slowly changing adversarial bandit algorithms are provably efficient for discounted MDPs

ian a. kash1, lev reyzin2 and zishun yu1

1 department of computer science, university of illinois at chicago
2 department of mathematics, statistics, and computer science, university of
illinois at chicago
{iankash, lreyzin, zyu32}@uic.edu

may 19, 2022

abstract

reinforcement learning (RL) generalizes bandit problems with additional difficulties on longer planning horizon and unknown transition kernel. we show that, under some mild assumptions, any slowly changing adversarial bandit algorithm enjoys near-optimal regret in adversarial bandits can achieve near-optimal (expected) regret in non-episodic discounted MDPs. The slowly changing property required by our generalization is mild, see e.g. (Even-Dar et al. 2009, Neu et al. 2010), we also show, for example, Exp3 (Auer et al. 2002) is slowly changing and enjoys near-optimal regret in MDPs.

1 introduction

Reinforcement learning (RL) and multi-armed bandits (MAB) are long-standing models that formalize sequential decision-making problems. RL generalizes MAB by modeling an environment that can change responding to the learner’s actions; this environment is modeled by an underlying Markov decision process (MDP).

As a generalization problem of MAB, RL has been considered a much harder problem partially due to 1) the unknown dynamics of MDP and 2) the goal of maximizing long-term return instead of immediate reward. However, recent advances in RL theory show that, RL methods can approach the optimal rate of $\Omega(\sqrt{T})$ regret, which matches the lower bounds in the MAB setting.

If RL is not necessarily a harder problem than bandits, we therefore ask the reverse question: can we convert arbitrary MAB algorithms to RL algorithms while preserving their guarantees, under some mild assumptions? Or, in other words: is there a reduction from reinforcement learning to multiarmed bandits? That is, can one trivially place a set of distinct bandit algorithms and have the set of bandits (as a whole) achieve sub-linear regret in MDPs without acquiring information from its co-learners except the shared global rewards? The answer turns out to be yes.

We prove that one could place $\tilde{O}(S)$ arbitrary slowly changing adversarial bandit algorithms to achieve $\tilde{O}\left(\frac{1}{\sqrt{1-\gamma}}\left(S^2 + SA\right)\sqrt{T}\right)$ regret under some relatively mild assumptions, see Section 2.1. Despite the black-box nature of the reduction, this achieves regret which is optimal in its dependence on $T$ (up to log factors). Furthermore, in contrast to much prior work, we achieve this in the non-episodic, discounted-reward setting.
1.1 Prior Work

As previously mentioned, a range of approaches have achieved regrets in RL settings whose dependence on $T$ is $\tilde{O}(\sqrt{T})$. Let $S = |S|$ and $A = |A|$ be the size of state and action space, respectively. Let $K$ denotes the number of episodes, $E$ to be the episode length, and $\tilde{O}(\cdot)$ omits the polylog factor in $T$. $\tilde{O}(\sqrt{E^2SAT + \text{poly}(ESA)})$ regret is enjoyed by UCB-VI Azar et al. [2017], $\tilde{O}(\sqrt{ESA})$ by UCB-Q-Hoeffding Jin et al. [2018], $\tilde{O}(\sqrt{E^2SAT + \text{poly}(ESA)})$ by UCB-M-Q Ménard et al. [2021] and Q-EarlySettled-Advantage Li et al. [2021], etc. Zhang et al. [2021] shows RL is not necessarily a more difficult problem by proposing a method that enjoys $O((\sqrt{SAK} + S^2A)\text{poly log}(SAEK))$ which escapes from the polynomial factor in $E$. Unlike these works our goal is not to provide the best regret bound possible but instead to provide a black-box reduction. Furthermore, we work in an infinite horizon setting rather than an episodic one.

Our technical approach builds on the analysis of online MDPs. The online MDP problem is originally defined by Even-Dar et al. [2009], where the key difference between online MDPs and MDPs is that the reward function is chosen by an adversary. In the case where the algorithm knows the MDP dynamics, Even-Dar et al. [2009] show that expert algorithms can run efficiently in such a adversarial reward MDP setting with full information feedback. Neu et al. [2010] then extend this result to the bandit feedback setting. In contrast to these works, we restricts to MDPs, address the more complex discounted reward rather than the average reward setting (which includes using a different notion of regret), provide a black-box reduction rather than relying on the a customized algorithm, and have a much lower computational complexity due to not requiring knowledge of the MDP dynamics. While Neu et al. [2010] also use delay-tolerant bandits, the purpose of the delay and the structure of feedback provided to the bandits is quite different. Some more recent work has examined online MDPs with bandit feedback and unknown dynamics Rosenberg and Mansour [2019], Jin et al. [2020]. These focus on the episodic setting and are less related to our approach.

A line of recent work on non-episodic discounted MDPs has measured the performance of algorithms using variants of a regret definition proposed by Liu and Su [2020] rather than sample complexity (although the two are related). Liu and Su [2020] shows that double Q-learning has $O(\sqrt{SAT}/(1 - \gamma)^{2.5})$ regret. He et al. [2020] proposed UCBVI-$\gamma$ that achieves $O(\sqrt{SAT}/(1 - \gamma)^{1.5})$ regret with a known reward function. Zhou et al. [2021b] achieves $O(\sqrt{SAT}/(1 - \gamma)^2)$ regret with a known reward function and feature mapping. Zhou et al. [2021a] use a version of this regret in an episodic setting. We adopt this notion of regret (specifically the version used by He et al. [2020]) and show that our method MAIN enjoys $\tilde{O}\left(\frac{\sqrt{T}}{(1+\gamma)(1-\gamma)}(S^2 + SA)\sqrt{T}\right)$ regret, without knowledge of the dynamics or rewards. Rather than a tight bound our key contribution is showing that any efficient, slowly changing adversarial bandit algorithm can be used in a black-box manner while being asymptotically optimal in terms of $T$.

The closest work, in terms of goal instead of technical approach, is Cheng et al. [2020], where the authors propose a reduction from RL to online learning while having a generative model. They do so by reformulating the primal-dual form of RL into a saddle point optimization and consider one player as an online learning problem. Our reduction is quite different and doesn’t require a generative model.
2 Preliminaries

2.1 Assumptions

We consider finite, discounted, infinite horizon MDPs with all rewards in $[0, 1]$. In addition, we make two assumptions that guarantee the MDP is "well behaved" in the sense that all states are likely to be visited often regardless of the policy and starting point. These assumptions have been used in prior work on online learning in MDPs such as that of Neu et al. [2010].

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

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

**Assumption 2.** There exists some fixed positive $\tau$ such that for any two arbitrary distributions $\mu$ and $\mu'$ over $S$,

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

where $\tau$ is the mixing time, we further assume $\tau \geq 1$.

2.2 MDP and Regret

We consider infinite-horizon discounted MDPs, denoted by a 5-tuple $(S, A, P, r, \gamma)$, where $S$ and $A$ are finite state and finite action space, respectively. Let $\Delta(X)$ be all probability distributions over space $X$, $P : S \times A \to \Delta(S)$ is the unknown stochastic transition function, a.k.a dynamics, $r : S \times A \to \Delta([0, 1])$ is the unknown reward function, and $\gamma \in [0, 1)$ is the discount factor. Given a (randomized) policy $\pi : S \to \Delta(A)$, the state value function (induced by $\pi$) $V^\pi(s)$ and state-action value function $Q^\pi(s, a)$ are defined as,

$$V^\pi(s) \triangleq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t | \pi, s_0 = s]$$

$$Q^\pi(s, a) \triangleq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t | \pi, s_0 = s, a_0 = a]$$

Our goal is to minimize a notion of regret introduced in recent studies of discounted MDPs, e.g. Liu and Su [2020], He et al. [2020], Zhou et al. [2021b]. They define the regret as the cumulative suboptimality $\delta_t \triangleq V^{\pi^*}(s_t) - V^{\pi_t}(s_t)$, where $\pi^* = \arg\max_\pi V^\pi(s)$ is the optimal policy that achieves the optimal values for all $s \in S$.

**Definition 3.** We define regret for discounted MDPs as

$$\text{Reg}(T) \triangleq \sum_{t=1}^{T} \delta_t = \sum_{t=1}^{T} V^{\pi^*}(s_t) - V^{\pi_t}(s_t).$$

This notion of regret penalizes poor policies only through their effects on the states actually visited, but the use of the value function indirectly penalizes poor policies in other states as well as the decisions of the policy in those states influence the value of state $s_t$.

2.3 Slowly-changing algorithms

Our main result requires the adversarial bandits placed in each state to be slowly changing, we now give a formal definition. To measure the changing rate of an algorithm, we first define the $1$-$\infty$
norm. For a conditional matrix $M(y|x)$,

$$||M||_{1,\infty} \triangleq \max_x \sum_y |M(y|x)|,$$

e.g. the difference between two polices $\pi(a|s) - \pi'(a|s)$.

We also define an algorithm $\mathcal{A} : \{S \times A \times [0,1]\}^{t-1} \times S \rightarrow \Delta(A), t = 1, 2, \ldots$, as a function that maps history $\mathcal{F}_{t-1} = \{s_1,a_1,r_1, \ldots s_{t-1},a_{t-1},r_{t-1},s_t\}$ to a policy $\pi_t$. Now we give the definition of slowly changing algorithms.

**Definition 4** (slowly changing). An algorithm $\mathcal{A}$ is slowly changing with a (non-increasing) rate of $c_T$ if

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

where $\pi_t$ is produced by $\mathcal{A}$.

Our analysis relies on our algorithm inheriting the slowly changing property from the bandits. This definition extends to stateless bandit algorithms, by using a singleton state set $S = \{s\}$ to measure the changing rate with the 1-$\infty$ norm.

The slowly changing assumption is a mild assumption, and like our assumptions about the MDP has been used in previous work on online learning in MDPs, for example Even-Dar et al. [2009], Neu et al. [2010]. For completeness we give a proof in Section 5.2 that Exp3 Auer et al. [2002] is slowly changing in this sense, a fact also commented on and indirectly used by Neu et al. [2010].

### 3 A black-box algorithm

We now introduce our framework, in Algorithm 1. It builds on a slowly changing bandit algorithm; we refer to this bandit algorithm as $\text{Local}$. The key idea of our reduction is to run one copy of $\text{Local}$ in each state to determine the policy for that state. Furthermore, we require this bandit algorithm tolerate receiving feedback with a delay. Robustness to delays allows us to wait to provide feedback to the algorithm until a time such that the difference between the return to that time and the return of the full trajectory is small. 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 a standard Monte Carlo update is small. Note that our algorithm requires $T$ which can be handled by standard doubling trick.

For clarity, we show our naming pattern in Table 1

| Name   | Description                      |
|--------|----------------------------------|
| MAIN   | our framework without specifying LOCAL |
| LOCAL  | arbitrary local bandit, required by MAIN |
| BOLD   | a black box algo. for delayed feedback |
| BASE   | arbitrary base bandit, required by BOLD |
| MAIN×BOLD | our framework with BOLD as LOCAL |
Algorithm 1 (MAIN) Black-Box Bandit for Infinite Horizon Discounted MDPs

1: Require: $\gamma \in [0, 1), T, H = \lceil \log_{\gamma} (1 - \gamma) / \sqrt{T} \rceil$, LOCAL
2: Initialize: \{LOCAL$_s$: $s = 1, 2, \ldots, |S|$\} \Comment{Initialize one instance for each state.}
3: for $t = 1, 2, \ldots, T$ do
4: \hspace{1em} Observe state $s_t$
5: \hspace{1em} Obtain action distr. $\pi_t$ (from LOCAL$_{s_t}$)
6: \hspace{1em} Draw $a_t \sim \pi_t$
7: \hspace{1em} Observe reward $r(s_t, a_t) \in [0, 1]$
8: \hspace{1em} if $t > H$ then
9: \hspace{2em} Set $\bar{g}_{t-H} = \sum_{k=t-H}^{t} \gamma^{k-H-t}r(s_k, a_k)$ \Comment{Discounted cumulative gain from $t - H$ to current $t$, capped by horizon $H$.}
10: \hspace{2em} Return $\bar{g}_{t-H}$ to LOCAL$_{s_{t-H}}$ as feedback. \Comment{$s_{t-H}$ is the state visited at time $t - H$.}
11: \hspace{1em} end if
12: end for

3.1 Monte Carlo estimator

In Algorithm 1, $\bar{g}_t$ is a shorthand of $\bar{g}_t(s_t, a_t)$ which is the discounted return of a rollout cut off after a horizon of $H$. We now give the definition of its expectation. In the later regret analysis in Section 4 we make use of $G^\pi_t$ instead of $\bar{g}_t$ as we care about the expected regret and the unbiasedness directly comes from the definition.

$$G^\pi_t(s_t, a_t) \triangleq \mathbb{E}\left[\sum_{k=t}^{t+H} \gamma^{k-t}r_k | \mathcal{A}, \mathcal{F}_{t-1}, S_t = s_t, A_t = a\right]$$

$$= \sum_{(s, a) \in \mathcal{S} \times \mathcal{A}} \prod_{k=t}^{t+H} \mathbb{P}(s_{k+1} | s_k, a_k) \pi_k(a_k | s_k) \mathbb{E}(r_k | s_k, a_k)$$

where $\mathcal{F}_{k-1} = \{s_1, a_1, r_1, \ldots, s_{k-1}, a_{k-1}, r_{k-1}, s_k\}$ and $\pi_k = \mathcal{A}(\mathcal{F}_{k-1})$, for $k = t, t + 1, \ldots$.

Note that $G^\pi_t$ is a non-stationary analogue of action-value function $Q^\pi$.\footnote{We use $\bar{G}$ instead of $G$ to emphasize the summation goes up to time $t + H$, in contrast to the infinity in the definition of $Q$. And $\bar{G}$ is an expected value instead of random variable to align the definition of $Q$.} The difference is that $Q^\pi(s, a)$ depends only on the stationary policy $\pi$ while $G^\pi_t$ depends on $(\mathcal{A}, \mathcal{F}_{t-1})$ as the policy continues to evolve. Therefore we use $\mathcal{A}$ and $t$ for super/sub-scripts. For brevity, and to align with its stationary counterpart $Q^\pi$, we typically write $G^\pi_t$ instead. As with $Q$, $G^\pi_t$ is well defined even at states and actions other than those visited at time $t$.

In our framework, LOCAL$_s$ receive discounted cumulative rewards as feedbacks. For $s \in \mathcal{S}$, we define feedback function as: $G^\pi_{s_t} = [G^\pi_t(s, a_1), \ldots, G^\pi_t(s, a_n)]^T$. More generally, we use $y$ denote any feedback function that might appear but not specified in the following sections, in contrast to the reward function $r$ of MDPs, .

4 Regret analysis

We provide an analysis that characterizes the regret of MAIN in terms of that of LOCAL. A complication is that at each time we are only in a single state so only a single bandit is updated. 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. To conclude our analysis, we show how the implementation of LOCAL can be reduced to a standard bandit problem. Thus, the outline of our analysis is as follows.

**Stage I**: Reduction from MAIN to sticky LOCAL bandits under delayed feedback.
1. We first show that learning in an MDP can be decomposed into a set of local bandit problems, one in each state $s$, with the true state-action value function $Q^\pi$ as the feedback function;
2. We prove (by Lemma 9) that our Monte Carlo estimator has a bias that asymptotically converges to zero as $T$ increases thanks to the slowly changing assumption;
3. We then prove that MAIN is efficient if slowly changing LOCAL is no-regret under delayed feedback and sticky policies;

**Stage II**: Black-box analysis of sticky LOCALs.
4. We show that in sticky setting, LOCAL can preserve its regret bound up to a $\tilde{O}\left(\frac{\tau(S+HA)}{\sqrt{\beta(1-\gamma)}}\right)$ factor;

**Stage III**: Black-box implementation of LOCAL.
5. To turn an arbitrary bandit into one that tolerates delays we use a construction BOLD due to Joulani et al. [2013]
6. As this construction is not slowly changing we construct a time-augmented MDP that restores this property.

4.1 Reduction from Main to sticky bandits

We address stage I in this subsection.

**Step 1.**

The first key ingredient of our reduction is that MDPs can be decomposed into a set of local regret minimization problems with the help of performance difference lemma.

**Lemma 5** (Performance difference lemma, Kakade and Langford [2002], Kakade [2003]). Let $M$ be an MDP, then for all stationary policies $\pi$ and $\pi'$, and for all $s_0$ and $\gamma$,

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

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

We define local regret as follows

**Definition 6.** For $s \in S$, the local regret with feedback function $y$ is defined as:

$$R_y^s(T) \triangleq \sum_{t=1}^T \left[ \mathbb{E}_{a^* \sim \pi^*(\cdot|s)} y_t^s(a^*) - \mathbb{E}_{a_t \sim \pi_t(\cdot|s)} y_t^s(a_t) \right] = \sum_{t=1}^T \sum_{a \in A} (\pi^*(a|s) - \pi_t(a|s)) y_t^s(a)$$

where, for the regret notion $R_y^s$, the subscript $s$ emphasizes the specific state $s$ where the regret is measured, and the superscript $y$ emphasizes the feedback function used to evaluate the regret.
Remark. Although feedback delays do not change the definition of regret, to emphasize the problem setting, we will use $R_y(T|z)$ for delayed setting with constant delay $z$.

The following lemma states that the expected regret for an MDP can be bounded by the cumulative expected regret in the local bandit problems.

**Lemma 7.** The global $\mathcal{R}_{\text{Reg}}(T)$ can be bounded by the sum of local regret (up to a $\frac{1}{1-\gamma}$ factor), while local regrets are measured w.r.t. the true state-value function $Q^\pi$. 

$$\mathcal{R}_{\text{Reg}}(T) \leq \frac{1}{1-\gamma} \sum_{s \in S} R_{Q^\pi_s}(T)$$

where feedback $Q^\pi_s = [Q^\pi(s,a_1), \ldots, Q^\pi(s,a_|A|)]^\top$.

Lemma 7 shows that, assuming the existence of an oracle that gives the true value function of the current policy, one could apply the true value function as the feedback function of local bandit problems and the total local regrets will form an upper bound on the expected regret in the original MDP. The proof of Lemma 7 is found in appendix B.1.

**Step 2.**

While the decomposition aligns our goal, to place one bandit player in one state, yet we have no control to the correctness of our Monte Carlo estimator. As

- The ground-truth feedback function $y_s^t = Q^\pi_t(s, \cdot)$, is determined by the set of local players only at time $t$ and the underlying MDP;
- Meanwhile $\bar{G}_s^t$ takes account of the future dynamics of all bandit learners.

Given our definition of slowly changing, it can be show that the slowly changing property of LOCAL is preserved by MAIN.

**Lemma 8.** MAIN is slowly changing if LOCAL is slowly changing.

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

$$||\pi_{t+1} - \pi_t||_{1,\infty} = ||\phi^1_{t+1} - \phi_t||_{1,\infty} \leq c$$

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

Given MAIN is slowly changing, we are able to prove Lemma 9, which is a key technical tool to show that the bias of our estimator relative to the true value function $Q^\pi_t$ can be controlled. The proof is found in appendix B.1.

**Lemma 9.** If MAIN is slowly changing with a rate of $c_T$, then

$$|\bar{G}_s^t(s,a) - Q^\pi_t(s,a)| \leq mc_T + \frac{1}{\sqrt{T}},$$

where $m = \frac{\gamma}{(1-\gamma)^2}S + \frac{(1+\gamma)^2}{(1-\gamma)^3}A$, $S = |S|$, and $A = |A|$.

**Step 3.**

Now we are ready to give the first main theorem, which bounds the expect regret of MAIN by the total regret of local delayed problems.
Theorem 10. Let \( m \) be the same as in Lemma 9 and \( c_T = \mathcal{O}(1/\sqrt{T}) \), let \( \mathcal{R}_{a}^\pi(T) \) be the expect local regret, under delayed feedback, in state \( s \in S \). For arbitrary choice of slowly changing LOCAL, MAIN has the following expected regret bound

\[
\mathcal{R}_{a}(T) \leq \tilde{O}\left(\frac{(1 + \gamma)\gamma^2}{(1 - \gamma)^4(S + A)\sqrt{T}}\right) + \sum_{s \in S} \mathcal{R}_{a}^\pi(T|H) \]

where \( \tilde{O} \) is the local feedback function (subject to delays).

Proof. Assume \( \{\pi_t : 1 \leq t \leq T\} \) is the sequence of polices obtained by running MAIN.

\[
\mathcal{R}_{a}(T) = \sum_{t=1}^{T} V^*(s_t) - V^\pi_t(s_t)
\]

\[
= \frac{1}{1 - \gamma} \sum_{t=1}^{T} \mathbb{E}_{\pi_t \sim d_{\pi_t}^s} \left[ Q^\pi_t(s, a) - V^\pi_t(s) \right]
\]

\[
= \frac{1}{1 - \gamma} \sum_{t=1}^{T} \sum_{s \sim d_{\pi_t}^s} \sum_{a \in A} \left[ \pi^*(a|s) - \pi_t(a|s) \right] Q^\pi_t(s, a)
\]

\[
\leq \frac{1}{1 - \gamma} \sum_{t=1}^{T} \sum_{s \sim d_{\pi_t}^s} \sum_{a \in A} \left[ \pi^*(a|s) \left( \tilde{G}^\pi_t(s, a) + \left| Q^\pi_t(s, a) - \tilde{G}^\pi_t(s, a) \right| \right)
\]

\[
- \pi_t(a|s) \left( \tilde{G}^\pi_t(s, a) - \left| Q^\pi_t(s, a) - \tilde{G}^\pi_t(s, a) \right| \right) \]

\[
\leq \frac{1}{1 - \gamma} \sum_{t=1}^{T} \sum_{s \sim d_{\pi_t}^s} \sum_{a \in A} \Delta^\pi(a|s) \tilde{G}^\pi_t(s, a) + 2 \left| Q^\pi_t(s, a) - \tilde{G}^\pi_t(s, a) \right|
\]

\[
\leq \frac{1}{1 - \gamma} \sum_{t=1}^{T} \left[ 2mc_T + \frac{2}{\sqrt{T}} + \mathbb{E}_{d_{\pi_t}^s} \sum_{a \in A} \Delta^\pi(a|s) \tilde{G}^\pi_t(s, a) \right]
\]

\[
\leq \frac{1}{1 - \gamma} \sum_{t=1}^{T} \left[ 2mc_T + \frac{2}{\sqrt{T}} + \sum_{s \in S} \sum_{a \in A} \Delta^\pi(a|s) \tilde{G}^\pi_t(s, a) \right]
\]

\[
\leq \frac{1}{1 - \gamma} \mathcal{O}(m\sqrt{T}) + \sum_{s \in S} \sum_{t=1}^{T} \mathcal{R}_{a}^\pi(T|H)
\]

Plugging in \( m \) concludes the proof.

where \( a \) follows from performance difference lemma by applying \( \pi' = \pi^* \) and \( \pi = \pi_t \); \( b \) comes from triangle inequality; in \( c \), we denote \( \Delta^\pi(a|s) = \pi^*(a|s) - \pi_t(a|s) \) for simplicity, and the inequality follows from \( \sum_a (\pi^*(a|s) + \pi_t(a|s)) = 2 \); \( d \) is from Lemma 9.
4.2 Sticky bandits

Step 4.

Note that Theorem 10 decomposes the regret into local regrets, however, we are not able to update each local bandit every timestep. As the sequence of $s_t$ is controlled by MAIN, each LOCAL can only act an unknown subset of the time.

However, thanks to the slowly changing property of MAIN, we are able to show that our actual feedback $\bar{G}^{\pi_t} (s,a)$ is also slowly changing.

**Lemma 11.** If MAIN is slowly changing with a non-increasing rate of $c_T$, we have

$$|\bar{G}^{\pi_{t+1}}(s,a) - \bar{G}^{\pi_t}(s,a)| \leq f(n)c_T$$

where $f(n) = \frac{1}{1-\gamma}(S + HA)n$

Therefore, we are able to approximately measure the potential regret (w.r.t. its current policy), even when a local bandit algorithm is not able to act, with the most recent regret.

**Definition 12.** Assuming $d_1$ is an arbitrary initial distribution, and $\pi_1, \pi_2, \ldots, \pi_{t-1}$ are arbitrary sequence of policies that are slowly changing. We define the $q_t(s)$ as the probability distribution over $S$.

$$q_t(s) = \Pr(S_t = s) = \{d_1P^{\pi_1}P^{\pi_2}\ldots P^{\pi_{t-1}}\}(s)$$

**Lemma 13.** Assume $T$ is sufficiently large such that $c_T\tau^2 < \beta/4$, then we have

$$q_t(s) \geq \beta/2, \text{ for } t > \tau \ln(8/\beta)$$

Then, let $v_t$ be the state distribution produced by running MAIN. Trivially by Lemma 13, we have

**Corollary 14.** Assume $T$ is sufficiently large such that $c_T\tau^2 < \beta/4$, for all $s \in S$. While running MAIN, starting from any state $s_t \in S$ at time $t$, we have

$$v_t(s) \geq \beta/2, \text{ for } t' > t + \tau \ln(8/\beta)$$

As $v_t$ is now uniformly bounded away from zero regardless of the current state if $T$ is sufficiently large, given Corollary 32. Therefore, the local sticky bandits are updated sufficiently often, in expectation, to achieve no-regret. Lemma 15 gives a regret bound for sticky bandits when the chosen bandit algorithm enjoys $O\left(\sqrt{(A+H)T\ln A}\right)$ regret in the standard bandit problem. Proof is found in Appendix B.2.

**Lemma 15.** If Assumption 1 and 2 are satisfied. For a slowly changing LOCAL bandit, with a rate of $O(1/\sqrt{T})$, enjoys $O(\sqrt{(A+H)T\ln A})$ expect regret in standard bandit problem under constant feedback delay $H$. It then enjoys

$$\bar{O}\left(\frac{\tau(S + HA)}{\sqrt{\beta(1 - \gamma)}} \sqrt{T}\right)$$

regret in the case of sticky setting.

Combining Theorem 10 and Lemma 15 leads to

**Corollary 16.** Let LOCAL be slowly changing with rate $c_T = O(1/\sqrt{T})$ and enjoys $O(\sqrt{(A+H)T\ln A})$ expected regret under constant delay $H$. If Assumption 1 and 2 are satisfied, and $T$ is sufficiently large such that $c_T\tau^2 < \beta/2$, then MAIN enjoys expected regret

$$\mathbb{R} \text{eg}(T) = \bar{O}\left(\frac{\tau (1 + \gamma)\gamma}{\sqrt{\beta} (1 - \gamma)^4} (S^2 + SA)\sqrt{T}\right)$$

(3)
4.3 Delayed feedback

We then address steps 5 and 6 in this part. Due to our construction, we first introduce feedback delays into our Algorithm 1. There has been extensive study of bandits with delayed feedback, see e.g. Joulani et al. [2013] for a summary. We leverage the result from Joulani et al. [2013], which bounds the regret of delayed problem for arbitrary bandit algorithm with its non-delayed guarantees. The authors propose a black-box bandit algorithm for (arbitrary) delay. In our case, we manually introduce a constant delay, therefore we present their algorithm specifically to handle constant delay in Algorithm 2.

**Algorithm 2** (BOLD\textsubscript{s}) Black-box Online Learning under (constant) Delayed feedback (for state \textit{s})

1. **Require:** constant delay \(H\)
2. **Initialize:** \(\text{BASE}\textsubscript{\textit{s}}^h : h = 1, 2, \ldots, H + 1\)
3. **for** \(t = 1, 2, \ldots, T\) **do**
   \(\triangleright t\) is global time of Main.
4. Set \(h_t = \lfloor t \mod (H + 1) \rfloor + 1\)
5. **if** \(s_t = s\) **then**
   6. Choose \(\text{BASE}\textsubscript{\textit{s}}^{h_t}\) to make prediction
    7. **end if**
8. **if** \(t > H\) and \(s_{t-H} = s\) **then**
   9. Receive feedback \(y_{t-H}\) for time \(t - H\)
10. Update \(\text{BASE}\textsubscript{\textit{s}}^{h_t-H}\) with \(y_{t-H}\)
11. **end if**
12. **end for**

The essence of the construction is using \(H + 1\) \textit{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 17.** (Joulani et al. [2013]) Suppose that the \textit{BASE} used in BOLD enjoys an (expected) regret bound \(f_{\text{BASE}}(T)\) in non-delayed setting. Assume, furthermore, that the delays are independent of the forecaster’s predictions. Then the expected regret of BOLD after \(N\) time steps satisfies

\[
\mathbb{E}[R] \leq (\mathbb{E}(G^*_T) + 1)f_{\text{BASE}}\left(\frac{T}{\mathbb{E}(G^*_T) + 1}\right)
\]

where \(G^*_t\) is maximal number of missing feedback during the first \(t\) time-steps. For constant delay \(H, \mathbb{E}(G^*_t) = H\) when \(t \geq H\).

We give an illustration of \textit{Main} \times BOLD in Figure 1.

**Step 5.**

Theorem 10 requires \textit{Main} to be slowly changing. Unfortunately, \textit{Main} \times BOLD is not slowly changing due to the switching mechanism used in its construction: each of the \(H + 1\) \textit{BASE} bandits could have arbitrarily different policies.

However, to preserve the slowly changing property we can augment the state space \(S\) by concatenating a time stamp \(h \in \mathcal{H} \triangleq \{1, 2, \ldots, H + 1\}\), i.e. \(\tilde{S} \triangleq S \times \mathcal{H}\).

Intuitively speaking, the switching mechanism by construction is now part of the state transition function and it is then possible to preserve the slowly changing property. Definition 18 gives a formal statement of time-argumented MDPs. (Note that if one specifies a (delay-robust) slowly changing bandit for \text{LOCAL} rather than BOLD, such expanded MDP construction is not necessary, given Lemma 8.)
**Definition 18.** Given a MDP $M = (S, A, P, r, \gamma)$, and let $H \triangleq \{1, 2, \ldots, H+1\}$, where $H$ is the constant delay constructed by MAIN. We define the $H$-argumented MDP as $\tilde{M} = (\tilde{S}, A, \tilde{P}, \tilde{r}, \gamma)$,

- $\tilde{S} \triangleq S \times H$
- $\tilde{P}(s \circ h, a, s' \circ h') \triangleq P(s, a, s') \mathbb{I}_{\{h' = [h+1 \mod (H+1)]\}}$
- $\tilde{r}(s \circ h, a) \triangleq r(s, a)$

where $s \in S, h \in H, a \in A, s \circ h \in \tilde{S}$, $\mathbb{I}_{\{\cdot\}}$ is indicator function and $\circ$ denotes concatenation.

By definition, we have

$$V^\pi(s \circ h) = V^\pi(s) \quad Q^\pi(s \circ h, a) = Q^\pi(s, a) \quad \tilde{\pi}^\ast(a|s \circ h) = \pi^\ast(a|s)$$

$$\text{Reg}(T) = \sum_{t=1}^{T} V^\pi(s_t) - V^\pi(s_t) = \sum_{t=1}^{T} V^{\tilde{\pi}^\ast}(s_t \circ h_t) - V^{\tilde{\pi}^\ast}(s_t \circ h_t)$$

where $\pi^\ast$ and $\tilde{\pi}^\ast$ are the optimal policies in the original MDP $M$ and its time-augmented surrogate $\tilde{M}$, respectively.

**Lemma 19.** MAIN $\times$ BOLD is slowly changing in $\tilde{M}$ if BASE is slowly changing.

Proof is found in appendix B.

**Step 6.**

Now we are ready to deliver the full reduction theorem, by running MAIN $\times$ BOLD in the surrogate MDPs.

**Theorem 20.** (Reduction) Let MDP $M$ that satisfy Assumption 1 and 2. Assume a slowly changing BASE, with a rate $c_T = O(1/\sqrt{T})$ enjoys $O(\sqrt{AT\ln A})$ regret in the standard (non-delayed) bandit.
setting, and $T$ is sufficiently large such that $c_T T^2 < \beta/4$. Running MAIN × BOLD in $\tilde{M}$ achieves the expect regret bound of

$$\text{Reg}(T) = \tilde{O}\left(\frac{\tau}{\sqrt{\beta}} \left(1 + \frac{1}{2}\right) \frac{1}{1-\gamma} (S^2 + SA) \sqrt{T}\right)$$

**Proof.** Applying Lemma 9 in $\tilde{M}$ yields

$$|\tilde{G}^{\pi_t}(\tilde{s}, a) - Q^{\pi_t}(\tilde{s}, a)| \leq mc_T + \frac{1}{\sqrt{T}}$$

where $m = \frac{\gamma}{(1-\gamma)^2} HS + \frac{1}{1-\gamma} A$.

By Theorem 10, and set delay $z = 0$ due to our BOLD construction.

$$\text{Reg}(T) \leq \frac{1}{1-\gamma} \left[ O(m \sqrt{T}) + \sum_{s \in \tilde{S}} \tilde{G}^{\pi_t}(T|0) \right]$$

By Lemma 15 and $|\tilde{S}| = HS$

$$\leq \frac{1}{1-\gamma} \left[ O(m \sqrt{T}) + \sum_{s \in \tilde{S}} \tilde{O}\left(\frac{\tau H(S + A)}{\sqrt{\beta} (1-\gamma)} \sqrt{T}\right) \right]$$

Plug in $m$ finishes the proof.

$$= \tilde{O}\left(\frac{\tau}{\sqrt{\beta}} \left(1 + \frac{1}{2}\right) \frac{1}{1-\gamma} (S^2 + SA) \sqrt{T}\right)$$

### 5 Discussion

We conclude by discussing possible extension of our results to episodic setting. We also examine EXP3 to give an example of the applicability of our reduction to a standard bandit algorithm.

#### 5.1 Episodic setting

Another interesting direction would be to adapt our approach to the episodic setting. One could easily convert an episodic setting with $N$ episodes and $L$ steps per episode to a non-episodic setting with $T = N \times L$ timesteps. However, the typical regret notion used for the episodic setting is slightly different Zhou et al. [2021a].

$$\text{Reg}(N) = \sum_{k=1}^{N} V^{\pi^*}(s_0^k) - V^{\pi_t}(s_0^k)$$

This can be bounded by our version of regret, which a non-episodic equivalent with a state reset every $L$ steps.

$$\leq \sum_{k=1}^{N} \sum_{t=1}^{L} V^{\pi^*}(s_t^k) - V^{\pi_t}(s_t^k).$$
Here, $V_\pi(s) = \mathbb{E}[\sum_{t=0}^L \gamma^t r_t | \pi, s_0 = s]$ as an episodic analogy to our previous definition, and $\gamma \in [0, 1)$. The state reset is essentially a change of occupancy measure, which is removed from Eq. (1) to Eq. (2) and is replaced by a sum over $\mathcal{S}$.

Intuitively, our approach could therefore be extended to this setting, but care would be needed due to the interaction of this resetting with our analysis of sticky bandits.

5.2 Case study: Exp3

We first introduce Exp3 to provide the relevant notation.

**Algorithm 3 Exp3**

1. **Require:** $\gamma \in [0, 1), \eta_T \in (0, 1], A = |\mathcal{A}|$
2. **Initialize:** $w_1(a) = 1$ for $a \in \mathcal{A}$
3. **for** $t = 1, 2, \ldots, T$ **do**
4. 
   \begin{align*}
   W_t &= \sum_{a=1}^{A} w_t(a) \\
   p_t(a) &= (1 - \eta_T) \frac{w_t(a)}{W_t} + \frac{\eta_T}{A}
   \end{align*}
5. 
   **Draw** $a_t$ **randomly accordingly to** $p_t$
6. 
   **Receive reward** $y_t(a) \in [0, \frac{1}{1-\gamma}]$
7. 
   **For** $a = 1, 2, \ldots, A$, **set**
   \begin{align*}
   \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{align*}
8. **end for**

5.2.1 Exp3 with delayed feedback

It is known that the optimal regret is $O(\sqrt{(A + z)T \ln A})$ for constant delay $z$ Cesa-Bianchi et al. [2019], 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 $O(\sqrt{A T + Z} \ln A)$ under delayed feedback, where $Z$ is the total delay. Therefore, using Exp3 as Base, LOCAL enjoys $O(\sqrt{(A + H)T \ln A})$ regret in Lemma 15 and Corollary 16.

5.2.2 Exp3 is slowly changing

Here we show that Exp3 meets the key slowly changing property with a rate of $\eta_T/A$

**Lemma 21.** Exp3 is slowly changing with a rate of $O(\eta_T)$ where the feedback $y_t$ is bounded between $[0, \frac{1}{1-\gamma}]$.

**Proof.** We observe it is sufficient to bound $p_{t+1}(a) - p_t(a)$ of the action chosen by the algorithm at time-step $t$. 

13
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.

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

$$= (1 - \eta T) \left( \frac{w_t e^{(1 - \gamma)\eta T \hat{y}_t / A}}{w_t + w_t e^{(1 - \gamma)\eta T \hat{y}_t / A} - w_t} \right)$$

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

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

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

$$\leq (1 - \eta T) \left( 2 \left( \frac{\eta T}{Ap_t} \right) \left( \frac{w_t}{W_t} \right) \right)$$

$$\leq 2\eta T / A. \quad (6)$$

Equation 5 follows from that $e^x - 1 < 2x$ for $0 \leq x \leq 1$. Equation 6 follows from $p_t \geq (1 - \eta T) \left( \frac{w_t}{W_t} \right)$.

We note that to achieve $\tilde{O}(\sqrt{AT})$ regret Exp3 is run with a learning rate of $\eta T = \tilde{O}(\sqrt{A/T})$, which means it is slowly changing with a rate of $c_T = \tilde{O}(\sqrt{1/AT})$.

**Acknowledgement**

This work was supported by the National Science Foundation grant CCF-1934915.
References

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

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In International Conference on Machine Learning, pages 263–272. PMLR, 2017.

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.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Delay and cooperation in nonstochastic bandits. 2019.

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, 2020.

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

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted mdps. arXiv preprint arXiv:2010.00587, 2020.

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.

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 Proc. 19th International Conference on Machine Learning. Citeseer, 2002.

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

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.

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.
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.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. 2019.

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

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via 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.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In Conference on Learning Theory, pages 4532–4576. PMLR, 2021a.

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, 2021b.
Appendix

A Technical tools

One can easily extend the slowly changing property 4 to a multi-step version,

**Lemma 22.** If an algorithm \( A \) is slowly changing with a non-increasing rate of \( c_T \), then

\[
||\pi_{t+k} - \pi_t||_{1,\infty} \leq kc_T \tag{7}
\]

**Proof.** Trivially by triangle inequality.

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

\[
\leq \sum_{i=0}^{k-1} c \leq kc_T \tag{9}
\]

\[\blacksquare\]

**Lemma 23.** Suppose \( ||\pi - \pi'||_{1,\infty} \leq c' \). Then, for any state distribution vector \( d \), we have

\[
||d\mathbb{P}^\pi - d\mathbb{P}^{\pi'}||_1 \leq c' \tag{10}
\]

where \( \mathbb{P}^\pi \) is the transition matrix induced from \( \pi \).

**Proof.**

\[
||d\mathbb{P}^\pi - d\mathbb{P}^{\pi'}||_1 = \sum_{s'} |d\mathbb{P}^\pi(s') - d\mathbb{P}^{\pi'}(s')| \tag{11}
\]

\[
= \sum_{s'} \left| \sum_s [d(s)\mathbb{P}^\pi(s,s') - d(s)\mathbb{P}^{\pi'}(s,s')] \right| \tag{12}
\]

\[
\leq \sum_{s'} \sum_s d(s)||\mathbb{P}^\pi(s,s') - \mathbb{P}^{\pi'}(s,s')|| \tag{13}
\]

\[
= \sum_{s'} \sum_s d(s)\sum_a |\mathbb{P}(s,a,s')\pi(a|s) - \mathbb{P}(s,a,s')\pi'(a|s)| \tag{14}
\]

\[
\leq \sum_{s'} \sum_s \sum_a d(s)|\mathbb{P}(s,a,s')\pi(a|s) - \pi'(a|s)| \tag{15}
\]

\[
= \sum_{s} d(s)\sum_a |\pi(a|s) - \pi'(a|s)| \tag{16}
\]

\[
\leq \sum_{s} d(s)||\pi - \pi'||_{1,\infty} = c' \tag{17}
\]

\[\blacksquare\]

**Lemma 24.** For any state distribution vectors \( d \) and \( d' \), we have

\[
||d\mathbb{P}^\pi - d'\mathbb{P}^\pi||_1 \leq ||d - d'||_1 \tag{18}
\]

where \( \mathbb{P}^\pi \) is the transition matrix induced from \( \pi \).
Proof.

\[ ||d^\pi - d'^{\pi'}||_1 = \sum_{s'} |d^\pi(s') - d'^{\pi'}(s')| \]  \hfill (19)

\[ = \sum_{s'} |\sum_s [d^\pi(s, s') - d'(s)\pi(s, s')]| \]  \hfill (20)

\[ \leq \sum_{s'} \sum_s |\pi(s, s')|d(s) - d'(s)| \]  \hfill (21)

\[ = \sum_s |d(s) - d'(s)| \sum_{s'} \pi(s, s') \]  \hfill (22)

\[ \leq \sum_{s'} \sum_s \pi(s, s') ||d - d'||_1 \]  \hfill (23)

Lemma 25. 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

\[ ||d^\pi - d'^{\pi'}||_1 \leq c' + \delta \]  \hfill (24)

Proof.

\[ ||d^\pi - d'^{\pi'}||_1 = ||d^\pi - d^{\pi'} + d^{\pi'} - d'^{\pi'}||_1 \]  \hfill (25)

\[ \leq ||d^\pi - d^{\pi'}||_1 + ||d^{\pi'} - d'^{\pi'}||_1 \]  \hfill (26)

by Lemma 24

\[ \leq ||d^\pi - d^{\pi'}||_1 + ||d - d'||_1 \]  \hfill (27)

by Lemma 23

\[ \leq c' + \delta \]  \hfill (28)

Lemma 26. 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

\[ ||d(\pi^1, \ldots, \pi^K) - d'(\pi'^1, \ldots, \pi'^K)||_1 \leq Kc' + \delta \]  \hfill (29)

Proof. We prove this by induction on \( K \),

\( K = 1 \): by Lemma 25

\[ ||d^{\pi^1} - d'^{\pi'^1}||_1 \leq c' + \delta \]  \hfill (30)

\( K = n \): assume we have,

\[ ||d(\pi^1, \ldots, \pi^n) - d'(\pi'^1, \ldots, \pi'^n)||_1 \leq nc' + \delta \]  \hfill (31)

\( K = n + 1 \):

\[ ||d(\pi^1, \ldots, \pi^{n+1}) - d'(\pi'^1, \ldots, \pi'^{n+1})||_1 \]  \hfill (32)

\[ = ||d_n^{\pi^{n+1}} - d'_n^{\pi'^{n+1}}|| \]  \hfill (33)

\[ \leq nc' + \delta + c' \]  \hfill (34)
**Corollary 27.** Let $\pi_t$ be slowly changing with non-increasing rate $c_T$, then we have $\|\pi_{t+1} - \pi_t\|_{1,\infty} \leq l c_T$. Apply Lemma 26 with $\pi_k = \pi_{t+k-1}$, $\pi'_k = \pi_t$ and $d = d''$, then

$$\|d (P^{\pi_t} \ldots P^{\pi_{t+K-1}}) - d(P^{\pi_t})^K\|_1 \leq K^2 c_T$$

**Corollary 28.** Let $\pi_t$ be slowly changing with non-increasing rate $c_T$, then we have $\|\pi_{t+n} - \pi_t\|_{1,\infty} \leq n c_T$. Apply Lemma 26 with $\pi_k = \pi_{t+k-1}$, $\pi'_k = \pi_{t+k-1}$ and $d = d''$, then

$$\|d (P^{\pi_{t+n}} \ldots P^{\pi_{t+n+K-1}}) - d(P^{\pi_{t+n}})^K\|_1 \leq K n c_T$$

### B Omitted proofs

#### B.1 Stage I: Main to Local

**Lemma 7.** The expected regret in MDPs can be reduced to cumulative expected regrets of local bandit problems,

$$\mathcal{R}_{\text{Reg}}(T) \leq \frac{1}{1-\gamma} \sum_{s \in S} Q^\pi(s) (T|H)$$

where feedback $Q^\pi = [Q^\pi(s, a_1), \ldots, Q^\pi(s, a_n)]^T$

**Proof.** Assume $\{\pi_t : 1 \leq t \leq T\}$ is the sequence of policies obtained by running by any algorithm $\mathcal{A}$.

$$\mathcal{R}_{\text{Reg}}(T) = \sum_{t=1}^{T} V^*(s_t) - V^\pi(s_t)$$

apply Lemma 5 with $\pi' = \pi^*$ and $\pi = \pi_t$

$$= \frac{1}{1-\gamma} \sum_{t=1}^{T} E_{s \sim d^\pi_t} \mathbb{E}_{a \sim \pi^*} [Q^\pi(s, a) - V^\pi(s)]$$

$$= \frac{1}{1-\gamma} \sum_{t=1}^{T} E_{s \sim d^\pi_t} \sum_{a \in A} (\pi^*(a|s) - \pi_t(a|s)) Q^\pi(s, a)$$

$$\leq \frac{1}{1-\gamma} \sum_{t=1}^{T} \sum_{s \in S} \sum_{a \in A} (\pi^*(a|s) - \pi_t(a|s)) Q^\pi(s, a)$$

by definition 6

$$= \frac{1}{1-\gamma} \sum_{s \in S} Q^\pi(s) (T)$$

**Lemma 29.** Let $G^\pi(s, a) = \mathbb{E}[\sum_{k=t}^{\infty} \gamma^{k-t} r_k | \mathcal{A}, F_{t-1}, S_t = s_t, A_t = a]$ be an infinite horizon expectation, in contrast to $G^\pi(s, a) = \mathbb{E}[\sum_{k=t}^{\infty} \gamma^{k-t} r_k | \mathcal{A}, F_{t-1}, S_t = s_t, A_t = a]$, which capped by horizon $H$.

$$|G^\pi(s, a) - G^\pi(s, a)| \leq \frac{1}{\sqrt{T}}$$
Proof:

\[|G^\pi_t(s, a) - G^\pi(s, a)| = \left| \sum_{k=t}^{t+H} \gamma^k r(s_k, a_k) \right| = \sum_{k=t}^{t+H} \gamma^k r(s_k, a_k) \]  

\[= \left| \sum_{k=t+1}^{\infty} \gamma^k r(s_k, a_k) \right| \leq \sum_{k=t+1}^{\infty} \gamma^k \leq \frac{\gamma^{H+1}}{1-\gamma} \]  

\[\leq \frac{\gamma^{\log_2 \frac{1}{1-\gamma}}}{1-\gamma} = \frac{1}{\sqrt{T}} \]  

Lemma 9. If MAIN is slowly changing with a non-increasing rate of \( c_T \), then

\[|G^\pi_t(s, a) - Q^\pi(s, a)| \leq m c_T + \frac{1}{\sqrt{T}} \]  

where \( m = \frac{\gamma}{1-\gamma} S + \frac{\gamma(1+\gamma)}{(1-\gamma)^2} A \)

Proof.

We will use the following fact for our proof. Let \( a_1, a_2, b_1, b_2 \in [0,1] \),

\[|a_1 b_1 - a_2 b_2| = |a_1 b_1 - a_1 b_2 + a_1 b_2 - a_2 b_2| \leq |a_1 b_1 - a_1 b_2| + |a_1 b_2 - a_2 b_2| \leq |b_1 - b_2| + |a_1 - a_2| \]

Let \( k = t, t+1, \ldots, G^\pi_t(s, a) \) defined as in Lemma 29. Noticing that

\( \bullet \) \( G^\pi_t(s, a) \) and \( Q^\pi(s, a) \) start from same policy \( \pi_t \), and same state distribution, i.e. \( d_t = d'_t = d_s \), where \( d_s \) means \( \Pr(S_t = s) = 1 \).

\( \bullet \) The future state distributions \( d'_k \) of \( G^\pi_t \) follows \( d'_{k+1} = d'_k \bar{Q}^\pi_t \), where \( \pi_k \) is slowly changing;

\( \bullet \) The future state distributions \( d_k \) of \( Q^\pi_t \) follows \( d_{k+1} = d_k \bar{Q}^\pi_t \) by definition.

By Lemma 22 and corollary 27, we have (we split cases to avoid invalid notion \( d_t(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_t-1}) \) when \( k = t \))

\[||\pi_k - \pi_t||_{1,\infty} \leq (k-t) c_T \]

\[
\left\{
\begin{array}{ll}
  k = t & ||d_t - d_t||_1 \leq (k-t)^2 c_T = 0 \\
  k > t & ||d_t(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_t-1}) - d_t(\mathbb{P}^{\pi_t})^{k-t}||_1 \leq (k-t)^2 c_T \\
\end{array}
\right.
\]

\[\left|G^\pi_t(s, a) - Q^\pi_t(s, a)\right| \leq \sum_{k=t}^{\infty} \sum_{x \in S} \sum_{i \in A} \gamma^{k-t} \left|d'_k(x)\pi_k(i|x) - d_k(x)\pi_t(i|x)\right| r(x, i) \]
given \(d'_k = d'_t(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_{k-1}})\) and \(d_k = d_t(\mathbb{P}^{\pi_t})^{k-t}\)

\[
\leq \sum_{k=t}^{\infty} \sum_{x,i} \gamma^{k-t} \left| d'_t(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_{k-1}})(x)\pi_k(i|x) - d_t(\mathbb{P}^{\pi_t})^{k-t}(x)\pi_t(i|x) \right|
\]

(56)

\[
\leq \sum_{k=t}^{\infty} \sum_{x,i} \gamma^{k-t} \left| d'_t(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_{k-1}})(x)\pi_k(i|x) - d_t(\mathbb{P}^{\pi_t})^{k-t}(x)\pi_k(i|x) \right|
\]

(57)

given \(|a_1b_1 - a_2b_2| \leq |a_1 - a_2| + |b_1 - b_2|\) and \(d'_t = d_t\) (starting from the same initial distributions), we have

\[
\leq \sum_{k=t}^{\infty} \sum_{x,i} \gamma^{k-t} \left\{ \left| \pi_k(i|x) - \pi_t(i|x) \right| + \left| d_t(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_{k-1}})(x) - d_t(\mathbb{P}^{\pi_t})^{k-t}(x) \right| \right\}
\]

(58)

\[
\leq \sum_{k=t}^{\infty} \gamma^{k-t} \left\{ S \max_x \sum_i |\pi_k(i|x) - \pi_t(i|x)| + \sum_x \sum_i |d_t(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_{k-1}})(x) - d_t(\mathbb{P}^{\pi_t})^{k-t}(x)| \right\}
\]

(59)

\[
\leq \sum_{k=t}^{\infty} \gamma^{k-t} \left\{ S \|\pi_k - \pi_t\|_1 + A \left| d_t(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_{k-1}}) - d_t(\mathbb{P}^{\pi_t})^{k-t}\right| \right\}
\]

(60)

given Eq.(52) and Eq. (53)

\[
\leq \sum_{k=t}^{\infty} \gamma^{k-t} \left\{ S(k-t)c_T + A(k-t)^2c_T \right\}
\]

(61)

by the limiting sum of arithmetico–geometric sequence

\[
\leq \left\{ \frac{\gamma}{(1-\gamma)^2} S + \frac{\gamma(1+\gamma)}{(1-\gamma)^3} A \right\} c_T
\]

(62)

Combining with Lemma 29 finishes the proof.

\[\blacklozenge\]

\subsection*{B.2 Stage II: Sticky bandits}

\textbf{Lemma 11.} If MAIN is slowly changing with a non-increasing rate of \(c_T\), then

\[
|\bar{G}^{\pi_{t+n}}(s,a) - \bar{G}^{\pi_t}(s,a)| \leq f(n)c_T
\]

(63)

where \(f(n) = \frac{1}{1-\gamma}(S + HA)n\)

\textbf{Proof:}

Similarly, \(\bar{G}^{\pi_{t+n}}(s,a)\) and \(\bar{G}^{\pi_t}(s,a)\) starts from the same state distribution \(d_t = d'_t = d_s\) but follows different sequences of policies.

By Lemma 22 and corollary 28, we have

\[
\|\pi_{t+n} - \pi_t\|_{1,\infty} \leq n c_T
\]

(64)

\[
\|d_t(\mathbb{P}^{\pi_{t+n}} \cdots \mathbb{P}^{\pi_{k+n}}) - d(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_k})\|_1 \leq n(k-t)c_T
\]

(65)

Let \(\pi_t\) be the sequence of policies produced by MAIN, \(d'_k = d_s(\mathbb{P}^{\pi_{t+n}} \cdots \mathbb{P}^{\pi_{k+n}})\), and \(d_k = d_s(\mathbb{P}^{\pi_t} \cdots \mathbb{P}^{\pi_k})\), where \(k = t, t+1, \ldots, t + H\)

\[
|\bar{G}^{\pi_{t+n}}(s,a) - \bar{G}^{\pi_t}(s,a) |
\]

(66)
\[
\begin{align*}
\sum_{k=t}^{t+H} \gamma^{k-t} \left[ d_k'(x) \pi_{k+n}(i|x) - d_k(x) \pi_k(i|x) \right] r(x,i)
\end{align*}
\]  
(67)

\[
\leq \sum_{k=t}^{t+H} \gamma^{k-t} \left[ d_s(\mathbb{P}^{\pi_{t+n}} \ldots \mathbb{P}^{\pi_k+n})(x) \pi_{k+n}(i|x) - d_s(\mathbb{P}^{\pi_t} \ldots \mathbb{P}^{\pi_k})(x) \pi_k(i|x) \right] r(x,i)
\]
(68)

\[
\leq \sum_{k=t}^{t+H} \gamma^{k-t} \left[ d_s(\mathbb{P}^{\pi_{t+n}} \ldots \mathbb{P}^{\pi_k+n})(x) \pi_{k+n}(i|x) - d_s(\mathbb{P}^{\pi_t} \ldots \mathbb{P}^{\pi_k})(x) \pi_k(i|x) \right]
\]
(69)

given \( |a_1 b_1 - a_2 b_2| \leq |a_1 - a_2| + |b_1 - b_2| \) for \( a_1, a_2, b_1, b_2 \in [0, 1] \), we have

\[
\leq \sum_{k=t}^{t+H} \gamma^{k-t} \left\{ |\pi_{k+n}(i|x) - \pi_k(i|x)| + |d_s(\mathbb{P}^{\pi_{t+n}} \ldots \mathbb{P}^{\pi_k+n})(x) - d_s(\mathbb{P}^{\pi_t} \ldots \mathbb{P}^{\pi_k})(x)| \right\}
\]
(70)

\[
\leq \sum_{k=t}^{t+H} \gamma^{k-t} \left\{ S \max_i |\pi_{k+n}(i|x) - \pi_k(i|x)| + \sum_i \sum_k |d_s(\mathbb{P}^{\pi_{t+n}} \ldots \mathbb{P}^{\pi_k+n})(x) - d_s(\mathbb{P}^{\pi_t} \ldots \mathbb{P}^{\pi_k})(x)| \right\}
\]
(71)

\[
\leq \sum_{k=t}^{t+H} \gamma^{k-t} \left( S \left| \pi_{k+n} - \pi_k \right|_{1,\infty} + A \left| d_s(\mathbb{P}^{\pi_{t+n}} \ldots \mathbb{P}^{\pi_k+n}) - d_s(\mathbb{P}^{\pi_t} \ldots \mathbb{P}^{\pi_k}) \right|_1 \right)
\]
(72)

given Eq. (64) and Eq. (65)

\[
\leq \sum_{k=t}^{t+H} \gamma^{k-t} \left\{ Snc_T + An(k-t)c_T \right\}
\]
(73)

by the limiting sum of geometric sequence and \( k - t \leq H \).

\[
\leq \frac{1}{1 - \gamma} (S + HA) n c_T
\]
(74)

**Definition 30.** Assuming \( d_1 \) is an arbitrary initial distribution, and \( \pi_1, \pi_2, \ldots, \pi_t \) are produced arbitrary sequence of policies that are slowly changing. We define the \( q_t(s) \) as the probability distribution over \( S \), and \( \mu_t(s) \) be the stationary distribution induced by \( \pi_t \). (Note \( \mu_t \) is not discounted, in contrast to discounted occupancy measure.)

\[
q_t(s) = \Pr(S_t = s) = \left\{ d_1 \mathbb{P}^{\pi_1} \mathbb{P}^{\pi_2} \ldots \mathbb{P}^{\pi_{t-1}} \right\}(s)
\]
(75)

\[
\mu_t(s) = \lim_{K \to \infty} \left\{ d_1(\mathbb{P}^{\pi_t})^K \right\}(s)
\]
(76)

**Lemma 31.** Given slowly changing by Def. 30, i.e. \( ||\pi_{t+1} - \pi_t||_{1,\infty} \leq c_T \), then we have \( ||q_t - \mu_t||_1 \leq c_T^2 \tau^2 + 2e^{-t/\tau} \)

**Proof.**

Let \( k + 1 \leq t \)

\[
||q_{k+1} - \mu_t||_1 = ||q_k \mathbb{P}^{\pi_k} - q_k \mathbb{P}^{\pi_t} + q_k \mathbb{P}^{\pi_k} - \mu_t \mathbb{P}^{\pi_t}||_1
\]
(77)

\[
\leq ||q_k \mathbb{P}^{\pi_k} - q_k \mathbb{P}^{\pi_t}||_1 + ||q_k \mathbb{P}^{\pi_t} - \mu_t \mathbb{P}^{\pi_t}||_1
\]
(78)
by Assumption 2

\[ \leq \|q_k^P - q_k^P\|_1 + e^{-1/\tau}\|q_k - \mu\|_1 \]  

(79)

by Lemma 23

\[ \leq (t - k)c_T + e^{-1/\tau}\|q_k - \mu\|_1 \]  

(80)

Then, by expanding the recursion we have

\[ \|q_t - \mu\|_1 = c_T \sum_{k=1}^{t-1} (t - k)e^{-(t-k)/\tau} + e^{-t/\tau}\|q_1 - \mu\|_1 \]  

(81)

by \(\|q_1 - \mu\|_1 \leq 2\)

\[ \leq c_T \sum_{k=1}^{t-1} (t - k)e^{-(t-k)/\tau} + 2e^{-t/\tau} \]  

(82)

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

\[ \leq c_T \tau^2 + 2e^{-t/\tau} \]  

(83)

**Lemma 13.** Assume \(T\) is sufficiently large such that \(c_T\tau^2 < \beta/4\), then we have

\[ q_t(s) \geq \beta/2, \quad \text{for} \quad t > \tau \ln(8/\beta) \]

**Proof.**

Given Lemma 31, we have

\[ \|q_t - \mu\|_1 \leq c_T^2 + 2e^{-t/\tau} \]  

(84)

By assuming \(c_T\tau^2 < \beta/4\) and let \(t > \tau \ln(8/\beta)\), we have

\[ \|q_t - \mu\|_1 \leq c_T^2 + 2e^{-t/\tau} < \frac{\beta}{4} + \frac{\beta}{4} = \frac{\beta}{2} \]  

(85)

By Assumption 1, we have \(\mu(s) \geq \beta\) for all \(s \in S\). Therefore, for any \(t > \tau \ln(8/\beta)\) and any \(s \in S\), we have

\[ q_t(s) \geq \mu(s) - |q_t(s) - \mu(s)| \]  

(86)

\[ \geq \mu(s) - \sum_{s'} |q_t(s') - \mu(s')| \]  

(87)

\[ = \mu(s) - \|q_t - \mu\|_1 \]  

(88)

\[ \geq \beta - \frac{\beta}{2} \]  

(89)

\[ = \frac{\beta}{2} \]  

(90)
Corollary 32. Assume $T$ is sufficiently large such that $c_T T^2 < \beta/4$, for all $s \in S$. While running MAIN, starting from any state $s_t \in S$ at time $t$, we have

$$v_T(t) \geq \beta/2, \quad \text{for } t' > t + \tau \ln(8/\beta)$$

Lemma 33. Let $\Omega_s = \{t^*_1, t^*_2, \ldots, t^*_n \}$ be the set of timesteps (in increasing order) that $s$ is visited. Denote $L_t^s$ be the step regret in $s$ at $t$. Then we have

\[ \text{Proof.} \]

Let $t_n^s \leq t < t_{n+1}^s$, $t^s_{|\Omega_s|+1} = T$, and $L_t^s = \sum_{a \in A} (\pi^*(a|s) - \pi_t(a|s)) G^{\pi^*}(s,a)$

we drop the superscript $s$ of $t_n^s$ for simplicity

$$\mathbb{E}_{\Omega_s} |L_t^s - L_{t_n}^s| = \mathbb{E}_{\Omega_s} \left| \sum_{a \in A} (\pi^*(a|s) - \pi_t(a|s)) G^{\pi^*}(s,a) - \sum_{a \in A} (\pi^*(a|s) - \pi_{t_n}(a|s)) G^{\pi^*}_{t_n}(s,a) \right|$$

use $\pi^u$, $\pi_{t,s}$ to denote the policy vector of state $s$, and $G^{\pi^u}_{t_n}$ for the feedback vector of state $s$

$$\leq \mathbb{E} \left| (\pi^u - \pi_{t,s} - G^{\pi^u}) - (\pi^u - \pi_{t_n,s} - G^{\pi^u}_{t_n}) \right|$$

$$= \mathbb{E} \left| (\pi^u - \pi_{t,s} - G^{\pi^u}) - (\pi^u - \pi_{t_n,s} - G^{\pi^u} - G^{\pi^u}) \right|$$

$$= \mathbb{E} \left| (\pi^u - \pi_{t,s} - G^{\pi^u}) - (\pi^u - \pi_{t_n,s} - G^{\pi^u} - G^{\pi^u}) \right|$$

$$= \mathbb{E} \left| (\pi_{t_n,s} - \pi_{t,s} - G^{\pi^u}) - (\pi_{t_n,s} - \pi_{t,s} - G^{\pi^u}) \right|$$

$$\leq \mathbb{E} \left\{ \left| (\pi_{t_n,s} - \pi_{t,s}, \frac{1}{1-\gamma}) \right| + \sum_{a} \left| (\pi^*(s,a) - \pi_{t_n}(s,a)) (G^{\pi^*}_{t_n}(s,a) - G^{\pi^*}(s,a)) \right| \right\}$$

given $||\pi_{t_n} - \pi_t||_{1,\infty} \leq (t - t_n)c_T$

$$\leq \mathbb{E} \left\{ \frac{1}{1-\gamma} (t - t_n)c_T + \sum_{a} \left| (\pi^*(s,a) - \pi_{t_n}(s,a)) (G^{\pi^*}_{t_n}(s,a) - G^{\pi^*}(s,a)) \right| \right\}$$

given $|G^{\pi^*}_{t_n}(s,a) - G^{\pi^*}(s,a)| \leq \frac{1}{1-\gamma} (S + HA)(t - t_n)c_T$

$$\leq \mathbb{E} \left\{ \frac{1}{1-\gamma} (t - t_n)c_T + \frac{2}{1-\gamma} (S + HA)(t - t_n)c_T \right\}$$

$$= \frac{2}{1-\gamma} \mathbb{E}(t - t_n)(S + HA + 0.5)c_T$$

Lemma 15. If Assumption 2 and 1 are satisfied. For a slowly changing LOCAL bandit, with a rate of $\mathcal{O}(1/\sqrt{T})$, enjoys $\mathcal{O}(\sqrt{(A+H)T} \ln A)$ expect regret in standard bandit problem under constant feedback delay $H$. Then in the case of sticky setting, it enjoys $\mathcal{O} \left( \tau \sqrt{(HA+S)T} \beta(1-\gamma) \right)$ regret.
Proof.

Let \( \Omega_s = \{t^s_1, t^s_2, \ldots, t^s_{|\Omega_s|}\} \) be the set of timesteps (in increasing order) that \( s \) is visited, and let \( t^s_{|\Omega_s|+1} = T+1 \). Let \( L^s_t \) be the step regret in \( s \) at \( t \), and we drop the superscript \( s \) of \( t^s_n \) for simplicity.

\[
\mathbb{E}_{\Omega_s} \mathcal{R}_s^{Q^*_t}(T|H) = \mathbb{E}_{\Omega_s} \sum_{t=1}^{T} L^s_t
\]

(100)

\[
= \sum_{t=1}^{T} L^s_t
\]

(101)

\[
= \mathbb{E}_{\Omega_s} \sum_{n=1}^{|\Omega_s|} \sum_{t=t_n}^{t_{n+1}-1} L^s_t
\]

(102)

\[
\leq \mathbb{E}_{\Omega_s} \sum_{n=1}^{|\Omega_s|} \sum_{t=t_n}^{t_{n+1}-1} \left[ L^s_t + |L^s_t - L^s_{t_n}| \right]
\]

(103)

Given lemma 33

\[
\leq \mathbb{E}_{\Omega_s} \sum_{n=1}^{|\Omega_s|} \sum_{t=t_n}^{t_{n+1}-1} L^s_t + \mathbb{E}_{\Omega_s} \sum_{n=1}^{|\Omega_s|} \sum_{t=t_n}^{t_{n+1}-1} \frac{2(t - t_n)}{(1 - \gamma)} \left( S + HA + \frac{1}{2} \right) c_T
\]

(104)

As \( 2(S + HA + 0.5)c_T/(1 - \gamma) \) does not depends on \( t \)

\[
= \mathbb{E}_{\Omega_s} \sum_{n=1}^{|\Omega_s|} \sum_{t=t_n}^{t_{n+1}-1} L^s_t + \mathbb{E}_{\Omega_s} \sum_{n=1}^{|\Omega_s|} \sum_{t=t_n}^{t_{n+1}-1} \frac{2}{(1 - \gamma)} \left( S + HA + \frac{1}{2} \right) c_T
\]

(105)

\[
= \mathbb{E}_{\Omega_s} \sum_{n=1}^{|\Omega_s|} \sum_{t=t_n}^{t_{n+1}-1} L^s_t + \frac{2}{(1 - \gamma)} \left( S + HA + \frac{1}{2} \right) c_T T
\]

(106)

By Corollary 32

\[
\leq \mathbb{E}_{\Omega_s} \sum_{n=1}^{|\Omega_s|} \left( \tau \ln(8/\beta) + 2/\beta \right) L^s_{t_n} + \frac{2}{(1 - \gamma)} \left( S + HA + \frac{1}{2} \right) c_T T
\]

(107)

\[
\leq \tilde{O}(\tau + 1/\beta) \mathbb{E}_{\Omega_s} \sum_{n=1}^{|\Omega_s|} L^s_{t_n} + \mathcal{O} \left( \frac{S + HA}{1 - \gamma} \sqrt{T} \right)
\]

(108)

\[
\leq \tilde{O} \left( (\tau + \frac{1}{\beta}) \mathcal{R}_s^{Q^*_t}(\beta T|H) \right) + \mathcal{O} \left( \frac{S + HA}{1 - \gamma} \sqrt{T} \right)
\]

(109)

Given the choice of \( \mathcal{R}_s^{Q^*_t}(T|H) = \mathcal{O}(\sqrt{(A + H)T \ln A}) \) in standard bandits.

\[
= \tilde{O} \left( (\tau + \frac{1}{\beta}) \sqrt{\beta (A + H) T} + \left( \frac{S + HA}{1 - \gamma} \sqrt{T} \right) \right)
\]

(110)

\[
= \tilde{O} \left( \frac{\tau(S + HA)}{\sqrt{\beta(1 - \gamma)}} \sqrt{T} \right)
\]

(111)
B.3 Stage III: Delayed feedback

Lemma 19. Main is slowly changing in \( \tilde{M} \) if Base is slowly changing.

Proof. Similar to Lemma 8, we use \( \phi^s \) to denote the policy of Local\(_s\). In addition, we use \( \phi^{s,h}_t \) to denote the policy of Base\(_h\) of 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.

\[
||\tilde{\pi}_{t+1} - \tilde{\pi}_t||_{1,\infty} = ||\phi^{s,h}_{t+1} - \phi^{s,h}_t||_{1,\infty} \leq c_T \tag{112}
\]

This simply follows the fact that only one Base\(_{s_t-H}^{h_t} \) is updated at time \( t \).  \( \blacksquare \)