbitrarily different policies.

However, one could preserve the slowly changing property by incorporating the timestep $h$ as part of the state. In other words, one could augment the state space $\mathcal{S}$ by concatenating a state $s$ with a time stamp $h \in \mathcal{H} := \{1, 2, \ldots, H+1\}$. Definition 17 gives a formal statement of $\mathcal{H}$-augmented MDPs. Similarly, constructing timestep (of episodes) as part of the state is often seen in episodic settings (Jin et al., 2018; Wang et al., 2021, etc.).

**Definition 17** *Given a MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$, and let $\mathcal{H} := \{1, 2, \ldots, H+1\}$. We define the $\mathcal{H}$-argumented MDP as $\tilde{M} = (\tilde{\mathcal{S}}, \mathcal{A}, \tilde{\mathbb{P}}, \tilde{r}, \gamma)$, where $\tilde{\mathcal{S}} := \mathcal{S} \times \mathcal{H}$, $\tilde{r}(s \circ h, a) := r(s, a)$, $\tilde{\mathbb{P}}(s \circ h, a, s' \circ h') := \mathbb{P}(s, a, s') \mathbb{1}\{h' = [h + 1 \bmod H + 1]\}$, where $\mathbb{1}\{\cdot\}$ is indicator function and $\circ$ denotes concatenation.*

**Lemma 18** *While applying Algorithm 2 as* LOCAL*, $\tilde{\pi}$ is slowly changing in $\tilde{\mathcal{M}}$, where $\tilde{\pi}_t(a|s \circ h) := \pi_t(a|s)$ and $\pi_t$ is produced by* MAIN*.*

Proof of Lemma 18 is deferred to Appendix I. The switching mechanism in Algorithm 2 is now part of the transition function and it is then possible to preserve the slowly changing property. The costs are that (1) we increased the cardinality of state space to $\mathcal{O}(HS)$, (2) the stationary distribution is now bounded below by $\mathcal{O}(\beta/H)$, as $\beta$ was a uniform bound which is therefore independent of $t$ and $h$. For completeness, Lemma 19 shows that $\tilde{\mathcal{M}}$ satisfies our assumptions on $\mathcal{M}$. The proof is deferred to Appendix J. Besides, as $\pi$ uniquely determines $\tilde{\pi}$, we therefore simply use $\pi$ to denote $\tilde{\pi}$ for brevity, for example we write $\tilde{\mu}^{\pi}$, $\tilde{\mathbb{P}}^{\pi}$ instead of $\tilde{\mu}^{\tilde{\pi}}$, $\tilde{\mathbb{P}}^{\tilde{\pi}}$.

**Lemma 19** *If assumption 1 and assumption 2 hold for an MDP $\mathcal{M}$, then for its $\mathcal{H}$-augmented counterpart $\tilde{\mathcal{M}}$,*

*(1) there is an unique stationary distribution $\tilde{\mu}^\pi$ for any $\pi$;*
*(2) $\inf_{\pi,\tilde{s}} \tilde{\mu}^\pi(\tilde{s}) \geq \beta/(H+1)$, where $\tilde{s} \in \tilde{\mathcal{S}}$;*
*(3) $\sup_\pi \|(\tilde{d} - \tilde{d}')\tilde{\mathbb{P}}^\pi\|_1 \leq e^{-1/\tau}\|\tilde{d} - \tilde{d}'\|_1$, for any $\tilde{d}$, $\tilde{d}'$.*

### 5.6. Main Theorem

We are now ready to present our main theorem. Theorem 20 concludes our reduction from RL to adversary bandits, by combining our prior results.

**Theorem 20** *When assumption 1 and assumption 2 hold, apply Algorithm 2 as* LOCAL*, suppose* BASE *of Algorithm 2 enjoys $\tilde{\mathcal{O}}(f(A)\sqrt{T})$ expect regret in standard adversary setting and is slowly changing with rate $c_T$, then* MAIN *enjoys an expect regret of*

$$\mathfrak{Regret}(T) = \tilde{\mathcal{O}}\left(\frac{H^{2.5}Sf(A)}{(1-\gamma)\beta}\sqrt{T} + \frac{\tau^2 H^4 S(S+A)}{(1-\gamma)^2\beta^3}c_T T\right).$$

*where $f(A)$ is the dependency on $A$ of running* BASE *in a standard adversarial non-delayed setting.*

**Proof** The regret analysis is structured into aforementioned components. Let's first consider the full-span regret that accumulates all local regrets $R^{\text{fs-mdp}} = \sum_{s\in\mathcal{S}} \mathfrak{R}_s(T)$.

1. Delayed Feedback: Given Corollary 16, $\text{LOCAL}_s$ has a regret of $\tilde{\mathcal{O}}(f(A)\sqrt{HT})$ for state $s$.
2. Sticky Bandits: Since the observed regret at state $s$ is now at most $\tilde{\mathcal{O}}(f(A)\sqrt{HT})$, applying $g(A,H) = f(A)\sqrt{H}$ for Corollary 14 leads to $\tilde{\mathcal{O}}\left(\frac{\sqrt{H}Sf(A)}{\beta}\sqrt{T} + \left(\frac{S+HA}{(1-\gamma)\beta^3} + \frac{\tau^2}{1-\gamma}\right)c_T T\right)$.
3. Objective Mismatch: Corollary 6 establishes that the additional error from objective mismatch is at most $\sum_s 2(mc_T T + \sqrt{T})$, where $m = H(S+HA)/(1-\gamma)$. It in turn leads to a bound of $\tilde{\mathcal{O}}\left(\frac{\sqrt{H}Sf(A)}{\beta}\sqrt{T} + \left(\frac{S+HA}{(1-\gamma)\beta^3} + \frac{\tau^2}{1-\gamma} + \frac{HS(S+HA)}{1-\gamma}\right)c_T T\right)$.
4. $\mathcal{H}$-augmented MDPs: As we expand $\mathcal{S}$ to $\tilde{\mathcal{S}}$, the cardinality of state space is then $\mathcal{O}(HS)$. To accommodate this expansion, we replace $S$ with $HS$ and $\beta$ with $\beta/H$, leading to a bound of $\tilde{\mathcal{O}}\left(\frac{H^{2.5}Sf(A)}{\beta}\sqrt{T} + \left(\frac{H^4(S+A)}{(1-\gamma)\beta^3} + \frac{\tau^2}{1-\gamma} + \frac{H^3 S(S+A)}{1-\gamma}\right)c_T T\right)$.

We now translate $R^{\text{fs-mdp}}$ into $\mathfrak{Regret}(T)$.

5. Lemma 4 establishes that $\mathfrak{Regret}(T) \leq \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}} \mathfrak{R}_s(T) = \frac{1}{1-\gamma}R^{\text{fs-mdp}}$, leading to the regret bound of $\tilde{\mathcal{O}}\left(\frac{H^{2.5}Sf(A)}{(1-\gamma)\beta}\sqrt{T} + \frac{\tau^2 H^4 S(S+A)}{(1-\gamma)^2\beta^3}c_T T\right)$.

This concludes our main result. ■

**Corollary 21** *Suppose the conditions in Theorem 20 are met. Given that $H = \mathcal{O}(\log\sqrt{T}/\log(1/\gamma))$, when $\gamma$ is close to 1, the bound presented in Theorem 20 becomes $\tilde{\mathcal{O}}\left(\frac{Sf(A)}{(1-\gamma)^{3.5}\beta}\sqrt{T} + \frac{\tau^2 S(S+A)}{(1-\gamma)^6\beta^3}c_T T\right)$.*

## 6. Case Study: EXP3

We further extend our discussion on our reduction by providing an example with a well-known exponential-weight bandit algorithm EXP3 (Auer et al., 2002).

### 6.1. EXP3 as BASE

We first present a regret bound while applying EXP3 as BASE in our reduction. In a standard adversarial non-delayed setting EXP3 has $\tilde{\mathcal{O}}(\sqrt{AT})$ regret (Auer et al., 2002) and a slowly-changing rate of $c_T = \tilde{\mathcal{O}}(\sqrt{1/AT})$. Applying Theorem 20 with aforementioned regret and changing rate leads to Corollary 22, and discussion on this rate $c_T$ can be found in Section 6.3 and Appendix K.

**Corollary 22** *Applying* EXP3 *as* BASE, MAIN *has a regret bound of* $\tilde{\mathcal{O}}\left(\frac{\tau^2 H^4 S(S+A)}{(1-\gamma)^2\beta^3}\sqrt{T}\right)$, *which becomes* $\tilde{\mathcal{O}}\left(\frac{\tau^2 S(S+A)}{(1-\gamma)^6\beta^3}\sqrt{T}\right)$, *when* $\gamma$ *is close to* 1.

### 6.2. EXP3 as LOCAL

It is known that the optimal regret is $\tilde{\mathcal{O}}(\sqrt{(A+z)T})$ for constant delay $z$ (Cesa-Bianchi et al., 2016), and remarkably, EXP3 achieves the optimal bound (Thune et al., 2019). Furthermore, for unrestricted delays, Bistritz et al. (2019) and Thune et al. (2019) show that EXP3 enjoys $\tilde{\mathcal{O}}(\sqrt{AT+Z})$ , where $Z$ is the total delay. EXP3 therefore enjoys $\tilde{\mathcal{O}}(\sqrt{(A+H)T})$ regret under our delay $H$.

In previous sections, we use Algorithm 2 as LOCAL, for the purpose of black-box reduction. However, LOCAL can be any delay-robust adversarial bandit algorithm, such as EXP3. Corollary 23 establishes the result when one use EXP3 as LOCAL, refining the dependency on $H$ compared to Corollary 22, which handles delays in a black-box fashion using Algorithm 2.

**Corollary 23** *When using* EXP3 *as* LOCAL, *the observed regret of* LOCAL *is* $\tilde{\mathcal{O}}(\sqrt{(A+H)T})$, *and* MAIN *meets the slowly changing condition without needing the* $\mathcal{H}$*-augmented trick, leading to a regret of* $\tilde{\mathcal{O}}\left(\frac{\tau^2(HS^2+H^2SA)}{(1-\gamma)^2\beta^3}\sqrt{T}\right)$, *which turns to* $\tilde{\mathcal{O}}\left(\frac{\tau^2(S^2+SA/(1-\gamma))}{(1-\gamma)^3\beta^3}\sqrt{T}\right)$ *when* $\gamma$ *is close to* 1.

### 6.3. EXP3 is Slowly Changing

It can be shown that EXP3 meets the slowly changing requirement with a rate of $\eta_T/A$, where $\eta_T$ is the learning rate of EXP3. See Appendix K for a pseudocode of EXP3 and the proof of Lemma 24.

**Lemma 24** *Let* $\eta_t$ *be the learning rate of* EXP3, EXP3 *is slowly changing with a rate of* $\mathcal{O}(\eta_T/A)$, *assuming the feedback is bounded within the range* $[0, 1/(1-\gamma)]$.

We note that to achieve $\tilde{\mathcal{O}}(\sqrt{AT})$ regret EXP3 is run with a learning rate of $\eta_T = \tilde{\mathcal{O}}(\sqrt{A/T})$, which means it is slowly changing with a rate of $c_T = \tilde{\mathcal{O}}(\sqrt{1/AT})$.

## 7. Conclusion

In this work, we explore the mathematical connections between RL and bandits, in a natural decentralized setting. Our result could serve as a theoretical tool to facilitate generalizing existing bandit results to MDPs, as demonstrated with the example of delayed feedback in Section 5. It can also be linked to multi-agent RL and Monte Carlo methods, as discussed in Section 1 and 2. However, our

results require additional assumptions, and the parameter dependencies, such as those on $S$ and $H$, could still be improved. We further extend our discussion on these limitations and future directions.

One limitation of our work is the need for two extra assumptions not typically needed for discounted infinite-horizon MDPs. These assumptions ensure that all states are visited sufficiently often, hence making the exploration in MDPs less difficult. Yet it remains unclear to us whether more aggressive local exploration or algorithm-dependent exploration incentives for local bandit learners could mitigate the need for these assumptions. However, from the perspective of Monte Carlo learning, our assumptions play a role akin to the exploring starts in the MCES algorithm, as both ensure adequate exploration. Hence, eliminating such assumptions could be an important direction for our framework with Monte Carlo evaluation. Another limitation of our result is the relatively large dependency on parameters such as the effective horizon $H$ and the state space size $S$. In Section 6, we show that the dependency on $H$ can be refined if one directly applies EXP3 as LOCAL. As for the dependency on $S$, we believe that, in our current framework, one could not do better than linear dependency on $S$, as accounting for possible policy changes on all states during the effective horizon $H$ unavoidably creates an additional $S$. While our work considers adversarial bandits, the varying feedback is in fact not caused by the environment but by the policy changes of its co-learners. Hence, alleviating the requirement of adversarial bandits to stochastic ones could be another important direction. For example, one may consider stochastic bandit algorithms with predefined policy-change times to address the non-stationary feedback, since it is known that stochastic bandits attain optimal bounds when the number of changes is known in advance (Auer et al., 2019). Moreover, while we focus on the tabular setting, prior work (Brown et al., 2019) has shown how algorithms using a regret minimizer in every state, such as CFR (Zinkevich et al., 2007), have practical implementations via function approximation, which could be another intriguing direction.

## Acknowledgments

We thank the reviewers and the meta-reviewer for assessing our paper and for their constructive feedback. This work is supported by the National Science Foundation (NSF) grant CCF-1934915 and the NSF grant ECCS-2217023. Zishun is supported in part by the National Institutes of Health (NIH) grant R01CA258827.

## References

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pages 322–331. IEEE, 1995.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pages 138–158. PMLR, 2019.

Peter L Bartlett and Ambuj Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42, 2009.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. Online exp3 learning in adversarial bandits with delayed feedback. *Advances in neural information processing systems*, 32, 2019.

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *International conference on machine learning*, pages 793–802. PMLR, 2019.

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134. PMLR, 2013.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Nicol'o Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622. PMLR, 2016.

Ching-An Cheng, Remi Tachet Combes, Byron Boots, and Geoff Gordon. A reduction from reinforcement learning to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3514–3524. PMLR, 2020a.

Ching-An Cheng, Jonathan Lee, Ken Goldberg, and Byron Boots. Online learning with continuous variations: Dynamic regret and reductions. In *International Conference on Artificial Intelligence and Statistics*, pages 2218–2228. PMLR, 2020b.

Romain Cravic, Nicolas Gast, and Bruno Gaujal. Decentralized model-free reinforcement learning in stochastic games with average-reward objective. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 1230–1238, 2023.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

Vektor Dewanto, George Dunn, Ali Eshragh, Marcus Gallagher, and Fred Roosta. Average-reward model-free reinforcement learning: a systematic review and literature mapping. *arXiv preprint arXiv:2010.08920*, 2020.

Zixuan Dong, Che Wang, and Keith Ross. On the convergence of monte carlo ucb for random-length episodic mdps. *arXiv preprint arXiv:2209.02864*, 2022.

S Rasoul Etesami. Learning stationary nash equilibrium policies in $n$-player stochastic games with independent chains via dual mirror descent. *arXiv preprint arXiv:2201.12224*, 2022.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Near optimal exploration-exploitation in non-communicating markov decision processes. *Advances in Neural Information Processing Systems*, 31, 2018a.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pages 1578–1586. PMLR, 2018b.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. *Advances in Neural Information Processing Systems*, 29, 2016.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted mdps. *Advances in Neural Information Processing Systems*, 34:22288–22300, 2021.

Benjamin Howson, Ciara Pike-Burke, and Sarah Filippi. Delayed feedback in episodic reinforcement learning. *arXiv preprint arXiv:2111.07615*, 2021.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2020.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*, 2022a.

Tiancheng Jin, Tal Lancewicki, Haipeng Luo, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret for adversarial mdp with delayed bandit feedback. *Advances in Neural Information Processing Systems*, 35:33469–33481, 2022b.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461. PMLR, 2013.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

J Zico Kolter and Andrew Y Ng. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning*, pages 513–520, 2009.

Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*, pages 320–334. Springer, 2012.

Tor Lattimore, Marcus Hutter, and Peter Sunehag. The sample-complexity of general reinforcement learning. In *International Conference on Machine Learning*, pages 28–36. PMLR, 2013.

Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Shengshi Li and Lin Yang. Horizon-free learning for markov decision processes and games: stochastically bounded rewards and improved bounds. In *International Conference on Machine Learning*, pages 20221–20252. PMLR, 2023.

Jun Liu. On the convergence of reinforcement learning with monte carlo exploring starts. *Automatica*, 129:109693, 2021.

Shuang Liu and Hao Su. Regret bounds for discounted mdps. *arXiv preprint arXiv:2002.05138*, 2020.

Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

Pierre Ménard, Omar Darwiche Domingues, Xuedong Shang, and Michal Valko. Ucb momentum q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pages 7609–7618. PMLR, 2021.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

Washim Uddin Mondal and Vaneet Aggarwal. Reinforcement learning with delayed, composite, and partially anonymous reward. *arXiv preprint arXiv:2305.02527*, 2023.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. In *Proceedings of the Twenty-Fourth Annual Conference on Neural Information Processing Systems*, 2010.

Ronald Ortner. Regret bounds for reinforcement learning via markov chain concentration. *Journal of Artificial Intelligence Research*, 67:115–128, 2020.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pages 2701–2710. PMLR, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. *Advances in neural information processing systems*, 30, 2017.

Aldo Pacchiano, Philip Ball, Jack Parker-Holder, Krzysztof Choromanski, and Stephen Roberts. Towards tractable optimism in model-based reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 1413–1423. PMLR, 2021.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR, 2018.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. *Advances in Neural Information Processing Systems*, 32, 2019.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 32, 2019.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888, 2006.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

István Szita and Csaba Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML*, 2010.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pages 770–805. PMLR, 2018.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. *Advances in Neural Information Processing Systems*, 32, 2019.

John N Tsitsiklis. On the convergence of optimistic policy iteration. *Journal of Machine Learning Research*, 3(Jul):59–72, 2002.

Che Wang, Shuhan Yuan, Kai Shao, and Keith W Ross. On the convergence of the monte carlo exploring starts algorithm for reinforcement learning. In *International Conference on Learning Representations*, 2021.

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

Yuanhao Wang, Kefan Dong, Xiaoyu Chen, and Liwei Wang. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. In *International Conference on Learning Representations*, 2020b.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR, 2021.

Anna Winnicki and R Srikant. On the convergence of policy iteration-based reinforcement learning with monte carlo policy evaluation. In *International Conference on Artificial Intelligence and Statistics*, pages 9852–9878. PMLR, 2023.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33:4358–4369, 2020.

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous q-learning. *IEEE Transactions on Information Theory*, 2023.

Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. PMLR, 2021.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. *Advances in Neural Information Processing Systems*, 32, 2019.

Zihan Zhang and Qiaomin Xie. Sharper model-free reinforcement learning for average-reward markov decision processes. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5476–5477. PMLR, 2023.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33: 15198–15207, 2020.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR, 2021.

Zihan Zhang, Xiangyang Ji, and Simon Du. Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory*, pages 3858–3904. PMLR, 2022.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20, 2007.

## Appendix A. A Hard Instance

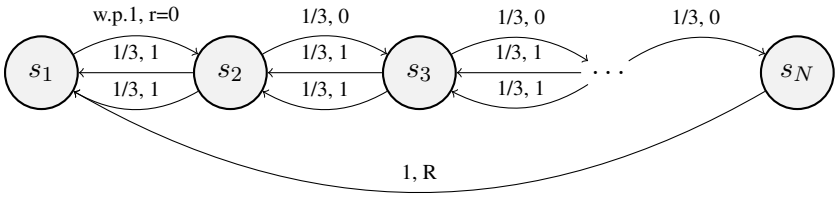


Figure 1: A hard instance for bandits, $\xrightarrow{p,r}$ denotes transition probability $p$ and reward $r$.

Consider the following deterministic MDP, where most nodes, except $s_1$ and $s_N$, have three actions $l_1, l_2$ and $r$, which stand for action going left and action going right, respectively. And let $S_0 = s_1$ w.p. 1. Going left with $l_1, l_2$ always admits a small reward 1, but the transition $s_N \to s_1$ has a large reward $R$.

Now we place in each state a bandit learner. As bandits are often initialized to assign equal probability to all actions, therefore $p(A_t = r) \leq p(A_t = l_i)$ prior to the first time hitting $s_N$. Therefore, one could consider a uniform policy $\pi(a|s_i) = 1/3$ for $a = l_1, l_2, r$, without losing generality.

As a result, the Markov chain induced by $\pi$ is equivalent to random walking on positive integers with a biased coin, prior to first hitting $s_N$. Let $M_n$ be the first hitting time of $s_n$, then $\mathbb{E}_\pi[M_n] = \mathcal{O}(2^n)$. Let $N$ be $\mathcal{O}(\sqrt{T})$, as if $N \geq T$ then no policy can be no-regret, (and for example, one could choose $R = \Omega(1/\gamma^T)$ so that the optimal policy is $\pi^*(r|s) = 1$). Then, we have $\mathbb{E}_\pi[M_N] = \mathcal{O}(2^{\sqrt{T}}) \geq \mathcal{O}(T)$. The expected first hitting time $\mathbb{E}_\pi[M_N]$ being $\mathcal{O}(2^{\sqrt{T}})$ implies bandits will have $\mathcal{O}(T)$ expected regret. Besides, $\mathcal{O}(T)$ first hitting time implies that this instance is an violation of our assumptions. Therefore, we need additional assumptions on MDPs made in Section 3.3.

But this instance will not be an issue for temperal difference approaches with UCB exploration, for example $\infty$-UCB (Wang et al., 2020b). UCB exploration assigns an exploration bonus to all $(s, a)$ pairs based on the number of visitations of $(s, a)$. Therefore, the states on the right will carry a larger bonus because they are rarely visited and the bonus will be propagated via temperal difference backupsto states on the left. As a result, $\infty$-UCB will be encouraged to choose $r$ for exploration, although one has to fine-tune the value of bonus. In contrast to our approach, one could consider UCB exploration is centralized as there is a central controller to compute exploration bonus for all state-action pairs $(s, a)$. As our reduction is in a decentralized setting, explorations purely rely on independent bandit learners, which leads to this additional difficulty.

## Appendix B. Technical Tools

We first introduce some technical tools, which are useful for our omitted proofs

**Lemma 25** $||\cdot||_{1,\infty}$ *is a norm*

**Proof.** Let $X$ and $Y$ be $n$ by $m$ matrices, and $X_{ij}$ be the element corresponding to row $i$ and col $j$

Triangle inequality:

$$||X+Y||_{1,\infty} = \max_i \sum_j |(X+Y)_{ij}| \leq \max_i \sum_j (|X_{ij}| + |Y_{ij}|) \tag{1}$$

$$\leq \max_i \sum_j |X_{ij}| + \max_i \sum_j |Y_{ij}| = ||X||_{1,\infty} + ||Y||_{1,\infty} \tag{2}$$

Absolute homogeneity:

$$||aX||_{1,\infty} = \max_i \sum_j |(aX)_{ij}| = |a| \max_i \sum_j |X_{ij}| = |a| \times ||X||_{1,\infty} \tag{3}$$

Positive definiteness ($||X||_{1,\infty} = 0 \iff X = \mathbf{0}$):

1. Let $||X||_{1,\infty} = 0$,

$$||X||_{1,\infty} = \max_i \sum_j |X_{ij}| = 0 \tag{4}$$

   implies $X_{ij} = 0$ for all $i, j$

2. Let $X = \mathbf{0}$

$$||\mathbf{0}||_{1,\infty} = \max_i \sum_j |0| = 0 \tag{5}$$

Non-negativity:

$$||X||_{1,\infty} = \max_i \sum_j |X_{ij}| \geq 0 \tag{6}$$

■

One can easily extend the slowly changing property in Definition 2 to a multi-step version,

**Lemma 26** *If an algorithm $\mathfrak{A}$ is slowly changing with a non-increasing rate of $c_T$, then*

$$||\pi_{t+k} - \pi_t||_{1,\infty} \leq k c_T \tag{7}$$

**Proof.** Trivially by triangle inequality.

$$||\pi_{t+k} - \pi_t||_{1,\infty} \leq \sum_{i=0}^{k-1} ||\pi_{t+i+1} - \pi_{t+i}||_{1,\infty} \leq \sum_{i=0}^{k-1} c_T \leq k c_T \tag{8}$$

■

It is useful to quantify the state distribution gap by following different policies, starting from the same initial state distribution.

**Lemma 27** *Suppose $||\pi - \pi'||_{1,\infty} \leq c$. Then, for any state distribution vector d, we have*

$$||d\mathbb{P}^{\pi} - d\mathbb{P}^{\pi'}||_1 \leq c \tag{9}$$

*where $\mathbb{P}^{\pi}$ is the transition matrix induced from $\pi$.*

**Proof.**

$$\begin{aligned}
||d\mathbb{P}^{\pi} - d\mathbb{P}^{\pi'}||_1 &= \sum_{s'} |d\mathbb{P}^{\pi}(s') - d\mathbb{P}^{\pi'}(s')| && (10)\\
&= \sum_{s'} |\sum_{s} [d(s)\mathbb{P}^{\pi}(s,s') - d(s)\mathbb{P}^{\pi'}(s,s')]| && (11)\\
&\leq \sum_{s'} \sum_{s} d(s)|\mathbb{P}^{\pi}(s,s') - \mathbb{P}^{\pi'}(s,s')| && (12)\\
&= \sum_{s'} \sum_{s} d(s)|\sum_{a} \mathbb{P}(s,a,s')\pi(a|s) - \mathbb{P}(s,a,s')\pi'(a|s)| && (13)\\
&\leq \sum_{s'} \sum_{s} \sum_{a} d(s)\mathbb{P}(s,a,s')|\pi(a|s) - \pi'(a|s)| && (14)\\
&= \sum_{s} d(s) \sum_{a} |\pi(a|s) - \pi'(a|s)| && (15)\\
&\leq \sum_{s} d(s)||\pi - \pi'||_{1,\infty} = c && (16)
\end{aligned}$$

■

Similarly, it is also helpful to bound the state distribution difference after following the same policy, if starting from different state distribution.

**Lemma 28** *For any state distribution vectors d and d′, we have*

$$||d\mathbb{P}^{\pi} - d'\mathbb{P}^{\pi}||_1 \leq ||d - d'||_1 \tag{17}$$

*where $\mathbb{P}^{\pi}$ is the transition matrix induced from $\pi$.*

**Proof.**

$$\begin{aligned}
||d\mathbb{P}^{\pi} - d'\mathbb{P}^{\pi}||_1 &= \sum_{s'} |d\mathbb{P}^{\pi}(s') - d'\mathbb{P}^{\pi}(s')| && (18)\\
&= \sum_{s'} |\sum_{s} [d(s)\mathbb{P}^{\pi}(s,s') - d'(s)\mathbb{P}^{\pi}(s,s')]| && (19)\\
&\leq \sum_{s'} \sum_{s} \mathbb{P}^{\pi}(s,s')|d(s) - d'(s)| && (20)\\
&= \sum_{s} |d(s) - d'(s)| \sum_{s'} \mathbb{P}^{\pi}(s,s') && (21)\\
&= ||d - d'||_1 && (22)
\end{aligned}$$

■

The case when starting from different distribution and following different policies for one step.

**Lemma 29** *Given policies $\pi$ and $\pi'$, and state distribution vectors $d$ and $d'$, if $||\pi - \pi'||_{1,\infty} \leq c$ and $||d - d'||_1 \leq \delta$, then we have*

$$||d\mathbb{P}^{\pi} - d'\mathbb{P}^{\pi'}||_1 \leq c + \delta \tag{23}$$

**Proof.**

$$||d\mathbb{P}^{\pi} - d'\mathbb{P}^{\pi'}||_1 = ||d\mathbb{P}^{\pi} - d\mathbb{P}^{\pi'} + d\mathbb{P}^{\pi'} - d'\mathbb{P}^{\pi'}||_1 \tag{24}$$

$$\leq ||d\mathbb{P}^{\pi} - d\mathbb{P}^{\pi'}||_1 + ||d\mathbb{P}^{\pi'} - d'\mathbb{P}^{\pi'}||_1 \tag{25}$$

by Lemma 28 and Lemma 27

$$\leq ||d\mathbb{P}^{\pi} - d\mathbb{P}^{\pi'}||_1 + ||d - d'||_1 \tag{26}$$

$$\leq c + \delta \tag{27}$$

■

## Appendix C. Key Technical Lemma

Extension to $n$-step case

**Lemma 30** *Given two set of policies (of equal size) $\{\pi_1, \ldots, \pi_k, \ldots, \pi_K\}$ and $\{\pi'_1, \ldots, \pi'_k, \ldots, \pi'_K\}$ and initial state distribution vectors $d$ and $d'$. If $||\pi_k - \pi'_k||_{1,\infty} \leq c$ and $||d - d'||_1 \leq \delta$, then we have*

$$||d(\mathbb{P}^{\pi_1} \cdots \mathbb{P}^{\pi_K}) - d'(\mathbb{P}^{\pi'_1} \cdots \mathbb{P}^{\pi'_K})||_1 \leq Kc + \delta \tag{28}$$

**Proof.** We prove this by induction on $K$,

$K = 1$: by Lemma 29

$$||d\mathbb{P}^{\pi_1} - d'\mathbb{P}^{\pi'_1}||_1 \leq c + \delta \tag{29}$$

$K = n$: assume we have,

$$||d(\mathbb{P}^{\pi_1} \cdots \mathbb{P}^{\pi_n}) - d'(\mathbb{P}^{\pi'_1} \cdots \mathbb{P}^{\pi'_n})||_1 \leq nc + \delta \tag{30}$$

$K = n + 1$:

$$||d(\mathbb{P}^{\pi_1} \cdots \mathbb{P}^{\pi_{n+1}}) - d'(\mathbb{P}^{\pi'_1} \cdots \mathbb{P}^{\pi'_{n+1}})||_1 \tag{31}$$

$$= ||d_n\mathbb{P}^{\pi_{n+1}} - d'_n\mathbb{P}^{\pi'_{n+1}}|| \tag{32}$$

$$\leq nc + \delta + c \tag{33}$$

■

**Corollary 31** *Let $\pi_t$ be slowly changing with non-increasing rate $c_T$, then we have $||\pi_{t+n} - \pi_t||_{1,\infty} \leq nc_T$. Apply Lemma 30 with $\pi_k = \pi_{t+n}$, $\pi'_k = \pi_t$ and $d = d'$, then*

$$||d(\mathbb{P}^{\pi_{t+n}})^K - d(\mathbb{P}^{\pi_t})^K||_1 \leq Knc_T \tag{34}$$

**Corollary 32** *Let $\pi_t$ be slowly changing with non-increasing rate $c_T$, then we have $||\pi_{t+n} - \pi_t||_{1,\infty} \leq nc_T$. Apply Lemma 30 with $\pi_k = \pi_{t+k-1}$, $\pi'_k = \pi_t$ and $d = d'$, then*

$$||d\underbrace{(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_{t+K-1}})}_{K \text{ transition kernels}} - d(\mathbb{P}^{\pi_t})^K||_1 \leq K^2 c_T \tag{35}$$

**Corollary 33** *Let $\pi_t$ be slowly changing with non-increasing rate $c_T$, then we have $||\pi_{t+n} - \pi_t||_{1,\infty} \leq nc_T$. Apply Lemma 30 with $\pi_k = \pi_{t+n+k-1}$, $\pi'_k = \pi_{t+k-1}$ and $d = d'$, then*

$$||d\underbrace{(\mathbb{P}^{\pi_{t+n}} \cdots \mathbb{P}^{\pi_{t+n+K-1}})}_{K \text{ kernels}} - d\underbrace{(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_{t+K-1}})}_{K \text{ kernels}}||_1 \leq Knc_T \tag{36}$$

These lemmas and corollaries describe how the state distribution would change by following different sequence of policies, which will be eventually used to prove Lemma 5 that bounds $|\bar{U}_t(s,a) - Q_t(s,a)|$ and Lemma 10 that bounds $|Q_{t+n}(s,a) - Q_t(s,a)|$.

## Appendix D. Proof of Lemma 4 (Decomposition Lemma)

We first introduce Performance Difference Lemma (Kakade and Langford, 2002; Kakade, 2003)

**Lemma 34 (Performance Difference Lemma.)** *Let $M$ be an MDP, then for all stationary policies $\pi$ and $\pi'$, and for all $s_0$ and $\gamma$,*

$$V^{\pi'}(s_0) - V^{\pi}(s_0) = \tfrac{1}{1-\gamma}\, \mathbb{E}_{s \sim d^{\pi'}_{s_0}}\, \mathbb{E}_{a \sim \pi'}[Q^{\pi}(s,a) - V^{\pi}(s)]$$

*where $d^{\pi'}_{s_0}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr(S_t = s | \pi', M, S_0 = s_0)$ is the normalized discounted occupancy measure starting from $s_0$ and following $\pi'$.*

**Lemma 4** *The expected regret in MDPs can be reduced to cumulative local regret with oracle feedback $Q_t$*

$$\mathfrak{Regret}(T) \leq \tfrac{1}{1-\gamma} \sum_{s \in \mathcal{S}} \mathfrak{R}_s(T) \tag{37}$$

**Proof.** Let $\{\pi_t : 1 \leq t \leq T\}$ be the sequence of policies obtained by running by any algorithm $\mathfrak{A}$.

$$\mathfrak{Regret}(T) = \sum_{t=1}^{T} V^*(s_t) - V_t(s_t) \tag{38}$$

apply Performance Difference Lemma 34 with $\pi' = \pi^*$ and $\pi = \pi_t$

$$= \tfrac{1}{1-\gamma} \sum_{t=1}^{T} \mathbb{E}_{s \sim d^{\pi^*}_{s_t}}\, \mathbb{E}_{a \sim \pi^*}[Q_t(s,a) - V_t(s)] \tag{39}$$

$$= \tfrac{1}{1-\gamma} \sum_{t=1}^{T} \mathbb{E}_{s \sim d^{\pi^*}_{s_t}} \sum_{a \in \mathcal{A}} (\pi^*(a|s) - \pi_t(a|s)) Q_t(s,a) \tag{40}$$

$$\leq \tfrac{1}{1-\gamma} \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} (\pi^*(a|s) - \pi_t(a|s)) Q_t(s,a) \tag{41}$$

by definition 3

$$= \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} \mathfrak{R}_s(T) \tag{42}$$

■

## Appendix E. Proof of Lemma 5 (Objective Mismatch)

As the local regret considers objective of state-value function $Q_t$ but our Monte Carlo estimator approximate the target $\bar{U}_t$, we now show the gap is bounded.

We define $\bar{Q}_t \coloneqq \mathbb{E}[\sum_{t=0}^{H} \gamma^t R_t | \pi, S_0 = s, A_0 = a]$, a finite horizon counterpart of $Q_t$, we will use this notation for proof in later sections as well. We first show the gap between $Q_t$ and $\bar{Q}_t$, as it is easier to compare $\bar{U}_t$ and $\bar{Q}_t$ because of the same finite horizon

**Lemma 35** *Let $\{X_i, i = 1, \cdots, \}$ be an arbitrary infinite sequence such that $X_i \in [0, 1]$ for all $i$, $\gamma$ be the discounted factor of a MDP, we have*

$$\left| \sum_{t=0}^{\infty} \gamma^t X_t - \sum_{t=0}^{H} \gamma^t X_t \right| \leq \frac{1}{\sqrt{T}} \tag{43}$$

**Proof.**

$$\left| \sum_{t=0}^{\infty} \gamma^t X_t - \sum_{t=0}^{H} \gamma^t X_t \right| = \left| \sum_{t=H+1}^{\infty} \gamma^t X_t \right| \tag{44}$$

$$= \sum_{t=H+1}^{\infty} \gamma^t = \frac{\gamma^{H+1}}{1-\gamma} \tag{45}$$

by the definition of $H$ in Algo. 1

$$\leq \frac{\gamma^{\log_\gamma (1-\gamma)/\sqrt{T}}}{1-\gamma} = \frac{1}{\sqrt{T}} \tag{46}$$

■

**Corollary 36** *As a result of Lemma 35, we have*

$$|\bar{Q}_t(s,a) - Q_t(s,a)| \leq \frac{1}{\sqrt{T}}, |\bar{U}_t(s,a) - U_t(s,a)| \leq \frac{1}{\sqrt{T}}. \tag{47}$$

Before giving the first key lemma, we first note a fact that $|a_1 b_1 - a_2 b_2| \leq |a_1 - a_2| + |b_1 - b_2|$ when $a_1, a_2, b_1, b_2 \in [0, 1]$. Let $a_1, a_2, b_1, b_2 \in [0, 1]$, we have

$$|a_1 b_1 - a_2 b_2| = |a_1 b_1 - a_1 b_2 + a_1 b_2 - a_2 b_2| \tag{48}$$

$$\leq |a_1 b_1 - a_1 b_2| + |a_1 b_2 - a_2 b_2| \tag{49}$$

$$\leq |a_1 - a_2| + |b_1 - b_2| \tag{50}$$

Now we are ready to give the first key lemma

**Lemma 5** *If* MAIN *is slowly changing with a non-increasing rate of $c_T$, then*

$$|\bar{U}_t(s,a) - Q_t(s,a)| \leq \frac{HS + H^2A}{1-\gamma} c_T + \frac{1}{\sqrt{T}} \tag{51}$$

**Proof.** Let $k = 0, 1, 2, \ldots$, and $U_t$ as defined in the main text. And recall that Corollary 32 described the state distribution gap after running different sequence of policies, starting from same distribution, shown as below

> **Corollary 32** *Let $\pi_t$ be slowly changing with non-increasing rate $c_T$, then we have $||\pi_{t+l} - \pi_t||_{1,\infty} \leq l c_T$. Apply Lemma 30 with $\pi_k = \pi_{t+k-1}$, $\pi'_k = \pi_t$ and $d = d'$, then*
>
> $$||d \underbrace{(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_{t+K-1}})}_{K \text{ transition kernels}} - d(\mathbb{P}^{\pi_t})^K||_1 \leq K^2 c_T \tag{52}$$

Noticing that the following facts perfectly fit the conditions of applying Corollary 32

- $\pi_t$ is used to evaluate $Q_t(s,a)$, $\{\pi_t, \ldots, \pi_{t+k}, \ldots\}$ are used to evaluate $U_t(s,a)$, and the initial policy is the same $\pi_t$
- The initial state are both $s$, because we are evaluating $Q_t(s,a)$ and $U_t(s,a)$. Let this deterministic distribution be $d_s$.
- The future state distributions $d_k$ (that used to evaluate $U_t$) follows $d_{k+1} = d_k \mathbb{P}^{\pi_{t+k}}$
- The future state distributions $d'_k$ (that used to evaluate $Q_t$) follows $d'_{k+1} = d'_k \mathbb{P}^{\pi_t}$

We introduce the symbols $d_s$ and $d_k$ instead of using $d_t$ and $d_{t+k}$ as to avoid confusion with the actual state distributions produced by running MAIN. $d_k$ indicates k steps in the future starting from $d_s$, where $s$ is **not** necessarily the actual visited state at time $t$. Therefore, the notions $d_k$ and $d_s$ are less attached with the distributions realized by algorithm.

Therefore, by Lemma 26 and Corollary 32, we have

$$||\pi_{t+k} - \pi_t||_{1,\infty} \leq k c_T \tag{53}$$

$$||d_s \underbrace{(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_{t+k-1}})}_{k \text{ kernels}} - d_s(\mathbb{P}^{\pi_t})^k||_1 \leq k^2 c_T \tag{54}$$

Now we are ready to bound the difference between objectives, note that we will use the prime notion $d_k$ and $d'_k$ as used in the bullet-points above. And we slightly abuse notation here by denoting an element of a vector by e.g. $d'_k(x)$, as both subscript and superscript are occupied

$$\left|\bar{U}_t(s,a) - Q_t(s,a)\right| \leq \left|\bar{U}_t(s,a) - \bar{Q}_t(s,a)\right| + \frac{1}{\sqrt{T}} \tag{55}$$

$$= \left|\sum_{k=0}^{H} \sum_{x \in \mathcal{S}} \sum_{i \in \mathcal{A}} \gamma^k \Big[d_k(x)\pi_{t+k}(i|x) - d'_k(x)\pi_t(i|x)\Big] r(x,i)\right| + \frac{1}{\sqrt{T}} \tag{56}$$

$$\leq \sum_{k=0}^{H} \sum_{x,i} \gamma^k \left|\Big[d_k(x)\pi_{t+k}(i|x) - d'_k(x)\pi_t(i|x)\Big] r(x,i)\right| + \frac{1}{\sqrt{T}} \tag{57}$$

by the assumption $r(s,a) \in [0,1]$

$$\leq \sum_{k=0}^{H} \sum_{x,i} \gamma^k \Big| d_k(x)\pi_{t+k}(i|x) - d'_k(x)\pi_t(i|x) \Big| + \frac{1}{\sqrt{T}} \tag{58}$$

by the fact $|a_1 b_1 - a_2 b_2| \leq |a_1 - a_2| + |b_1 - b_2|$ for $a_i, b_i \in [0,1]$

$$\leq \sum_{k=0}^{H} \sum_{x,i} \gamma^k \Big\{ \Big| d_k(x) - d'_k(x) \Big| + \Big| \pi_{t+k}(i|x) - \pi_t(i|x) \Big| \Big\} + \frac{1}{\sqrt{T}} \tag{59}$$

$$\leq \sum_{k=0}^{H} \gamma^k \Big\{ \sum_{x} \sum_{i} \Big| d_k(x) - d'_k(x) \Big| + \sum_{x} \max_{x} \sum_{i} \Big| \pi_{t+k}(i|x) - \pi_t(i|x) \Big| \Big\} + \frac{1}{\sqrt{T}} \tag{60}$$

$$= \sum_{k=0}^{H} \gamma^k \Big\{ A||d_k - d'_k||_1 + S||\pi_{t+k} - \pi_t||_{1,\infty} \Big\} + \frac{1}{\sqrt{T}} \tag{61}$$

given $d_k = d_s(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_{t+k-1}})$ and $d'_k = d_s(\mathbb{P}^{\pi_t})^k$

$$= \sum_{k=0}^{H} \gamma^k \Big\{ A \Big|\Big| d_s(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_{t+k-1}}) - d_s(\mathbb{P}^{\pi_t})^k \Big|\Big|_1 + S||\pi_{t+k} - \pi_t||_{1,\infty} \Big\} + \frac{1}{\sqrt{T}} \tag{62}$$

follow Eq.(53) and Eq. (54)

$$\leq \sum_{k=0}^{H} \gamma^k (Skc_T + Ak^2 c_T) + \frac{1}{\sqrt{T}} \tag{63}$$

by $k \leq H$ and the sum of geometric sequence

$$\leq \frac{HS + H^2 A}{1-\gamma} c_T + \frac{1}{\sqrt{T}} \tag{64}$$

■

## Appendix F. Proof of Lemma 10

**Lemma 10** *If* MAIN *is slowly changing with a non-increasing rate of $c_T$, then*

$$|Q_{t+n}(s,a) - Q_t(s,a)| \leq \frac{1}{1-\gamma} (S + HA)\, nc_T + \frac{2}{\sqrt{T}} \tag{65}$$

**Proof:** Recall

> **Corollary 31** *Let $\pi_t$ be slowly changing with non-increasing rate $c_T$, then we have $||\pi_{t+l} - \pi_t||_{1,\infty} \leq lc_T$. Apply Lemma 30 with $\pi_k = \pi_{t+n}$, $\pi'_k = \pi_t$ and $d = d'$, then*
>
> $$||d(\mathbb{P}^{\pi_{t+n}})^K - d(\mathbb{P}^{\pi_t})^K||_1 \leq Knc_T \tag{66}$$

By Lemma 26 and corollary 31, we have

$$||\pi_{t+n} - \pi_t||_{1,\infty} \leq nc_T \tag{67}$$

$$||d_s(\mathbb{P}^{\pi_{t+n}})^k - d_s(\mathbb{P}^{\pi_t})^k||_1 \leq nkc_T \tag{68}$$

$$\left|Q_{t+n}(s,a) - Q_t(s,a)\right| \leq \left|\bar{Q}_{t+n}(s,a) - \bar{Q}_t(s,a)\right| + \left|\bar{Q}_{t+n}(s,a) - Q_{t+n}(s,a)\right| + \left|\bar{Q}_t(s,a) - Q_t(s,a)\right| \tag{69}$$

by Corollary 36, we have

$$\leq \Big|\sum_{k=0}^{H}\sum_{x\in\mathcal{S}}\sum_{i\in\mathcal{A}}\gamma^k\Big[d'_k(x)\pi_{t+n}(i|x) - d_k(x)\pi_t(i|x)\Big]r(x,i)\Big| + \frac{2}{\sqrt{T}} \tag{70}$$

$$\leq \sum_{k=0}^{H}\sum_{x}\sum_{i}\gamma^k\Big|\Big[d'_k(x)\pi_{t+n}(i|x) - d_k(x)\pi_t(i|x)\Big]r(x,i)\Big| + \frac{2}{\sqrt{T}} \tag{71}$$

$r(x,i) \in [0,1]$

$$\leq \sum_{k=0}^{H}\sum_{x}\sum_{i}\gamma^k\Big|d'_k(x)\pi_{t+n}(i|x) - d_k(x)\pi_t(i|x)\Big| + \frac{2}{\sqrt{T}} \tag{72}$$

by the fact $|a_1b_1 - a_2b_2| \leq |a_1 - a_2| + |b_1 - b_2|$ for $a_i, b_i \in [0,1]$

$$\leq \sum_{k=0}^{H}\sum_{x}\sum_{i}\gamma^k\Big\{\Big|d'_k(x) - d_k(x)\Big| + \Big|\pi_{t+n}(i|x) - \pi_t(i|x)\Big|\Big\} + \frac{2}{\sqrt{T}} \tag{73}$$

$$\leq \sum_{k=0}^{H}\gamma^k\Big\{\sum_{x}\sum_{i}\Big|d'_k(x) - d_k(x)\Big| + \sum_{x}\max_{x}\sum_{i}\Big|\pi_{t+n}(i|x) - \pi_t(i|x)\Big|\Big\} + \frac{2}{\sqrt{T}} \tag{74}$$

$$= \sum_{k=0}^{H}\gamma^k\Big\{A||d'_k - d_k||_1 + S||\pi_{t+n} - \pi_t||_{1,\infty}\Big\} + \frac{2}{\sqrt{T}} \tag{75}$$

$$= \sum_{k=0}^{H}\gamma^k\Big\{A\Big|\Big|d_s(\mathbb{P}^{\pi_{t+n}})^k - d_s(\mathbb{P}^{\pi_t})^k\Big|\Big|_1 + S||\pi_{t+n} - \pi_t||_{1,\infty}\Big\} + \frac{2}{\sqrt{T}} \tag{76}$$

$$\leq \sum_{k=0}^{H}\gamma^k\Big(nS + nkA\Big)c_T + \frac{2}{\sqrt{T}} \tag{77}$$

by $k \leq H$ and the sum of geometric sequences

$$\leq \frac{1}{1-\gamma}\left(S + HA\right)nc_T + \frac{2}{\sqrt{T}} \tag{78}$$

■

## Appendix G. Proof of Lemma 11

We follow the proof by Even-Dar et al. (2009); Neu et al. (2010), to give a lemma that $\nu_t$ tracks the stationary distribution $\mu_t$ slowly

**Lemma 11** *If the sequence of policies $\{\pi_1, \dots \pi_t\}$ is slowly changing with rate $c_T$, then $||\nu_t - \mu_t||_1 \leq \tau(\tau+1)c_T + 2e^{-(t-1)/\tau}$*

**Proof.** Suppose $k+1 \leq t$

$$||\nu_{k+1} - \mu_t||_1 = ||\nu_k \mathbb{P}^{\pi_k} - \nu_k \mathbb{P}^{\pi_t} + \nu_k \mathbb{P}^{\pi_t} - \mu_t \mathbb{P}^{\pi_t}||_1 \tag{79}$$

$$\leq ||\nu_k \mathbb{P}^{\pi_k} - \nu_k \mathbb{P}^{\pi_t}||_1 + ||\nu_k \mathbb{P}^{\pi_t} - \mu_t \mathbb{P}^{\pi_t}||_1 \tag{80}$$

by Assumption 2

$$\leq ||\nu_k \mathbb{P}^{\pi_k} - \nu_k \mathbb{P}^{\pi_t}||_1 + e^{-1/\tau}||\nu_k - \mu_t||_1 \tag{81}$$

by Lemma 27

$$\leq (t-k)c_T + e^{-1/\tau}||\nu_k - \mu_t||_1 \tag{82}$$

Then, by expanding the recursion we have

$$||\nu_t - \mu_t||_1 \leq c_T \sum_{k=1}^{t-1}(t-k)e^{-(t-k-1)/\tau} + e^{-(t-1)/\tau}||\nu_1 - \mu_t||_1 \tag{83}$$

by $||\nu_1 - \mu_t||_1 \leq 2$

$$\leq c_T \sum_{k=1}^{t-1}(t-k)e^{-(t-k-1)/\tau} + 2e^{-(t-1)/\tau} \tag{84}$$

notice that $\sum_{k=1}^{t-1}(t-k)e^{-(t-k-1)/\tau} \leq \int_0^\infty (k+1)e^{-k/\tau}dk = \tau^2$

$$\leq \tau(\tau+1)c_T + 2e^{-(t-1)/\tau} \tag{85}$$

■

## Appendix H. Proof of Lemma 13 (Full-Span Regret)

We first give two technical lemmas, by consider a single sticky bandit

**Lemma 37** *Let $X_i$ be the timestep of $i$-th time that a sticky bandit could react, and suppose each time $t$, the probability of the bandit could react is at least $\beta$, while assuming each draw is independent*

$$\mathbb{E}_{X_{i+1}|X_i}[X_{i+1} - X_i] \leq 1/\beta \tag{86}$$

**Proof.** As $X_{i+1} - X_i \geq 1$

$$\mathbb{E}_{X_{i+1}|X_i}(X_{i+1} - X_i) \leq \beta \times 1 + (1-\beta)\Big[\mathbb{E}_{X_{i+1}|X_i}(X_{i+1} - X_i) + 1\Big] \tag{87}$$

$$\Rightarrow \tag{88}$$

$$\mathbb{E}_{X_{i+1}|X_i}(X_{i+1} - X_i) \leq 1/\beta \tag{89}$$

■

**Lemma 38** *Let $X_i$ be the timestep of $i$-th time that a sticky bandit could react, and suppose each time $t$, the probability of the bandit could react is at least $\beta$, while assuming each draw is independent*

$$\mathbb{E}_{X_{i+1}|X_i}(X_{i+1} - X_i)^2 \leq \frac{2}{\beta^3} \tag{90}$$

**Proof.**

$$\mathbb{E}_{X_{i+1}|X_i}(X_{i+1} - X_i)^2 = \sum_{t=X_i+1}^{\infty} \Pr(X_{i+1} = t)(t - X_i)^2 \tag{91}$$

as $\Pr(X_{i+1} = t) \geq \beta$ for all $t$

$$\leq \sum_{t=X_i+1}^{\infty} \left[(1-\beta)^{t-X_i-1} \times 1\right](t - X_i)^2 \tag{92}$$

let $k = X_i + 1$, $q = 1 - \beta$ and given $X_i > 0$, we have

$$= \sum_{t=k}^{\infty} q^{t-k}(t - k + 1)^2 \tag{93}$$

$$= \sum_{t=0}^{\infty} q^{t}(t + 1)^2 \tag{94}$$

let $S_n = \sum_{t=0}^{n} q^t(t+1)^2$

$$= \lim_{n\to\infty} S_n \tag{95}$$

$$= \lim_{n\to\infty} \frac{1-q}{1-q} S_n \tag{96}$$

observe that $S_n - qS_n = 1 - q^{n+1}(n+1)^2 + \sum_{t=1}^{n}(2t+1)q^t$

$$= \frac{1}{1-q} \lim_{n\to\infty} \left[1 - q^{n+1}(n+1)^2 + \sum_{t=1}^{n}(2t+1)q^t\right] \tag{97}$$

$$= \frac{1}{1-q} \lim_{n\to\infty} \sum_{t=0}^{n}(2t+1)q^t \tag{98}$$

where $\sum_{t=0}^{n}(2t+1)q^t$ is an arithmetico-geometric series, by the sequence sum of arithmetico–geometric series

$$\leq \frac{\beta^2 - 3\beta + 2}{(1-\beta)\beta^3} \leq \frac{2}{\beta^3} \tag{99}$$

■

Now we are ready to bound the full-span regret by the observed regret.

**Lemma 13** *If Assumption 1 and 2 are satisfied, and the* LOCAL *learner is slowly changing with a rate $c_T$, we have*

$$\tilde{R}^{\textit{un-mdp}}(T) \leq \frac{\tilde{R}^{\textit{ob-mdp}}(T)}{\beta} + \frac{S + HA}{\beta^3(1-\gamma)} c_T T + 4S\sqrt{T}. \tag{100}$$

**Proof.** Recall that the definitions of observed/unobserved regret in MDPs are,

$$\tilde{R}^{\text{ob-mdp}} \coloneqq \sum_{s\in\mathcal{S}} \sum_{t=1}^{T} \mu_t \langle \pi_s^* - \pi_{t,s}, Q_{t,s} \rangle \tag{101}$$

$$= \sum_{s\in\mathcal{S}} \mathbb{E}_{\{Y_i^s\}} \sum_{t\in\{Y_i^s\}} \langle \pi_s^* - \pi_{t,s}, Q_{t,s} \rangle \tag{102}$$

$$\tilde{R}^{\text{un-mdp}} \coloneqq \sum_{s\in\mathcal{S}} \sum_{t=1}^{T} (1-\mu_t) \langle \pi_s^* - \pi_{t,s}, Q_{t,s} \rangle \tag{103}$$

$$= \sum_{s\in\mathcal{S}} \mathbb{E}_{\{Y_i^s\}} \sum_{t\notin\{Y_i^s\}} \langle \pi_s^* - \pi_{t,s}, Q_{t,s} \rangle \tag{104}$$

where $Y_i^s$ is a random variable that stands for the time step $t$ the bandit at $s$ was allowed to pull, and $\pi_{t,s} \coloneqq \pi_t(\cdot|s)$ and $Q_{t,s} \coloneqq Q_t(s,\cdot)$ denote vectors corresponding to state $s$ for the sake of space.

One could divide the sequence into segments $(Y_i^s, Y_{i+1}^s)$, and without loss of generality, we assume bandit at $s$ is pulled $N_s$ times in total, define $Y_0^s = 1$

$$\tilde{R}^{\text{un-mdp}} = \sum_s \mathbb{E}_{\{Y_i^s\}} \sum_{i=0}^{N_s} \sum_{t\in(Y_i^s, Y_{i+1}^s)} \Big[ \langle \pi_s^* - \pi_{t,s}, Q_{t,s} \rangle \Big] \tag{105}$$

because of the sticky setting

$$= \sum_s \mathbb{E}_{\{Y_i^s\}} \sum_{i=0}^{N_s} \sum_{t\in(Y_i^s, Y_{i+1}^s)} \Big[ \langle \pi_s^* - \pi_{Y_i^s,s}, Q_{t,s} \rangle \Big] \tag{106}$$

given $|Q_{t+n}(s,a) - Q_t(s,a)| \leq \frac{S+HA}{1-\gamma} n c_T + \frac{2}{\sqrt{T}}$, let $b(n) = \frac{S+HA}{1-\gamma} n c_T$

$$= \sum_s \mathbb{E}_{\{Y_i^s\}} \sum_{i=0}^{N_s} \sum_{t\in(Y_i^s, Y_{i+1}^s)} \Big[ \langle \pi_s^* - \pi_{Y_i^s,s}, Q_{Y_i^s,s} \rangle + 2b(t - Y_i^s) \Big] + \sum_s \sum_{t=1}^{T} \frac{4}{\sqrt{T}} \tag{107}$$

by the chain rule

$$= \sum_s \mathbb{E}_{Y^s_{N_s}|Y^s_{N_s-1},\ldots,Y^s_0} \ldots \mathbb{E}_{Y^s_1|Y^s_0} \sum_{i=0}^{N_s} \sum_{t\in(Y^s_i,Y^s_{i+1})} \left[\langle \pi^*_s - \pi_{Y^s_i,s}, Q_{Y^s_i,s}\rangle + 2b(t-Y^s_i)\right] + 4S\sqrt{T} \tag{108}$$

as the expected length of $[Y^s_i, Y^s_{i+1})$ is independent of $Y^s_j$ conditioned on $Y^s_i$ for $j > i+1$

$$= \sum_s \sum_i \mathbb{E}_{Y^s_{i+1}|Y^s_i,\ldots,Y^s_0} \sum_{t\in(Y^s_i,Y^s_{i+1})} \left[\langle \pi^*_s - \pi_{Y^s_i,s}, Q_{Y^s_i,s}\rangle + 2b(t-Y^s_i)\right] + 4S\sqrt{T} \tag{109}$$

because of the Markov property

$$= \sum_s \sum_i \mathbb{E}_{Y^s_{i+1}|Y^s_i} \sum_{t\in(Y^s_i,Y^s_{i+1})} \left[\langle \pi^*_s - \pi_{Y^s_i,s}, Q_{Y^s_i,s}\rangle + 2b(t-Y^s_i)\right] + 4S\sqrt{T} \tag{110}$$

by Lemma 37

$$\leq \sum_s \left[ \sum_i \frac{\langle \pi^*_s - \pi_{Y^s_i,s}, Q_{Y^s_i,s}\rangle}{\beta} + \sum_i \mathbb{E}_{Y^s_{i+1}|Y^s_i} \sum_{t\in(Y^s_i,Y^s_{i+1})} 2b(t-Y^s_i)\right] + 4S\sqrt{T} \tag{111}$$

$$= \sum_s \left[ \sum_i \frac{\langle \pi^*_s - \pi_{Y^s_i,s}, Q_{Y^s_i,s}\rangle}{\beta} + \sum_i \mathbb{E}_{Y^s_{i+1}|Y^s_i} \sum_{t\in(Y^s_i,Y^s_{i+1})} 2\frac{S+HA}{1-\gamma}(t-Y^s_i)c_T\right] + 4S\sqrt{T} \tag{112}$$

$$= \sum_s \left[ \sum_i \frac{\langle \pi^*_s - \pi_{Y^s_i,s}, Q_{Y^s_i,s}\rangle}{\beta} + \frac{S+HA}{1-\gamma} \sum_i \mathbb{E}_{Y^s_{i+1}|Y^s_i}(Y^s_{i+1}-Y^s_i)^2 c_T\right] + 4S\sqrt{T} \tag{113}$$

by Lemma 38

$$\leq \sum_s \left[ \sum_i \frac{\langle \pi^*_s - \pi_{Y^s_i,s}, Q_{Y^s_i,s}\rangle}{\beta} + \frac{S+HA}{1-\gamma} \sum_i \frac{2}{\beta^3} c_T\right] + 4S\sqrt{T} \tag{114}$$

$\sum_i \langle \pi^*_s - \pi_{Y^s_i,s}, Q_{Y^s_i,s}\rangle$ is the observed regret at $s$, and summing over $s$ and $i$ results in $T$ steps

$$= \frac{\tilde{R}^{\text{ob-mdp}}(T)}{\beta} + \frac{2(S+HA)}{\beta^3(1-\gamma)} c_T T + 4S\sqrt{T} \tag{115}$$

■

## Appendix I. Proof of Lemma 18

We first show that the slowly changing property of LOCAL is preserved by MAIN.

**Lemma 39** MAIN *is slowly changing if* LOCAL *is slowly changing.*

**Proof.** We use $\phi^s$ to denote the *policy* of $\text{LOCAL}_s$ (to be distinguished from $\pi$ defined over $\mathcal{S} \times \mathcal{A}$). The *state space* of $\phi^s$ is the singleton $\{s\}$. We have $||\phi^s_{t+1} - \phi^s_t||_{1,\infty} \leq c_T$ because $\text{LOCAL}_s$ is slowly changing.

$$||\pi_{t+1} - \pi_t||_{1,\infty} = ||\phi^{s_{t-H}}_{t+1} - \phi^{s_{t-H}}_t||_{1,\infty} \leq c_T$$

This simply follows the fact that only $\text{LOCAL}_{s_{t-H}}$ is updated at time $t$. ■

**Lemma 18** *While applying Algo. 2 as* LOCAL, *$\tilde{\pi}$ is slowly changing in $\tilde{\mathcal{M}}$, where $\tilde{\pi}_t(a|s \circ h) \coloneqq \pi_t(a|s)$ and $\pi_t$ is produced by* MAIN.

**Proof.** Similar to Lemma 39, we use $\phi^s$ to denote the *policy* of $\text{LOCAL}_s$. In addition, we use $\phi^{s,h}_t$ to denote the *policy* of $\text{BASE}^h_s$ of $\text{LOCAL}_s$ at time $t$. We have $||\phi^{s,h}_{t+1} - \phi^{s,h}_t||_{1,\infty} \leq c_T$, given the slowly changing BASE assumption.

$$||\tilde{\pi}_{t+H+1} - \tilde{\pi}_{t+H}||_{1,\infty} = \max_{s \circ h} \sum_a |\tilde{\pi}_{t+H+1}(a|s \circ h) - \tilde{\pi}_{t+H}(a|s \circ h)| \tag{116}$$

by the fact that only $\text{BASE}^{h_t}_{s_t}$ is updated at $t + H$

$$= \sum_a |\tilde{\pi}_{t+H+1}(a|s_t \circ h_t) - \tilde{\pi}_{t+H}(a|s_t \circ h_t)| \tag{117}$$

$$= \sum_a |\phi^{s_t,h_t}_{t+H+1} - \phi^{s_t,h_t}_{t+H}| \tag{118}$$

$$= ||\phi^{s_t,h_t}_{t+H+1} - \phi^{s_t,h_t}_{t+H}||_{1,\infty} \leq c_T \tag{119}$$

■

## Appendix J. Proof of Lemma 19 (Assumptions Hold for $\tilde{\mathcal{M}}$)

By definition of $\tilde{\mathcal{M}}$, we have

$$\tilde{\pi}(a|s \circ h) \coloneqq \pi(a|s) \tag{120}$$

$$\tilde{\mathbb{P}}^{\tilde{\pi}} \coloneqq \begin{array}{c} \\ \mathcal{S} \circ 1 \\ \mathcal{S} \circ 2 \\ \vdots \\ \mathcal{S} \circ (H+1) \end{array} \begin{array}{c} \begin{array}{cccc} \mathcal{S} \circ 1 & \mathcal{S} \circ 2 & \cdots & \mathcal{S} \circ (H+1) \end{array} \\ \begin{pmatrix} & \mathbb{P}^\pi & & \\ & & \ddots & \\ & & & \mathbb{P}^\pi \\ \mathbb{P}^\pi & & & \end{pmatrix} \end{array} \tag{121}$$

where $\mathcal{S} \circ h$ stands for concatenating all elements in the set $\mathcal{S}$ with $h$.

For brevity, we use $\tilde{\mathbb{P}}^\pi$ instead of $\tilde{\mathbb{P}}^{\tilde{\pi}}$ as $\pi$ is sufficient to avoid confusion.

**Lemma 40** *(Stationary distribution $\tilde{\mu}^\pi$) For any $\mathcal{H}$-augmented MDP $\tilde{\mathcal{M}}$, if the MDP $\mathcal{M}$ before augmentation satisfies assumption 2, then there is an unique stationary distribution*

$$\tilde{\mu}^\pi = \frac{1}{H+1}\left[\mu^\pi, \dots, \mu^\pi\right] \tag{122}$$

**Proof.**
**(1) Existence:**

Let $x_h$ be an row vector (with size $1 \times (H+1)S$ ) that is defined as and $\mathbf{0}$ be size $1 \times S$

$$x_h \coloneqq [\underbrace{\mathbf{0}, \mathbf{0} \dots}_{h-1 \text{ zero vectors}}, \mu^\pi \dots, \mathbf{0}] \tag{123}$$

Consider a convex combination of $\{x_h : h = 1, \dots, H+1\}$ multiplied by $\tilde{\mathbb{P}}^\pi$

$$\left(\sum_h \alpha_h x_h\right)\tilde{\mathbb{P}}^\pi = \sum_h \left(\left[\underbrace{\mathbf{0}, \mathbf{0} \dots}_{h-1 \text{ zero vectors}}, \alpha_h \mu^\pi \dots, \mathbf{0}\right]\tilde{\mathbb{P}}^\pi\right) \tag{124}$$

as density at block $h$ always be pushed to block $h+1$

$$= \sum_h \left(\left[\underbrace{\mathbf{0}, \mathbf{0} \dots}_{h \text{ zero vectors}}, \alpha_h \mu^\pi \dots, \mathbf{0}\right]\right) \tag{125}$$

This equality implies that

$$[\alpha_1 \mu^\pi, \alpha_2 \mu^\pi, \dots \alpha_{H+1}\mu^\pi] = [\alpha_{H+1}\mu^\pi, \alpha_1 \mu^\pi, \dots \alpha_H \mu^\pi] \tag{126}$$

which implies

$$\alpha_h = \frac{1}{H+1} \qquad \text{for } h = 1, 2, \dots, H+1 \tag{127}$$

Therefore,

$$\tilde{\mu}^\pi = \frac{1}{H+1}\left[\mu^\pi, \dots, \mu^\pi\right] \tag{128}$$

**(2) Uniqueness:**

Suppose there exists a row vector $d$ such that $d\tilde{\mathbb{P}}^\pi = d$ and $d \neq \tilde{\mu}^\pi$. Let divide $d$ into $H+1$ blocks as well

$$d = [d_1, d_2, \dots, d_{H+1}] \tag{129}$$

Multiplying $d$ with the transition kernel $\tilde{\mathbb{P}}^\pi$

$$d\tilde{\mathbb{P}}^\pi = [d_1, d_2, \dots, d_{H+1}]\tilde{\mathbb{P}}^\pi \tag{130}$$

$$= [d_1, d_2, \dots, d_{H+1}] \begin{pmatrix} & \mathbb{P}^\pi & & \\ & & \ddots & \\ & & & \mathbb{P}^\pi \\ \mathbb{P}^\pi & & & \end{pmatrix} \tag{131}$$

$$= [d_{H+1}\mathbb{P}^\pi, d_1\mathbb{P}^\pi, \dots, d_H\mathbb{P}^\pi] \tag{132}$$

which implies

$$d_h \mathbb{P}^\pi = d_{h+1} \tag{133}$$

the stationary distribution of $\mathbb{P}^\pi$ is unique, i.e. $\mu^\pi$, and $\|d\|_1 = 1$, which implies

$$d_h = \frac{1}{H+1}\mu^\pi \tag{134}$$

which contradicts our assumption that $d \neq \tilde{\mu}^\pi$, therefore $\tilde{\mu}^\pi$ is the unique stationary distribution of $\tilde{\mathbb{P}}^\pi$. ■

**Corollary 41** *For any $\mathcal{H}$-augmented MDP $\tilde{\mathcal{M}}$, if the MDP $\mathcal{M}$ before augmentation satisfies assumption 1 and assumption 2, then the stationary distributions $\tilde{\mu}^\pi(s)$ are uniformly bounded away from zero,*

$$\inf_{\pi,s} \tilde{\mu}^\pi(s) \geq \frac{\beta}{H+1} \text{ for some } \beta > 0. \tag{135}$$

**Proof.** Trivially implied by Lemma 40. ■

**Lemma 42** *If assumption 2 holds, then for any two arbitrary distributions $\tilde{d}$ and $\tilde{d}'$ over $\tilde{\mathcal{S}}$, we have*

$$\sup_\pi \|(\tilde{d} - \tilde{d}')\tilde{\mathbb{P}}^\pi\|_1 \leq e^{-1/\tau}\|\tilde{d} - \tilde{d}'\|_1,$$

*where $\tau$ is the same as in assumption 2.*

**Proof.**

Let $X_{:,j}$ be the $j$-th column of matrix $X$

$$\sup_\pi \|(\tilde{d} - \tilde{d}')\tilde{\mathbb{P}}^\pi\|_1 = \sup_\pi \sum_{s\circ h \in \tilde{\mathcal{S}}} \left|(\tilde{d} - \tilde{d}')\tilde{\mathbb{P}}^\pi_{:,s\circ h}\right| \tag{136}$$

As $\tilde{\mathbb{P}}^\pi_{:,s\circ h}$ is in the form of $[0, \ldots, (\mathbb{P}^\pi_{:,s})^\top, \ldots, 0]^\top$, and let $y_h$ be the $h$-th block of row vector $y$, if one divide $y$ into $H+1$ equal blocks (where $H+1$ comes from our construction in Algorithm 2)

$$\leq \sup_\pi \sum_h \sum_s \left|(\tilde{d} - \tilde{d}')_h \mathbb{P}^\pi_{:,s}\right| \tag{137}$$

By $\sup_\pi \sum_h \leq \sum_h \sup_\pi$

$$\leq \sum_h \sup_\pi \sum_s \left|(\tilde{d} - \tilde{d}')_h \mathbb{P}^\pi_{:,s}\right| \tag{138}$$

By assumption 2

$$\leq \sum_h e^{-1/\tau}\|(\tilde{d} - \tilde{d}')_h\|_1 \tag{139}$$

$$= e^{-1/\tau}\|\tilde{d} - \tilde{d}'\|_1 \tag{140}$$

■

## Appendix K. Proof of Lemma 24 (EXP3 is Slowly-Changing)

We first give Algorithm 3 to provide relevant notations.

**Algorithm 3:** EXP3

**Require:** $\gamma \in [0,1)$, $\eta_T \in (0,1]$, $A = |\mathcal{A}|$
**Initialize:** $w_1(a) = 1$ for $a \in \mathcal{A}$
**for** $t = 1, 2, \ldots, T$ **do**
  Set $W_t = \sum_{a=1}^{A} w_t(a)$
    $p_t(a) = (1-\eta_T) w_t(a)/W_t + \eta_T/A$
  Draw $a_t$ randomly accordingly to $\mathbf{p}_t$
  Receive reward $y_t(a_t) \in [0, \frac{1}{1-\gamma}]$
  For $a = 1, 2, \ldots, A$, set

$$\hat{y}_t(a) = \begin{cases} y_t(a)/p_t(a), \text{if } a = a_t \\ 0, \text{otherwise} \end{cases}$$

$$w_{t+1}(a) = w_t(a) \exp((1-\gamma)\eta_T \hat{y}_t(a)/A)$$

**end**

**Lemma 24** EXP3 *is slowly changing with a rate of* $\mathcal{O}(\eta_T/A)$*, assuming the feedback* $y_t$ *is bounded within the range* $[0, 1/(1-\gamma)]$.

**Proof.** We observe it is sufficient to bound $p_{t+1}(a) - p_t(a)$ of the action $a$ chosen by the algorithm at time-step $t$. We then fix an arbitrary action $a$ to be chosen (and whose weight is updated) and drop it from the notation below w.r.t. $p$, $w$, $\hat{y}$, etc.

$$p_{t+1} - p_t = (1-\eta_T)\left(\frac{w_{t+1}}{W_{t+1}} - \frac{w_t}{W_t}\right) \tag{141}$$

$$= (1-\eta_T)\left(\frac{w_t e^{(1-\gamma)\eta_T \hat{y}_t/A}}{W_t + w_t(e^{\eta_T \hat{y}_t/A} - 1)} - \frac{w_t}{W_t}\right) \tag{142}$$

$$\leq (1-\eta_T)\left(\frac{w_t e^{(1-\gamma)\eta_T \hat{y}_t/A}}{W_t} - \frac{w_t}{W_t}\right) \tag{143}$$

$$= (1-\eta_T)\left(\frac{w_t(e^{(1-\gamma)\eta_T \hat{y}_t/A} - 1)}{W_t}\right) \tag{144}$$

$$\leq (1-\eta_T)\left(2(1-\gamma)\left(\frac{\eta_T \hat{y}_t}{A}\right)\left(\frac{w_t}{W_t}\right)\right) \tag{145}$$

$$\leq (1-\eta_T)\left(2\left(\frac{\eta_T}{A p_t}\right)\left(\frac{w_t}{W_t}\right)\right) \tag{146}$$

$$\leq 2\eta_T/A. \tag{147}$$

(145) follows from that $e^x - 1 < 2x$ for $0 \leq x \leq 1$ and (147) follows from $p_t \geq (1-\eta_T)(w_t/W_t)$.

As mentioned in Section 6, to achieve $\tilde{\mathcal{O}}(\sqrt{AT})$ regret, EXP3 is run with a learning rate of $\eta_T = \tilde{\mathcal{O}}(\sqrt{A/T})$, which means it is slowly changing with a rate of $c_T = \tilde{\mathcal{O}}(\sqrt{1/AT})$.