Near-Optimal Regret for Adversarial MDP with Delayed Bandit Feedback

Tiancheng Jin tiancheng.jin@usc.edu
University of Southern California

Tal Lancewicki lancewicki@mail.tau.ac.il
Tel Aviv University

Haipeng Luo haipengl@usc.edu
University of Southern California

Yishay Mansour mansour.yishay@gmail.com
Tel Aviv University and Google Research

Aviv Rosenberg avivros007@gmail.com
Tel Aviv University

Abstract

The standard assumption in reinforcement learning (RL) is that agents observe feedback for their actions immediately. However, in practice feedback is often observed in delay. This paper studies online learning in episodic Markov decision process (MDP) with unknown transitions, adversarially changing costs, and unrestricted delayed bandit feedback. More precisely, the feedback for the agent in episode k is revealed only in the end of episode $k+d^k$, where the delay d^k can be changing over episodes and chosen by an oblivious adversary. We present the first algorithms that achieve near-optimal $\sqrt{K+D}$ regret, where K is the number of episodes and $D=\sum_{k=1}^{K} d^k$ is the total delay, significantly improving upon the best known regret bound of $(K+D)^{2/3}$.

Keywords: Regret minimization, Adversarial MDPs, Delayed feedback, Reinforcement Learning, Online Learning.

1. Introduction

Delayed feedback has become a fundamental challenge that sequential decision making algorithms must face in almost every real-world application. Notable examples include communication between agents (Chen et al., 2020), video streaming (Changuel et al., 2012) and robotics (Mahmood et al., 2018). Broadly, delays occur either for computational reasons, e.g., in autonomous vehicles and wearable technology, or when they are an inherent part of the environment like healthcare, finance and recommendation systems.

Although a prominent challenge in practice, there is only limited theoretical literature on delays in reinforcement learning (RL). Recently, Howson et al. (2021) studied regret minimization in episodic Markov decision processes (MDPs) but assume that the delays (and costs) are *stochastic*, i.e., sampled i.i.d from a fixed (unknown) distribution, which is a limiting assumption since it does not allow dependencies between costs and delays that are very common in practice. The case of *adversarial* delays and costs was also studied recently (Lancewicki et al., 2020). However, they focus on *full-information* feedback where the learner observes the entire cost function, which is not realistic in many applications, and obtain only sub-optimal regret bounds for *bandit* feedback (where the learner observes only the costs on the traversed trajectory).

In this paper we significantly advance our understanding of delayed feedback in adversarial MDPs with bandit feedback. More precisely, we consider episodic MDPs with unknown transition function, adversarially changing costs and unrestricted delayed bandit feedback, i.e., the learner observes the costs suffered in episode k only in the end of episode $k+d^k$ where the sequence of delays $\{d^k\}_{k=1}^K$ are chosen by an oblivious adversary. We develop the first algorithms for this setting that achieve near-optimal regret and provide a major improvement over the currently best known regret bound (Lancewicki et al., 2020) - see Table 1 for more details.

Table 1: Regret bounds for Adversarial MDPs with unknown transition and unrestricted delayed bandit feedback. K is the number of episodes, D is the total delay, H is the horizon, S is the number of states and A is the number of actions. Algorithms presented in this paper appear in grey.

Algorithm	Regret	Efficient	Regret w.h.p
D-OPPO Lancewicki et al. (2020)	$\widetilde{O}(HS\sqrt{A}K^{2/3} + H^2 D^{2/3})$	✓	✓
Delayed Hedge	$\widetilde{O}(H^2 S\sqrt{AK} + H^{3/2}\sqrt{SD})$	✗	✓
Delayed UOB-FTRL	$\widetilde{O}(H^2 S\sqrt{AK} + H^{3/2} SA\sqrt{D})$	✓	✗
Delayed UOB-REPS	$\widetilde{O}(H^2 S\sqrt{AK} + (HSA)^{1/4} \cdot H\sqrt{D})$	✓	✓
Lower bound Lancewicki et al. (2020)	$\Omega(H^{3/2}\sqrt{SAK} + H\sqrt{D})$		

In the following paragraph we provide an overview of our contributions and the structure of the paper. In Section 3 we devise an inefficient Hedge (Freund and Schapire, 1997) based algorithm that treats every deterministic policy as an arm. This can be seen as a warm-up – a relatively simple and elegant solution that shows that order $\sqrt{K+D}$ regret is attainable with delayed bandit feedback. Moreover, our adaptation of Hedge to the setting of adversarial MDP with unknown transition and bandit feedback presents highly non-trivial algorithmic and technical features that may be of independent interest. Then, we focus on the pressing question: *Can delayed bandit feedback be handled both optimally and efficiently?* We answer this affirmatively by presenting two efficient algorithms with near-optimal regret. Through our unique analysis and algorithmic design, we shed light on the great challenges of handling efficiently delayed bandit feedback. In Section 4 we consider a relatively standard algorithm we call Delayed UOB-FTRL, based on the Follow the Regularized Leader (FTRL) framework, and focus on a unique novel analysis that may be of independent interest. As seen in Table 1, our analysis of Delayed UOB-FTRL shows regret similar to the inefficient Delayed Hedge. However, it has worse dependence on S and A, and has regret guarantee on expectation rather than with high probability (w.h.p). In Section 5 we propose our final solution which is mainly algorithmic: we introduce the algorithm Delayed UOB-REPS that has a novel importance-sampling estimator which generalizes the standard estimator and accommodates it to the delays. This approach allows us to follow the path of more standard analysis, but most importantly, ensures w.h.p the best regret so far (see Table 1). The first term of the regret bound matches the best known regret for adversarial MDP with non-delayed bandit feedback (Jin et al., 2020a), while the second term matches the lower bound of Lancewicki et al. (2020) up to a factor of $(HSA)^{1/4}$.

1.1 Additional Related Work

Delays in RL. While delays are popular in the practical RL literature (Schuitema et al., 2010; Liu et al., 2014; Changuel et al., 2012; Mahmood et al., 2018; Derman et al., 2021), there is limited theoretical literature on the subject. Most previous work (Katsikopoulos and Engelbrecht, 2003; Walsh et al., 2009) considered constant delays in observing the current state. However, the challenges in that setting are different than the ones considered in this paper (see Lancewicki et al. (2020) for more details). As discussed in the introduction, most related to this paper are the recent works of Lancewicki et al. (2020) and Howson et al. (2021).

Delays in multi-arm bandit (MAB). Delays were extensively studied in MAB and optimization both in the stochastic setting (Agarwal and Duchi, 2012; Vernade et al., 2017, 2020; Pike-Burke et al., 2018; Cesa-Bianchi et al., 2018; Zhou et al., 2019; Gael et al., 2020; Lancewicki et al., 2021; Cohen et al., 2021a), and the adversarial setting (Quanrud and Khashabi, 2015; Cesa-Bianchi et al., 2016; Thune et al., 2019; Bistritz et al., 2019; Zimmert and Seldin, 2020; Ito et al., 2020; Gyorgy and Joulani, 2021; van der Hoeven and Cesa-Bianchi, 2021). However, as discussed in Lancewicki et al. (2020), delays introduce new challenges in MDPs that do not appear in MAB.

Regret minimization in RL. There is a rich literature on regret minimization in both stochastic (Jaksch et al., 2010; Azar et al., 2017; Jin et al., 2018, 2020b; Yang and Wang, 2019; Zanette et al., 2020a,b) and adversarial (Zimin and Neu, 2013; Rosenberg and Mansour, 2019a,b, 2021; Jin et al., 2020a; Jin and Luo, 2020; Cai et al., 2020; Shani et al., 2020; Luo et al., 2021; Jin et al., 2021) MDPs. Note that regret minimization in standard episodic MDPs is a special case of the model considered in this paper where $d^k = 0$ for every episode k.

2. Preliminaries

We consider the problem of learning adversarial MDPs under delayed feedback. A finite-horizon episodic MDP is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, p, \{c^k\}_{k=1}^K)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces of sizes $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$, respectively, H is the horizon (i.e., episode length) and K is the number of episodes. $p : \mathcal{S} \times \mathcal{A} \times [H] \to \Delta_{\mathcal{S}}$ is the *transition function* which defines the transition probabilities. That is, $p_h(s'|s,a)$ is the probability to move to state s' when taking action a in state s at time h. $\{c^k : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]\}_{k=1}^K$ are *cost functions* which are chosen by an *oblivious adversary*, such that $c_h^k(s,a)$ is the cost of taking action a in state s at time h of episode k.

A *policy* $\pi : \mathcal{S} \times [H] \to \Delta_{\mathcal{A}}$ is a function such that $\pi_h(a|s)$ is the probability to take action a when visiting state s at time h. The value $V_h^{\pi,p'}(s;c)$ is the expected cost of π with respect to cost function c and transition function p' starting from state s in time h, i.e., $V_h^{\pi,p'}(s;c) = \mathbb{E}^{\pi,p'}\left[\sum_{h'=h}^H c_{h'}(s_{h'}, a_{h'}) \mid s_h = s\right]$, where $\mathbb{E}^{\pi,p'}[\cdot]$ denotes the expectation with respect to policy π and transition function p', that is, $a_{h'} \sim \pi_{h'}(\cdot \mid s_{h'})$ and $s_{h'+1} \sim p'_{h'}(\cdot \mid s_{h'}, a_{h'})$.

Learner-environment interaction. At the beginning of episode k, the learner picks a policy π^k, and starts in an initial state $s_1^k = s_{\text{init}}$. In each time $h \in [H]$, it observes the current state s_h^k, draws an action from the policy $a_h^k \sim \pi_h^k(\cdot|s_h^k)$ and transitions to the next state $s_{h+1}^k \sim p_h(\cdot|s_h^k, a_h^k)$. The feedback of episode k contains the cost function over the agent's trajectory $\{c_h^k(s_h^k, a_h^k)\}_{h=1}^H$ (i.e., bandit feedback). This feedback is observed only at the end of episode $k + d^k$, where the *delays* $\{d^k\}_{k=1}^K$ are unknown and chosen by the oblivious adversary together with the costs. If $d^k = 0$ for all k, this model scales down to standard online learning in adversarial MDP.

Occupancy measure. Given a policy π and a transition function p', the *occupancy measure* $q^{\pi,p'} \in [0,1]^{HS^2A}$ is a vector, where $q_h^{\pi,p'}(s,a,s')$ is the probability to visit state s at time h, take action a and transition to state s'. We also denote $q_h^{\pi,p'}(s,a) = \sum_{s'} q_h^{\pi,p'}(s,a,s')$ and $q_h^{\pi,p'}(s) = \sum_a q_h^{\pi,p'}(s,a)$. By Rosenberg and Mansour (2019a), the occupancy measure encodes the policy and the transition function through the following relations:

$$\pi_h(a \mid s) = \frac{q_h^{\pi,p'}(s,a)}{q_h^{\pi,p'}(s)} \quad ; \quad p'_h(s' \mid s,a) = \frac{q_h^{\pi,p'}(s,a,s')}{q_h^{\pi,p'}(s,a)}.$$

The set of all occupancy measures with respect to an MDP $\mathcal{M}$ is denoted by $\Delta(\mathcal{M})$. Importantly, the value of a policy from the initial state can be written as the dot product between its occupancy measure and the cost function, i.e., $V_1^{\pi,p'}(s_{\text{init}}; c) = \langle q^{\pi,p'}, c\rangle$. Whenever p' is omitted from the notations $q^{\pi,p'}$ and $V^{\pi,p'}$, this means that they are with respect to the true transition function p.

Regret. The learner's performance is measured by the *regret* which is the difference between the cumulative expected cost of the learner and the best fixed policy in hindsight:

$$R_K = \sum_{k=1}^K V_1^{k,\pi^k}(s_{\text{init}}) - \min_\pi \sum_{k=1}^K V_1^{k,\pi}(s_{\text{init}}) = \sum_{k=1}^K \langle q^{\pi^k}, c^k\rangle - \min_{q \in \Delta(\mathcal{M})} \sum_{k=1}^K \langle q, c^k\rangle,$$

where $V_h^{k,\pi}(s) = V_h^{\pi,p}(s; c^k)$.

Confidence set. Since the transition function is unknown, we maintain standard Bernstein-based confidence sets $\mathcal{P}^k$ for each episode k that contain p with high-probability. For the exact definition of $\mathcal{P}^k$ see Algorithm 5, and the fact that $p \in \mathcal{P}^k$ for every k w.h.p is proved for example in Jin et al. (2020a) (for more details see the appendix). Using $\mathcal{P}^k$ we can define a confidence set of occupancy measures by

$$\Delta(\mathcal{M}, k) = \{q^{\pi,p'} \mid \pi \in (\Delta_{\mathcal{A}})^{\mathcal{S} \times [H]}, p' \in \mathcal{P}^k\},$$

which is a polytope with polynomial constraints as shown in (Rosenberg and Mansour, 2019a). Note that as long as $p \in \mathcal{P}^k$, $\Delta(\mathcal{M}) \subseteq \Delta(\mathcal{M}, k)$.

Additional notations. In general, episode indices always appear as superscripts and in-episode steps as subscripts. $n_h^k(s,a,s')$ denotes the total number of visits at state s in which the agent took action a at time h and transitioned to s' by the end of episode $k-1$. $\bar{p}_h^k(s'|s,a) = n_h^k(s,a,s') / \max\{n_h^k(s,a), 1\}$ is the empirical mean estimation of $p_h(s'|s,a)$, where $n_h^k(s,a) = \sum_{s'} n_h^k(s,a,s')$. $\mathcal{F}^k = \{j : j + d^j = k\}$ denotes the set of episodes such that their

Algorithm 1 Delayed Hedge

1: **Initialization:** Set ω^1 to be the uniform distribution over all deterministic policies.
2: **for** $k = 1, 2, ..., K$ **do**
3: Execute policy π^k sampled from ω^k, observe trajectory $\{s_h^k, a_h^k\}_{h=1}^H$.
4: Update confidence set $\mathcal{P}^k$, compute upper occupancy bound u^k and exploration bonus b^k by:

$$u_h^k(s,a) = \max_{p' \in \mathcal{P}^k} \sum_{\pi \in \Omega} \omega^k(\pi) q_h^{\pi,p'}(s,a) \quad ; \quad b^k(\pi) = \max_{p' \in \mathcal{P}^k} \|q^{\pi,\bar{p}^k} - q^{\pi,p'}\|_1.$$

5: **for** $j : j + d^j = k$ **do**
6: Observe costs $\{c_h^j(s_h^j, a_h^j)\}_{h=1}^H$, compute loss estimator $\hat{c}^j$ defined in Equation (1), and estimated loss by $\widehat{\ell}^j(\pi) = \langle q^{\pi,\bar{p}^j}, \widehat{c}^j \rangle$.
7: **end for**
8: Update policy distribution ω^{k+1} by: $\omega^{k+1}(\pi) \propto \omega^k(\pi) \cdot \exp\left(\eta b^k(\pi) - \eta \sum_{j:j+d^j=k} \widehat{\ell}^j(\pi)\right)$.
9: **end for**

feedback arrives in the end of episode k. The notations $\widetilde{O}(\cdot)$ and $\lesssim$ hide constant and poly-logarithmic factors including $\log(K/\delta)$ for some confidence parameter δ, the indicator of event E is denoted by $\mathbb{I}\{E\}$, and $x \vee y = \max\{x, y\}$.

Simplifying assumptions. Throughout this paper we assume that K and $D = \sum_{k=1}^K d^k$ are known and that the maximal delay $d_{max} = \max_k d^k \leq \sqrt{D}$. Both of these assumptions are made only for simplicity of presentation and can be easily relaxed using standard *doubling* and *skipping* procedures as shown for example by Thune et al. (2019); Lancewicki et al. (2020); Bistritz et al. (2021). In addition, we focus on the case of non-delayed trajectory feedback, where the learner observes the trajectory immediately at the end of the episode and only the feedback regarding the cost is delayed. Delayed trajectory feedback mainly affects approximation errors and the ideas presented in Lancewicki et al. (2020) for handling such delay apply to our case as well.

3. Delayed Hedge

In this section, we consider running a Hedge-based algorithm over all $\Omega = \mathcal{A}^{\mathcal{S} \times [H]}$ deterministic policies. Algorithm 1, which we call Delayed Hedge, is inefficient but gives the first order-optimal regret bounds for adversarial MDP with delayed bandit feedback. Although the main issue that Delayed Hedge tackles is delayed feedback, we note that there are many additional challenges introduced by the unknown transitions and the bandit feedback when we maintain a distribution over policies instead of a single stochastic policy.

Delayed Hedge maintains a distribution ω^k over deterministic policies (starting from a uniform distribution), and in the beginning of episode k samples a policy π^k to execute. Thus, the expected loss incurred in episode k is $\sum_{\pi \in \Omega} \omega^k(\pi) \langle q^{\pi,p}, c^k \rangle$. The algorithm updates the distribution ω^k based on the exponential weights update, for which we need to compute an estimated loss for every policy $\pi \in \Omega$.

To do so, first we estimate the cost in each state-action pair. Due to unknown dynamics, following (Jin et al., 2020a) we use the confidence sets to compute optimistic importance weighted estimator that will induce exploration:

$$\hat{c}_h^k(s,a) = \frac{c_h^k(s,a)\mathbb{I}\{s_h^k = s, a_h^k = a\}}{u_h^k(s,a) + \gamma}, \tag{1}$$

where $u_h^k(s,a) = \max_{p' \in \mathcal{P}^k} \sum_{\pi \in \Omega} \omega^k(\pi) q_h^{\pi,p'}(s,a)$ is an upper occupancy bound on the probability to visit (s,a) in step h of episode k, and γ is a small bias added for high probability regret (Neu, 2015).

Then, we use the empirical transition function $\bar{p}^k$ to compute the estimated loss $\widehat{\ell}^k(\pi) = \langle q^{\pi,\bar{p}^k}, \widehat{c}^k \rangle$ for each policy π. To ensure optimism, we introduce the exploration bonus $b^k(\pi) = \max_{p' \in \mathcal{P}^k} \|q^{\pi,\bar{p}^k} - q^{\pi,p'}\|_1$. As long as the real transition function p is in confidence set $\mathcal{P}^k$, optimism is indeed ensured in the sense that $\langle q^{\pi,\bar{p}^k}, c \rangle - b^k(\pi)$ is always no more than the true cost $\langle q^{\pi,p}, c \rangle$ for any policy π and $[0,1]$-valued cost function c.

Algorithm 2 Delayed UOB-FTRL

```
Initialization: Set π^1 to be uniform policy.
for k = 1, 2, ..., K do
    Execute policy π^k, observe trajectory {s_h^k, a_h^k}_{h=1}^H, update confidence set P^k and compute upper occupancy
    bound u_h^k(s,a) = max_{p'∈P^k} q_h^{π^k,p'}(s,a).
    for j : j + d^j = k do
        Observe costs {c_h^j(s_h^j, a_h^j)}_{h=1}^H and compute the standard loss estimator ĉ^j by
            ĉ_h^j(s,a) = c_h^j(s,a) I{s_h^j = s, a_h^j = a} / (u_h^j(s,a) + γ).        (2)
    end for
    Compute occupancy measure by: q^{k+1} = argmin_{q ∈ ∩_{j=1}^{k+1} Δ(M,j)} ⟨q, Σ_{j+d^j≤k} ĉ^j⟩ + φ(q),
    where φ(q) = (1/η) Σ_{h,s,a,s'} q_h(s,a,s') log q_h(s,a,s') is the Shannon entropy regularizer.
    Update policy: π_h^{k+1}(a | s) = q_h^{k+1}(s,a) / q_h^{k+1}(s).
end for
```

With the estimated loss and the exploration bonus for each π, the distribution ω^{k+1} is now updated in a manner similar to that of Gyorgy and Joulani (2021): $\omega^{k+1}(\pi) \propto \omega^k(\pi) \cdot \exp\left(\eta b^k(\pi) - \eta \sum_{j:j+d^j=k} \widehat{\ell^j}(\pi)\right)$. Note that all information required for this update has been received by the learner at the end of episode k. With the help of all these definitions, we prove the following regret bound for Delayed Hedge, and defer the details to Appendix A including the complete algorithm and regret analysis.

Theorem 1 *With appropriate choices of parameters, Delayed Hedge ensures* $R_K = \widetilde{O}\left(H^2 S\sqrt{AK} + H^{3/2}\sqrt{SD}\right)$ *with high probability (w.h.p.).*

4. Delayed UOB-FTRL

In this section, we adjust the UOB-REPS algorithm (Jin et al., 2020a) to delayed feedback and present the Delayed UOB-FTRL algorithm (Algorithm 2) - the first efficient algorithm to attain order-optimal regret for adversarial MDP with delayed bandit feedback. The proof is based on a novel analysis without additional changes to the algorithm. Namely, we use standard loss estimators (defined in Equation (2)). Our algorithm is based on the Follow-the-Regularized-Leader (FTRL) framework, which is widely used for deriving online learning algorithm in adversarial environments. Notable examples are Zimin and Neu (2013) that applies FTRL over occupancy measure space to solve the adversarial MDP problem with known transition, and Zimmert and Seldin (2020) that uses FTRL to achieve optimal regret for MAB with delayed feedback.

In our context, in the beginning of episode k, FTRL computes $q^k = \arg\min_{q \in \cap_{j=1}^{k} \Delta(\mathcal{M},j)} \langle q, \widehat{L}_k^{obs} \rangle + \phi(q)$ and then the policy π^k to be played in the episode is extracted from q^k. Here, $\widehat{L}_k^{obs} = \sum_{j+d^j<k} \widehat{c}^j$ is the cumulative losses observed prior to episode k, and $\phi(q) = \frac{1}{\eta}\sum_{h,s,a,s'} q_h(s,a,s') \log q_h(s,a,s')$ is the Shannon entropy regularizer. Thus, our algorithm can be regarded as a direct extension to MDP of FTRL for delayed feedback. However, unlike the successes in MAB, it is highly unclear whether optimal regret could be obtained in adversarial MDPs with FTRL even if the transition function is known.

In Theorem 2, we show that Delayed UOB-FTRL enjoys order-optimal regret. Through the key steps of the analysis, we shall take a closer look at the key reason why traditional analysis fails: in occupancy measure space, the interplay between different entries of loss functions is significantly harder to analyze. Thus, many critical properties used in Zimmert and Seldin (2020) do not hold anymore. The complete algorithm and proof are deferred to Appendix B.

Theorem 2 *With appropriate choices of parameters, Delayed UOB-FTRL (Algorithm 2) ensures* $\mathbb{E}[R_K] = \widetilde{O}\big(H^2 S\sqrt{AK} + HSA\sqrt{HD}\big)$.

Proof sketch of Theorem 2 Let $q^\star = q^{\pi^\star, p}$ be the occupancy measure associated with the optimal policy $\pi^\star$. We adopt the regret decomposition of Jin et al. (2020a):

$$R_K = \underbrace{\sum_{k=1}^{K} \left\langle q^{\pi^k} - q^k, c^k \right\rangle}_{\textsc{Est}} + \underbrace{\sum_{k=1}^{K} \left\langle q^k, c^k - \widehat{c}^k \right\rangle}_{\textsc{Bias}_1} + \underbrace{\sum_{k=1}^{K} \left\langle q^k - q^\star, \widehat{c}^k \right\rangle}_{\textsc{Reg}} + \underbrace{\sum_{k=1}^{K} \left\langle q^\star, \widehat{c}^k - c^k \right\rangle}_{\textsc{Bias}_2}.$$

$\textsc{Est}$, $\textsc{Bias}_1$ and $\textsc{Bias}_2$ are standard and bounded in Jin et al. (2020a) w.h.p by $\widetilde{O}(\gamma HSAK + H^2 S\sqrt{AK} + H/\gamma)$. Now, we focus on bounding $\textsc{Reg}$. To this end, we denote by $\widehat{L}_k = \sum_{j=1}^{k-1} \widehat{c}^k$ the non-delayed cumulative loss, and introduce the convex conjugate functions $F_k^\star$ with respect to the regularizer $\phi(\cdot)$: $F_k^\star(x) = -\min_{q\in\Delta(\mathcal{M},k)} \{\phi(q) - \langle x, q\rangle\}$.

We now use $F_k^\star$ to decompose $\textsc{Reg}$ into the following three terms as

$$\begin{aligned} &\sum_{k=1}^{K} -F_k^\star(-\widehat{L}_k^{\text{obs}}) + \left\langle q^k, \widehat{c}^k \right\rangle + F_k^\star(-\widehat{L}_k^{\text{obs}} - \widehat{c}^k) + \sum_{k=1}^{K} -F_k^\star(-\widehat{L}_k - \widehat{c}^k) + F_k^\star(-\widehat{L}_k) - \left\langle q^\star, \widehat{c}^k \right\rangle \\ &+ \sum_{k=1}^{K} \Big\{ -F_k^\star(-\widehat{L}_k^{\text{obs}} - \widehat{c}^k) + F_k^\star(-\widehat{L}_k^{\text{obs}}) - \Big(-F_k^\star(-\widehat{L}_k - \widehat{c}^k) + F_k^\star(-\widehat{L}_k) \Big) \Big\}. \end{aligned} \tag{3}$$

The first term is associated with the unseen loss $\widehat{c}^k$. It is relatively standard and bounded by $\widetilde{O}(\eta HSAK)$ w.h.p. The second term can be regarded as the regret of a "cheating" algorithm which does not suffer delay and sees one step into the future. This term can be bounded by $\widetilde{O}(H/\eta)$ similarly to Gyorgy and Joulani (2021). The third term which only relates to delayed feedback, is the most critical object in the analysis.

In the previous work of Zimmert and Seldin (2020) for *multi-arm bandit*, the authors managed to rewrite and then upper bound the delay-caused term for every episode k by

$$\int_0^1 \left\langle \widehat{c}^k, \nabla F_k^\star(-\widehat{L}_k^{\text{obs}} - x\widehat{c}^k) - \nabla F_k^\star(-\widehat{L}_k - x\widehat{c}^k) \right\rangle dx \le \eta \sum_{i\in[N]} p^k(i) \cdot \widehat{c}^k(i) \cdot \Big(\widehat{L}_k(i) - \widehat{L}_k^{\text{obs}}(i) \Big),$$

where $[N]$ is the set of arms and $p^k(i)$ is the probability that the algorithm chooses arm i in episode k. Here, the first step uses Newton-Leibniz theorem and the differentiability of convex conjugates, and the second step follows directly from (Zimmert and Seldin, 2020, Lemma 3). Importantly, the second step is largely based on the specific structure of the simplex (over which MAB algorithms operate), which yields the simple behavior of FTRL-based algorithms (e.g., EXP3). Specifically, it is based on the following observation. Suppose that we increase the cumulative loss of arm i. Now consider the behavior of $p(i')$, the probability of taking arm i' where p is computed from the FTRL framework. One can verify that $p(i')$ will increase for $i' \neq i$ and decrease for $i' = i$. In other words, the relationship between any pair of arms is competitive, and this property is critical to achieve the optimal regret with delayed feedback in Zimmert and Seldin (2020).

However, this property does not hold for MDPs because the constraints of the transition function can dictate positive correlation between entries of the occupancy measure. Similarly, consider two state-action pairs (s, a, h) and (s', a', h') from different states. It is highly unclear whether increasing the cumulative loss of (s, a, h) will increase or decrease the probability $q_{h'}(s', a')$ of reaching s' in time h' and taking action a'. In fact, the relation is related to the specific transition function of the MDP. For example, the FTRL algorithm will decrease the probability in the cases where taking action a at state s in step h is necessary to reach (s', a', h'), and will increase in other cases where not taking action a at state s of step h is necessary.

Therefore, an alternative analysis is required in our case. Specifically, we are able to bound the delayed-caused term by

$$\int_0^1 \left\langle \widehat{c}^k, \nabla F_k^\star(-\widehat{L}_k^{\text{obs}} - x\widehat{c}^k) - \nabla F_k^\star(-\widehat{L}_k - x\widehat{c}^k) \right\rangle dx \le 2 \left\| \widehat{c}^k \right\|_{\nabla^{-2}\phi(\xi)} \left\| \widehat{L}_k - \widehat{L}_k^{\text{obs}} \right\|_{\nabla^{-2}\phi(\xi)}$$
$$\le 2\eta \sum_{j=1, j+d^j \ge k}^{k-1} \left(\sum_{h,s,a} \widehat{c}_h^k(s,a) \right) \cdot \left(\sum_{h,s,a} \widehat{c}_h^j(s,a) \right)$$

where the first step uses the properties of convex conjugates for some valid occupancy measure ξ (See Lemma 14 for more details) with $\|x\|_M = \sqrt{x^\top M x}$ being the matrix norm for any vector x and positive definite matrix M, and the second step follows from the facts that $\nabla^{-2}\phi(\xi)$ is a diagonal matrix with values $\{\eta \cdot \xi_h(s,a) : \forall (h,s,a)\}$ on its diagonal and $\xi_h(s,a) \le 1$.

While we managed to overcome the complex dependencies between different states in the MDP, it comes at the price of a looser regret bound. The final bound does not have $q_h^k(s,a)$ in the summations which leads to an extra factor of SA. This follows from the application of Hölder's inequality and the relaxation of intermediate occupancy measure ξ.

Taking the summation over all episodes, we have that the third term in Equation (3) is bounded by $\widetilde{O}(\eta H^2 S^2 A^2 D)$ in expectation. Finally, with proper choice of the parameters η, γ and δ, combining the bounds for EST, BIAS$_1$, BIAS$_2$ and the three terms in Equation (3) finishes the proof. ■

5. Delayed UOB-REPS with Delay-adapted Estimator

Finally, we present our last algorithm, Delayed UOB-REPS equipped with our novel importance sampling estimator which we call *delay-adapted importance sampling estimator*. The algorithm appears as Algorithm 3 and in its full version together with the analysis for known and unknown dynamics in Appendices C and D.

Much like Delayed UOB-FTRL, the algorithm is efficient; but it outperforms Delayed UOB-FTRL in two important aspects: (i) it guarantees high-probability regret bound (and not only expected regret), and (ii) the delay term in its regret bound is tighter. In fact, as long as $A \le S$ (which happens in most cases), it obtains an improvement even on the regret of the inefficient Delayed Hedge algorithm.

To maintain the occupancy measures q^k from which the executed policies π^k are extracted, Delayed UOB-REPS uses the Online Mirror Decent (OMD) update rule:

$$q^{k+1} = \operatorname*{arg\,min}_{q \in \Delta(\mathcal{M}, k+1)} \eta \left\langle q, \sum_{j: j+d^j = k} \hat{c}^j \right\rangle + \mathrm{KL}(q \parallel q^k),$$

where η is a learning rate and $\mathrm{KL}(q \parallel q') = \sum_{h,s,a} q_h(s,a) \ln \frac{q_h(s,a)}{q'_h(s,a)} + q'_h(s,a) - q_h(s,a)$ is the unnormalized KL-divergence. We note that OMD is standard in the O-REPS literature, and has similar guarantees to FTRL. In this case, OMD will be much more useful than FTRL because we can utilize its update rule to prove certain properties for the relation between consecutive occupancy measures (see Lemma 29).

We do not use the standard importance sampling estimator, but the following delay-adapted estimator:

$$\hat{c}_h^k(s,a) = \frac{c_h^k(s,a) \mathbb{I}\{s_h^k = s, a_h^k = a\}}{\max\{u_h^k(s,a), u_h^{k+d^k}(s,a)\} + \gamma}. \tag{4}$$

The delay-adapted estimator specifically tackles one of the main technical challenges in analyzing algorithms under delayed feedback (especially in MDPs) – bound their stability. It is a biased estimator, and in fact has larger bias than the standard importance sampling estimator, but allows us to directly control the stability of the algorithm.

To describe the intuition behind the delay-adapted estimator, let us first consider a fixed delay $d^k = d$. The policy π^{k+d} is updated based on the episodes $1, ..., k-1$. Thus, playing π^{k+d} at episode k is equivalent to running OMD on the same loss estimators but in a non-delayed environment. Standard analysis for delayed feedback (e.g., Thune et al.

Algorithm 3 Delayed UOB-REPS with Delay-adapted Estimator

Initialization: Set π^1 to be uniform policy.
for $k = 1, 2, ..., K$ **do**
  Execute policy π^k, observe trajectory $\{s_h^k, a_h^k\}_{h=1}^H$, update confidence set $\mathcal{P}^k$ and compute upper occupancy bound $u_h^k(s,a) = \max_{p' \in \mathcal{P}^k} q_h^{\pi^k, p'}(s,a)$.
  for $j : j + d^j = k$ **do**
    Observe costs $\{c_h^j(s_h^j, a_h^j)\}_{h=1}^H$ and compute the delay-adapted cost estimator $\hat{c}^j$ by Equation (4).
  end for
  Update occupancy measure by: $q^{k+1} = \arg\min_{q \in \Delta(\mathcal{M}, k+1)} \eta \left\langle q, \sum_{j \in \mathcal{F}^k} \hat{c}^j \right\rangle + \mathrm{KL}(q \parallel q^k)$,
  where $\mathrm{KL}(q \parallel q') = \sum_{h,s,a} q_h(s,a) \ln \frac{q_h(s,a)}{q'_h(s,a)} + q'_h(s,a) - q_h(s,a)$.
  Update policy: $\pi_h^{k+1}(a \mid s) = {q_h^{k+1}(s,a)}/{q_h^{k+1}(s)}$.
end for

(2019); Bistritz et al. (2019) for MAB or Lancewicki et al. (2020) for MDPs) utilizes this fact to bound the regret with respect to the estimated cost by the sum of: (i) the regret of playing π^{k+d}; (ii) the "drift" between the playing π^{k+d} and π^k:

$$\sum_{k=1}^K \langle q^k - q^\star, \hat{c}^k \rangle \lesssim \underbrace{\sum_{k=1}^K \langle q^k - q^{k+d}, \hat{c}^k \rangle}_{\textsc{Drift}} + \frac{H}{\eta} + \underbrace{\eta \sum_{h,s,a,k} q_h^{k+d}(s,a) \hat{c}_h^k(s,a)^2}_{\textsc{Stability}} . \tag{5}$$

The term $\frac{H}{\eta}$ is usually referred to as the $\textsc{Penalty}$, and the bound (i) $\leq$ $\textsc{Penalty}$ + $\textsc{Stability}$ is by standard OMD guarantees. The *standard* importance sampling estimator defined in Equation (2) is approximately unbiased (ignoring γ and transition approximation errors), so the left-hand-side of Equation (5) is approximately the regret in expectation. On the other hand, to bound the $\textsc{Stability}$ term, one needs to control the ratio ${q_h^{k+d}(s,a)}/{q_h^k(s,a)}$ since $\hat{c}_h^k(s,a)$ has $q_h^k(s,a)$ in the denominator and not $q_h^{k+d}(s,a)$ (for simplicity we ignore the bias between q^k and u^k).

In MAB, this ratio is essentially bounded by a constant, but the proof heavily relies on the simple update form of OMD on the simplex (i.e., EXP3), as explained in Section 4. However, it still remains unclear whether this ratio is bounded by a constant when running OMD or FTRL on a more general convex set such as $\Delta(\mathcal{M})$. While in the proof of Theorems 1 and 2 we are able to avoid bounding the ratio in the stability term itself by using a "cheating" regret approach, a similar issue re-appears in the drift term. In Theorem 1 we bound the ratio between distributions by utilizing the simple update form (for the specific argument see Equation (20) in Appendix A), and in Theorem 2 we solve this issue with the help of convex conjugates (specifically, Hölder's inequality with respect to the Hessian of the regularizer at an intermediate occupancy measure ξ), but this comes at the cost of expected regret guarantees and looser bound on the delay term of the regret.

The main idea of the delay-adapted estimator is to re-weight the cost of episode k using both q^{k+d} and q^k. The first allows us to control the stability and avoids the need to bound the ratio ${q_h^{k+d}(s,a)}/{q_h^k(s,a)}$, while the second keeps the bias sufficiently small. More precisely, we re-weight using their maximum, which remarkably, causes the estimator's bias to scale similarly to the $\textsc{Drift}$ term.

Finally, there are a few important points to notice with respect to our new estimator before we analyze the regret of Algorithm 3 in Theorem 3. First, since the estimator $\hat{c}^k$ is computed only in the end of episode $k + d^k$ (when the feedback from episode k arrives), we have already computed both u^k and u^{k+d^k} at that point and the estimator is well-defined. Second, it generalizes the standard importance sampling estimator and adapts it to the delays. That is, whenever there is no delay, our estimator is identical to the standard importance sampling estimator. Third, there is no additional computational cost in computing the new estimator since we compute u^k for every k anyway. Moreover, there is no additional space complexity because every algorithm for adversarial environments with delayed feedback keeps the probabilities to play actions in episode k until its feedback is received in the end of episode $k + d^k$.

Theorem 3 *With appropriate choices of parameters, Delayed UOB-REPS with the delay-adapted estimator (Algorithm 3) ensures with high probability that* $R_K = \widetilde{O}\big(H^2 S \sqrt{AK} + (HSA)^{1/4} \cdot H\sqrt{D}\big)$.

The second term in the regret improves the guarantee of Delayed UOB-FTRL with the standard estimator by a factor of $H^{1/4}(SA)^{3/4}$. It also improves Delayed Hedge by $(HS)^{1/4}$, but on the other hand has an extra factor $A^{1/4}$. Generally, this term is tight up to the $(HSA)^{1/4}$ factor (Lancewicki et al., 2020). The first term in the regret matches the state-of-the-art regret bound for non-delayed adversarial MDPs (Jin et al., 2020a). In Appendix C we consider the case of known transitions, and present Delayed O-REPS with the delay-adapted estimator that achieves the following regret bound. It has similar delay term but its first term is optimal up to poly-log factors (Zimin and Neu, 2013).

Theorem 4 *Assume that the transition function is known to the learner. With high probability, Delayed O-REPS with the delay-adapted estimator (Algorithm 7) ensures that* $R_K = \widetilde{O}\big(H\sqrt{SAK} + (HSA)^{1/4} \cdot H\sqrt{D}\big)$.

We conclude the section with a proof sketch of our main theorem (for the unknown transition case).

Proof sketch of Theorem 3 We first break the regret as follows:

$$R_K = \underbrace{\sum_{k=1}^{K}\langle q^{\pi^k} - q^k, c^k\rangle}_{\text{EST}} + \underbrace{\sum_{k=1}^{K}\langle q^k, c^k - \hat{c}^k\rangle}_{\text{BIAS}_1} + \underbrace{\sum_{k=1}^{K}\langle q^\star, \hat{c}^k - c^k\rangle}_{\text{BIAS}_2} + \underbrace{\sum_{k=1}^{K}\langle q^k - q^{k+d^k}, \hat{c}^k\rangle}_{\text{DRIFT}} + \underbrace{\sum_{k=1}^{K}\langle q^{k+d^k} - q^\star, \hat{c}^k\rangle}_{\text{REG}}.$$

EST is the standard transition approximation error term which is bounded w.h.p by $\widetilde{O}(H^2 S\sqrt{AK})$ (Jin et al., 2020a). For BIAS_2 we use the fact that the delay-adapted estimator is always smaller than the standard estimator and bound it by $\widetilde{O}(H/\gamma)$ similarly to Jin et al. (2020a).

The real advantage of the estimator appears in the REG term. Similar to the fixed delay case, we can bound REG by,

$$\frac{H}{\eta} + \underbrace{\eta \sum_{k,h,s,a} q_h^{k+d^k}(s,a)\hat{c}_h^k(s,a)\left(\sum_{j\in\mathcal{F}^{k+d^k}} \hat{c}_h^j(s,a)\right)}_{\text{STABILITY}} \le \frac{H}{\eta} + \eta \sum_{k,h,s,a}\sum_{j\in\mathcal{F}^{k+d^k}} \hat{c}_h^j(s,a),$$

where the inequality above is exactly where we utilize the delay-adapted estimator, as by its definition $\hat{c}_h^k(s,a) \le 1/u_h^{k+d^k}(s,a) \le 1/q_h^{k+d^k}(s,a)$, where the last inequality holds w.h.p. Then, using a standard concentration of $\hat{c}_h^k(s,a)$ around $c_h^k(s,a) \le 1$ we get that STABILITY $\lesssim \eta(HSAK + d_{max}/\gamma)$. Importantly, the concentration arguments hold only because the maximum of u^k and u^{k+d^k} appears in the estimator's denominator. If it were only u^{k+d^k}, we could not have bounded the distance between the estimator $\hat{c}^k$ and the real cost c^k.

For the DRIFT term, let $\widetilde{H}^k$ be the realization of all episodes j such that $j + d^j < k$. Note that u^k and u^{k+d^k} are completely determined by the history $\widetilde{H}^{k+d^k}$, and on the other hand, the k-th episode is not part of this history. Next, we take the absolute value on each element of $q^k - q^{k+d^k}$ and apply a concentration bound to obtain: DRIFT $\lesssim \sum_{k=1}^K \mathbb{E}\big[\langle|q^k - q^{k+d^k}|, \hat{c}^k\rangle \mid \widetilde{H}^{k+d^k}\big] + \frac{H}{\gamma}$. The specific definition of the history $\widetilde{H}^{k+d^k}$ is crucial because now we have:

$$\begin{aligned}\text{DRIFT} &\lesssim \sum_{k=1}^{K}\mathbb{E}\left[\langle|q^k - q^{k+d^k}|, \hat{c}^k\rangle \mid \widetilde{H}^{k+d^k}\right] + \frac{H}{\gamma} = \sum_{k=1}^{K}\langle|q^k - q^{k+d^k}|, \mathbb{E}\left[\hat{c}^k \mid \widetilde{H}^{k+d^k}\right]\rangle + \frac{H}{\gamma} \\ &\le \sum_{k=1}^{K}\|q^k - q^{k+d^k}\|_1 + \frac{H}{\gamma} \le \sum_{k=1}^{K}\sum_{j=1}^{d^k}\|q^j - q^{j+1}\|_1 + \frac{H}{\gamma} \lesssim \sum_{k=1}^{K}\sum_{j=1}^{d^k}\sqrt{\text{KL}(q^j \parallel q^{j+1})} + \frac{H}{\gamma},\end{aligned}$$

where the third step follows since w.h.p $\mathbb{E}\big[\hat{c}_h^k(s,a) \mid \widetilde{H}^{k+d^k}\big] = \frac{q_h^{\pi^k}(s,a)c_h^k(s,a)}{\max\{u_h^k(s,a), u_h^{k+d^k}(s,a)\}} \le 1$, the forth step uses the triangle inequality, and the last is by Pinsker inequality. Finally, we utilize the OMD update (which uses KL as regularization) to obtain a bound on $\text{KL}(q^j \parallel q^{j+1})$ and finally a bound $\widetilde{O}(\eta\sqrt{H^3 SA}(D+K) + H/\gamma)$ on the DRIFT term. For BIAS_1, we apply a similar concentration on the cost estimators around $\mathbb{E}\big[\hat{c}^k \mid \widetilde{H}^{k+d^k}\big]$ and show that BIAS_1

is mainly bounded by,

$$\sum_k \|\max\{u^{k+d^k}, u^k\} - q^k\|_1 + \gamma HSAK \leq 2\sum_k \|u^k - q^k\|_1 + \sum_k \|q^{k+d^k} - q^k\|_1 + \gamma HSAK,$$

where the maximum is taken element-wise. For last, the first sum is bounded similarly to the EST term while the second sum is bounded similarly to the DRIFT term. Summing the regret from the different terms and optimizing over η and γ completes the proof. ■

6. Conclusions and Future Work

In this paper we made a substantial contribution to the literature on delayed feedback in RL. We presented the first algorithms that achieve near-optimal regret bounds for the challenging setting of adversarial MDP with delayed bandit feedback. Our key algorithmic contribution is a novel delay-adapted importance sampling estimator, and we develop various new techniques to analyze delayed bandit feedback in adversarial MDPs.

We leave a few interesting questions open for future work. First, there is still a gap of $(HSA)^{1/4}$ in the delay term between our upper bounds and the lower bound of Lancewicki et al. (2020). Second, it remains an open question whether our new estimator is necessary to obtain optimal regret in the presence of delays, or is it possible to achieve optimal regret with standard algorithms. Finally, our algorithms are based on the O-REPS framework but it remains an important open problem to achieve $\widetilde{O}\left(\sqrt{K+D}\right)$ regret with policy optimization (PO) methods that are widely used in practice, and were recently shown to achieve near-optimal regret in adversarial MDP with non-delayed bandit feedback (Luo et al., 2021).

Acknowledgments

Acknowledgements TL, YM and AR have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grant number 993/17), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University. HL is supported by NSF Award IIS-1943607 and a Google Faculty Research Award.

References

Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5451–5452. IEEE, 2012.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. Online exp3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, pages 11349–11358, 2019.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. No discounted-regret learning in adversarial bandits with delays. *arXiv preprint arXiv:2103.04550*, 2021.

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.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anonymous feedback. In *Conference On Learning Theory*, pages 750–773, 2018.

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, 2016.

Nesrine Changuel, Bessem Sayadi, and Michel Kieffer. Online learning for qoe-based video streaming to mobile receivers. In *2012 IEEE Globecom Workshops*, pages 1319–1324. IEEE, 2012.

Baiming Chen, Mengdi Xu, Zuxin Liu, Liang Li, and Ding Zhao. Delay-aware multi-agent reinforcement learning. *arXiv preprint arXiv:2005.05441*, 2020.

Alon Cohen, Amit Daniely, Yoel Drori, Tomer Koren, and Mariano Schain. Asynchronous stochastic optimization robust to arbitrary delays. *arXiv preprint arXiv:2106.11879*, 2021a.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34, 2021b.

Esther Derman, Gal Dalal, and Shie Mannor. Acting in delayed environments with non-stationary markov policies. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Manegueu Anne Gael, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pages 3348–3356. PMLR, 2020.

Andras Gyorgy and Pooria Joulani. Adapting to delays and data in adversarial multi-armed bandits. In *International Conference on Machine Learning*, pages 3988–3997. PMLR, 2021.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

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

Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Delay and cooperation in nonstochastic linear bandits. *Advances in Neural Information Processing Systems*, 33:4872–4883, 2020.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 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, 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020b.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in neural information processing systems*, 2020.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *Advances in Neural Information Processing Systems*, 2021.

Konstantinos V Katsikopoulos and Sascha E Engelbrecht. Markov decision processes with delays and asynchronous cost collection. *IEEE transactions on automatic control*, 48(4):568–574, 2003.

Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. Learning adversarial markov decision processes with delayed feedback. *arXiv preprint arXiv:2012.14843*, 2020.

Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 5969–5978. PMLR, 2021.

Shichao Liu, Xiaoyu Wang, and Peter Xiaoping Liu. Impact of communication delays on secondary frequency control in an islanded microgrid. *IEEE Transactions on Industrial Electronics*, 62(4):2021–2031, 2014.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34, 2021.

A Rupam Mahmood, Dmytro Korenkevych, Brent J Komer, and James Bergstra. Setting up a reinforcement learning task with a real-world robot. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4635–4640. IEEE, 2018.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28:3168–3176, 2015.

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.

Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. *Advances in neural information processing systems*, 28:1270–1278, 2015.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pages 5478–5486. PMLR, 2019a.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. *Advances in Neural Information Processing Systems*, 32:2212–2221, 2019b.

Aviv Rosenberg and Yishay Mansour. Stochastic shortest path with adversarially changing costs. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 2936–2942. ijcai.org, 2021. doi: 10.24963/ijcai.2021/404. URL https://doi.org/10.24963/ijcai.2021/404.

Erik Schuitema, Lucian Buşoniu, Robert Babuška, and Pieter Jonker. Control delay in reinforcement learning for real-time dynamic systems: a memoryless approach. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3226–3231. IEEE, 2010.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613. PMLR, 2020.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, pages 6541–6550, 2019.

Dirk van der Hoeven and Nicolò Cesa-Bianchi. Nonstochastic bandits and experts with arm-dependent delays. *arXiv preprint arXiv:2111.01589*, 2021.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckner. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pages 9712–9721. PMLR, 2020.

Thomas J Walsh, Ali Nouri, Lihong Li, and Michael L Littman. Learning and planning in environments with delayed feedback. *Autonomous Agents and Multi-Agent Systems*, 18(1):83, 2009.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964, 2020a.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020b.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*, pages 5197–5208, 2019.

Alexander Zimin. *Online Learning in Markovian Decision Processes*. PhD thesis, Central European University, 2013.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 1583–1591, 2013.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, pages 3285–3294. PMLR, 2020.

Appendix

Table of Contents

Algorithm 4 Delayed Hedge

Input: State space $\mathcal{S}$, Action space $\mathcal{A}$, Horizon H, Number of episodes K, Learning rate $\eta > 0$, Exploration parameter $\gamma > 0$, Confidence parameter $\delta > 0$.

Initialization: Set $\omega^1(\pi) = \frac{1}{|\Omega|}$ for every deterministic policy $\pi \in \Omega$; set $n_h^1(s,a) = 0, n_h^1(s,a,s')$ for every $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$ and $\mathcal{P}^1$ be the set of all transition functions.

for $k = 1, 2, ..., K$ **do**

Play a randomly sampled policy from distribution ω^k and observe trajectory $\{(s_h^k, a_h^k)\}_{h=1}^H$.

Compute upper occupancy bound $u_h^k(s,a) = \max_{p' \in \mathcal{P}^k} \sum_{\pi \in \Omega} \omega^k(\pi) q_h^{p',\pi}(s,a)$.

Define confidence set $\mathcal{P}^{k+1}$ by Algorithm 5.

for $j : j + d^j = k$ **do**

Observe feedback $\{c_h^j(s_h^j, a_h^j)\}_{h=1}^H$.

Compute loss estimator $\widehat{c}_h^j(s,a) = \frac{c_h^j(s,a)\mathbb{I}\{s_h^j = s, a_h^j = a\}}{u_h^j(s,a) + \gamma}$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

end for

Update probability distribution over policy space:

$$\omega^{k+1}(\pi) \propto \omega^k(\pi) \cdot \exp\left(\eta \cdot b^k(\pi) - \eta \sum_{j:j+d^j=k} \widehat{\ell}^j(\pi)\right), \forall \pi \in \mathcal{A}^{\mathcal{S}\times[H]}$$

where $\widehat{\ell}^j(\pi) = \sum_{h=1}^H \sum_{s,a} q_h^{\pi,\bar{p}^j}(s,a)\widehat{c}_h^j(s,a)$ denotes the loss suffered by policy π with respect to the loss estimator $\widehat{c}^j$ and transition function $\bar{p}^j$, $b^k(\pi) = \max_{p' \in \mathcal{P}^k} \left\| q^{\pi,\bar{p}^k} - q^{\pi,p'} \right\|_1$ is the exploration bonus for policy π at episode k.

end for

Algorithm 5 Update confidence set

Input: trajectory $\{(s_h^k, a_h^k)\}_{h=1}^H$.

Update visit counters: $n_h^{k+1}(s_h^k, a_h^k) \leftarrow n_h^k(s_h^k, a_h^k) + 1, n_h^{k+1}(s_h^k, a_h^k, s_{h+1}^k) \leftarrow n_h^k(s_h^k, a_h^k, s_{h+1}^k) + 1$ for every $h \in [H]$.

Compute empirical transitions function $\bar{p}^{k+1}$: $\bar{p}_h^{k+1}(s' \mid s,a) = \frac{n_h^{k+1}(s,a,s')}{n_h^{k+1}(s,a) \vee 1} \qquad \forall (s,a,s',h)$.

Define confidence sets $\mathcal{P}^{k+1}$ such that $p' \in \mathcal{P}^{k+1}$ if and only if, for every (s,a,s',h), p' ensures $\sum_{s'} p'_h(s'|s,a) = 1$ and:

$$\left|p'_h(s'|s,a) - \bar{p}_h^{k+1}(s'|s,a)\right| \leq \sqrt{\frac{16\bar{p}_h^{k+1}(s'|s,a)\log\frac{10HSAK}{\delta}}{n_h^{k+1}(s,a) \vee 1}} + \frac{10\log\frac{10HSAK}{\delta}}{n_h^{k+1}(s,a) \vee 1}.$$

Appendix A. Delayed Hedge

In this section, we consider running Hedge over the policy space, that is, the set of all deterministic policies. We propose Algorithm 4 with unknown transition and bandit feedback, which ensures $\widetilde{\mathcal{O}}\left(\sqrt{K} + \sqrt{D}\right)$ regret as shown in Theorem 5 (ignoring dependence on other parameters).

Theorem 5 *With $\eta = \gamma = \sqrt{\frac{S\iota}{HD+HSAK}}$, Algorithm 4 ensures that*

$$R_K = O\left(H^2 S\sqrt{AK\iota} + H^{3/2}\sqrt{SD\iota} + H^3 S^3 A\iota^3 + H^2 d_{max}\iota\right).$$

with probability at least $1 - 64\delta$ and the coefficient $\iota = \log\frac{HSAK}{\delta}$.

A.1 Proof of the Main Theorem

Proof of Theorem 5 We first decompose the regret decomposition as:

$$R_K = \sum_{k=1}^{K} \left\langle \omega^k - \omega^\star, \ell^k \right\rangle = \underbrace{\sum_{k=1}^{K} \left\langle \omega^k, \ell^k - \widehat{\ell}^k + b^k \right\rangle}_{\text{EST}} + \underbrace{\sum_{k=1}^{K} \left\langle \omega^k - \omega^\star, \widehat{\ell}^k - b^k \right\rangle}_{\text{REG}} + \underbrace{\sum_{k=1}^{K} \left\langle \omega^\star, \widehat{\ell}^k - b^k - \ell^k \right\rangle}_{\text{BIAS}}. \tag{6}$$

By combining Lemmas 6 to 8, we arrive at the following bound of regret with learning rate η, exploration parameter γ and confidence parameter δ, with probability at least $1 - 64\delta$ that

$$\begin{aligned} R_K = \mathcal{O}\left(\frac{HS\ln(A)}{\eta} + \eta H^2 \left(D + H^2 SAK\right) + \gamma HSAK + \left(\frac{\eta}{\gamma} H^2 \left(d_{max} + 1\right) + \frac{H}{\gamma} \right) \iota \right) \\ + \mathcal{O}\left(H^2 S \sqrt{AK\iota} + H^3 S^3 A \ln K \iota^2 \right). \end{aligned} \tag{7}$$

Setting the learning rate and exploration parameter $\eta = \gamma = \sqrt{\frac{S\ln(A)}{HD + HSAK}}$, one can verify that the regret R_K is bounded by $\mathcal{O}\left(H^2 S \sqrt{AK\iota} + H^{3/2}\sqrt{SD\iota} + H^3 S^3 A \iota^3 + H^2 d_{max} \iota \right)$. ■

Throughout the rest of this section, we will bound the three terms separately in Lemmas 6 to 8.

A.2 Bound on the Bias of the Cost Estimator (BIAS in Equation (6))

Lemma 6 (BIAS) *With probability at least* $1 - 7\delta$*, Algorithm 4 ensures that* $\text{BIAS} = \mathcal{O}\left(\frac{H\iota}{\gamma}\right)$.

Proof Similar to the analysis in Jin et al. (2020a), we have BIAS bounded by

$$\begin{aligned} \sum_{k=1}^{K} \left\langle \omega^\star, \widehat{\ell}^k - b^k - \ell^k \right\rangle &= \sum_{k=1}^{K} \left\langle q^{\pi^\star, \bar{p}^k}, \widehat{c}^k \right\rangle - \sum_{k=1}^{K} b^k(\pi^\star) - \sum_{k=1}^{K} \left\langle q^{\pi^\star, p}, c^k \right\rangle \\ &\le \mathcal{O}\left(\frac{H}{\gamma} \log\left(\frac{HSA}{\delta} \right) \right) + \sum_{k=1}^{K} \left\langle q^{\pi^\star, \bar{p}^k} - q^{\pi^\star, p}, c^k \right\rangle - b^k(\pi^\star) \\ &\le \mathcal{O}\left(\frac{H}{\gamma} \log\left(\frac{HSA}{\delta} \right) \right) + \sum_{k=1}^{K} \left\| q^{\pi^\star, \bar{p}^k} - q^{\pi^\star, p} \right\|_1 - b^k(\pi^\star) \\ &= \mathcal{O}\left(\frac{H}{\gamma} \log\left(\frac{HSA}{\delta} \right) \right), \end{aligned} \tag{8}$$

where the second step applies Lemma 14 of Jin et al. (2020a) with probability at least $1 - 6\delta$; the third step applies Hölder's inequality; the last step follows from the event $p \in \cap_k \mathcal{P}^k$ which holds with probability at least $1 - \delta$, and the definition of exploration bonus $b^k(\pi)$. ■

A.3 Bound on the Transition Estimation Error (EST in Equation (6))

Lemma 7 (EST) *With probability at least* $1 - 8\delta$*, Algorithm 4 ensures that*

$$\text{EST} = \mathcal{O}\left(\gamma HSAK + H^2 S \sqrt{AK \log \iota} + S^3 H^3 A \ln K \iota \right).$$

Proof Observe that, $\sum_{k=1}^{K}\left\langle \omega^k, \ell^k - \widehat{\ell}^k + b^k \right\rangle$ can be upper bounded under the event that $p \in \cap_k \mathcal{P}^k$ by

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)\left(\left\langle q^{\pi,p}, c^k\right\rangle - \left\langle q^{\pi,\bar{p}^k}, \widehat{c}^k\right\rangle\right) + \sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)b^k(\pi) \\
&= \sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)\left\langle q^{\pi,p} - q^{\pi,\bar{p}^k}, c^k\right\rangle + \sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)\left\langle q^{\pi,\bar{p}^k}, c^k - \widehat{c}^k\right\rangle + \sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)b^k(\pi) \\
&\le \sum_{k=1}^{K}\left\langle q^k, c^k - \widehat{c}^k\right\rangle + 2\sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)b^k(\pi) \qquad (9)
\end{aligned}$$

where $q^k = \sum_{\pi\in\Omega}\omega^k(\pi)q^{\pi,\bar{p}^k}$ is the estimated occupancy measure at episode k, and the second step follows from the definition of b^k and Hölder's inequality.

Note that, $\left\langle q^k, \widehat{c}^k\right\rangle$ is bounded by H because $\bar{p}^k \in \mathcal{P}^k$ and $u_h^k(s,a) \ge q_h^k(s,a)$ by its definition. Thus, with the help of Azuma's inequality, we have with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\left\langle q^k, \mathbb{E}^k\left[\widehat{c}^k\right] - \widehat{c}^k\right\rangle \le \mathcal{O}\left(H\sqrt{K\ln\left(\frac{1}{\delta}\right)}\right).$$

where $E^k[\cdot] = E[\cdot \mid \mathcal{H}^k]$ and $\mathcal{H}^k$ is the history of episodes 1, ..., $k-1$. We then focus on the term $\sum_{k=1}^{K}\left\langle q^k, c^k - \mathbb{E}^k\left[\widehat{c}^k\right]\right\rangle$ and rewrite it as

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h,s,a} q_h^k(s,a)c_h^k(s,a)\left(1 - \frac{\mathbb{E}^k\left[\mathbb{I}\{s_h^k = s, a_h^k = a\}\right]}{u_h^k(s,a)+\gamma}\right) \\
&= \sum_{k=1}^{K}\sum_{h,s,a} q_h^k(s,a)c_h^k(s,a)\left(1 - \frac{\widehat{q}_h^k(s,a)}{u_h^k(s,a)+\gamma}\right) \\
&= \sum_{k=1}^{K}\sum_{h,s,a} \frac{q_h^k(s,a)}{u_h^k(s,a)+\gamma}\left(u_h^k(s,a) - \widehat{q}_h^k(s,a) + \gamma\right)c_h^k(s,a) \\
&\le \gamma HSAK + \sum_{k=1}^{K}\sum_{h,s,a}\left|u_h^k(s,a) - \widehat{q}_h^k(s,a)\right| \qquad (10)
\end{aligned}$$

where $\widehat{q}^k = \sum_{\pi\in\Omega}\omega^k(\pi)q^{\pi,p}$ is the occupancy measure with the true transition p, and the last step comes from the fact that $u_h^k(s,a) \ge q_h^k(s,a)$ for all state-action pairs according to its definition.

Fixed the state-action pair (s,a) and let $p' \in \mathcal{P}^k$ be the transition function that yields $u_h^k(s,a)$ for simplicity. Then, we have the following inequality under the event $p \in \bigcap_k \mathcal{P}^k$ that

$$\begin{aligned}
u_h^k(s,a) - \widehat{q}_h^k(s,a) &= \sum_{\pi\in\Omega}\omega^k(\pi)\left(q_h^{\pi,p'}(s,a) - q_h^{\pi,p}(s,a)\right) \\
&= \sum_{\pi\in\Omega}\omega^k(\pi)\sum_{m=0}^{h-1}\sum_{x,y,z} q_m^{\pi,p}(x,y)\cdot\left(p_m(z|x,y) - p'_m(z|x,y)\right)\cdot q_{h|m+1}^{\pi,p'}(s,a|z) \\
\Rightarrow \left|u_h^k(s,a) - \widehat{q}_h^k(s,a)\right| &\le \sum_{\pi\in\Omega}\omega^k(\pi)\sum_{m=0}^{h-1}\sum_{s,a,s'} q_m^{\pi,p}(x,y)\cdot\epsilon_m^k(z|x,y)\cdot q_{h|m+1}^{\pi,p'}(s,a|z)
\end{aligned}$$

where the second step follows from (Jin et al., 2021, Lemma D.3.1) with the conditional occupancy measure $q_{h|m+1}^{\pi,p'}(s,a|z)$ being the conditional probability of visiting state-action pair (s,a) at step h from state z at state $m+1$ with policy π

and transition p'; the third step comes from taking the absolute value of both sides and the fact that

$$\epsilon_h^k(s'|s,a) \triangleq \mathcal{O}\left(\min\left\{1, \sqrt{\frac{p_h^k(s'|s,a)\iota}{n_h^k(s,a)\vee 1}} + \frac{\iota}{n_h^k(s,a)\vee 1}\right\}\right) \geq |p_h'(s'|s,a) - p_h(s'|s,a)| \tag{11}$$

for any transition tuple (s,a,s') and step h under the event $p \in \bigcap_k \mathcal{P}^k$ according to (Jin et al., 2021, Lemma D.3.3). In addition, we have $q_{h|m+1}^{\pi,p'}(s,a|z) - q_{h|m+1}^{\pi,p}(s,a|z)$ bounded by

$$\begin{aligned}
&\sum_{o=m+1}^{h-1}\sum_{u,v,w} q_{o|m+1}^{\pi,p}(u,v|z)\cdot\left(p_o^k(w|u,v) - p_o'(w|u,v)\right)\cdot q_{h|o+1}^{\pi,p'}(s,a|w) \\
&\leq \pi_h(a|s)\sum_{o=m+1}^{h-1}\sum_{u,v,w} q_{o|m+1}^{\pi,p}(u,v|z)\cdot\left|p_o^k(w|u,v) - p_o'(w|u,v)\right| \\
&\leq \pi_h(a|s)\sum_{o=m+1}^{h-1}\sum_{u,v} q_{o|m+1}^{\pi,p}(u,v|z)\cdot\min\left\{2, \sum_w \epsilon_o^k(w|u,v)\right\}
\end{aligned}$$

where the first step uses the fact that $q_{h|o+1}^{\pi,p'}(s,a|w) \leq \pi_h(a|s)\cdot q_{h|o+1}^{\pi,p'}(s|w) = \pi_h(a|s)$; the second step follows from similar argument above; the last step uses the fact that $\sum_w \left|p_o^k(w|u,v) - p_o'(w|u,v)\right| \leq 2$.

Combining these inequalities, we have the second term of Equation (10), $\sum_{k=1}^k \sum_{h,s,a}\left|u_h^k(s,a) - \widehat{q}_h^k(s,a)\right|$, bounded by

$$\begin{aligned}
&\sum_{k=1}^K\sum_{\pi\in\Omega}\omega^k(\pi)\sum_{h,s,a}\sum_{m=0}^{h-1}\sum_{x,y,z} q_m^{\pi,p}(x,y)\cdot\epsilon_m^k(z|x,y)\cdot q_{h|m+1}^{\pi,p}(s,a|z) \\
&+\sum_{k=1}^K\sum_{\pi\in\Omega}\omega^k(\pi)\sum_{h,s,a}\sum_{m=0}^{h-1}\sum_{x,y,z}\sum_{o=m+1}^{h-1}\sum_{u,v} q_m^{\pi,p}(x,y)\cdot\epsilon_m^k(z|x,y)\cdot q_{o|m+1}^{\pi,p}(u,v|z)\cdot\min\left\{2,\sum_w \epsilon_o^k(w|u,v)\right\}\cdot\pi_h(a|s).
\end{aligned} \tag{12}$$

Note that, the first term of Equation (12) can be bounded (under the event $p \in \bigcap_k \mathcal{P}^k$) as

$$\begin{aligned}
&\sum_{k=1}^K\sum_{\pi\in\Omega}\omega^k(\pi)\sum_{h,s,a}\sum_{m=0}^{h-1}\sum_{x,y,z} q_m^{\pi,p}(x,y)\cdot\epsilon_m^k(z|x,y)\cdot q_{h|m+1}^{\pi,p}(s,a|z) \\
&= \sum_{k=1}^K\sum_{\pi\in\Omega}\omega^k(\pi)\sum_{m=0}^{H}\sum_{x,y,z} q_m^{\pi,p}(x,y)\cdot\epsilon_m^k(z|x,y)\cdot\left(\sum_{h=m+1}^{H}\sum_{s,a} q_{h|m+1}^{\pi,p}(s,a|z)\right) \\
&\leq H\sum_{k=1}^K s\sum_{\pi\in\Omega}\omega^k(\pi)\sum_{m=0}^{H}\sum_{x,y,z} q_m^{\pi,p}(x,y)\cdot\epsilon_m^k(z|x,y) \\
&= H\sum_{k=1}^K\sum_{m=0}^{H}\sum_{x,y}\left(\sum_{\pi\in\Omega}\omega^k(\pi)q_m^{\pi,p}(x,y)\right)\cdot\left(\sum_z \epsilon_m^k(z|x,y)\right) \\
&= H\sum_{k=1}^K\sum_{m=0}^{H}\sum_{x,y}\widehat{q}_m^k(x,y)\cdot\left(\sum_z \epsilon_m^k(z|x,y)\right) \\
&= \mathcal{O}\left(H\sum_{k=1}^K\sum_{m=0}^{H}\sum_{x,y}\widehat{q}_m^{\pi,p}(x,y)\left(\sqrt{\frac{S\iota}{n_m^k(x,y)\vee 1}} + \frac{S\iota}{n_m^k(x,y)\vee 1}\right)\right) \\
&= \mathcal{O}\left(H^2 S\sqrt{AK\log\iota}\right)
\end{aligned} \tag{13}$$

where the second steps follows from the fact that $\sum_{s,a} q^{\pi,p}_{h|m+1}(s,a|z) = 1$ for any policy π and step $h \geq m+1$; the fourth step uses the definition of $\widehat{q}^k$, the true occupancy measure at episode k; the fifth step uses the properties of ϵ^k under the event $p \in \bigcap_k \mathcal{P}^k$; the final step applies Lemma 10 of Jin et al. (2020a), which yields a high probability bound with the help of a standard Bernstein-type concentration inequality for martingale.

Observing that $\sum_{h=o+1}^{H} \sum_{s,a} \pi_h(a|s) \leq SH$, we can reorder the summation and bound the second term of Equation (12) by SH multiplying

$$\sum_{k=1}^{k} \sum_{\pi\in\Omega} \omega^k(\pi) \sum_{m=0}^{h-1} \sum_{x,y,z} \sum_{o=m+1}^{h-1} \sum_{u,v} q^{\pi,p}_m(x,y) \cdot \epsilon^k_m(z|x,y) \cdot q^{\pi,p}_{o|m+1}(u,v|z) \cdot \min\left\{2, \sum_w \epsilon^k_o(w|u,v)\right\}.$$

Similar to the proof in Appendix B.2 of Jin et al. (2020a), we can further rewrite and bound the term above by

$$\begin{aligned}
&\mathcal{O}\left(\sum_{k=1}^{K} \sum_{\pi\in\Omega} \omega^k(\pi) \sum_{m=0}^{H-1} \sum_{x,y,z} \sum_{o=m+1}^{H} \sum_{u,v,w} q^{\pi,p}_m(x,y) \cdot \sqrt{\frac{p_m(z|x,y)\iota}{n^k_m(x,y)\vee 1}} \cdot q^{\pi,p}_{o|m+1}(u,v|z) \cdot \sqrt{\frac{p_o(w|u,v)\iota}{n^k_o(u,v)\vee 1}}\right) \\
&+ \mathcal{O}\left(\sum_{k=1}^{K} \sum_{\pi\in\Omega} \omega^k(\pi) \sum_{m=0}^{H-1} \sum_{x,y,z} \frac{q^{\pi,p}_m(x,y)\iota}{n^k_m(x,y)\vee 1} \left(\sum_{o=m+1}^{H} \sum_{u,v} q^{\pi,p}_{o|m+1}(u,v|z) \min\left\{\sum_w \epsilon^k_o(w|u,v), 2\right\}\right)\right) \\
&+ \mathcal{O}\left(\sum_{k=1}^{K} \sum_{\pi\in\Omega} \omega^k(\pi) \sum_{o=0}^{H} \sum_{u,v,w} \left(\sum_{m=0}^{o-1} \sum_{x,y,z} q^{\pi,p}_m(x,y) \cdot p_m(z|x,y) \cdot q^{\pi,p}_{o|m+1}(u,v|z)\right) \cdot \frac{\iota}{n^k_o(u,v)\vee 1}\right)
\end{aligned}$$

by using the property of ϵ^k_h as in Equation (11) and the fact that $\sqrt{xy} \leq x+y$ for any $x, y > 0$, therefore, $\epsilon^k_h(s'|s,a) \leq \mathcal{O}\left(p_h(s'|s,a) + \frac{\iota}{n^k_h(s,a)\vee 1}\right)$ holds for any (s,a,s').

Clearly, the later two are able to be reformulated and then bounded as

$$\begin{aligned}
&\mathcal{O}\left(\sum_{k=1}^{K} \sum_{\pi\in\Omega} \omega^k(\pi) \sum_{m=0}^{H-1} \sum_{x,y,z} \frac{q^{\pi,p}_m(x,y)\iota}{n^k_m(x,y)\vee 1} \left(\sum_{o=m+1}^{H} \sum_{u,v} q^{\pi,p}_{o|m+1}(u,v|z) \min\left\{\sum_w \epsilon^k_o(w|u,v), 2\right\}\right)\right) \\
&\quad + \mathcal{O}\left(\sum_{k=1}^{K} \sum_{\pi\in\Omega} \omega^k(\pi) \sum_{o=0}^{H} \sum_{u,v,w} \left(\sum_{m=0}^{o-1} \sum_{x,y,z} q^{\pi,p}_m(x,y) \cdot p_m(z|x,y) \cdot q^{\pi,p}_{o|m+1}(u,v|z)\right) \cdot \frac{\iota}{n^k_o(u,v)\vee 1}\right) \\
&= \mathcal{O}\left(H \sum_{k=1}^{K} \sum_{\pi\in\Omega} \omega^k(\pi) \sum_{m=0}^{H-1} \sum_{x,y,z} \frac{q^{\pi,p}_m(x,y)\iota}{n^k_m(x,y)\vee 1} + H \sum_{k=1}^{K} \sum_{\pi\in\Omega} \omega^k(\pi) \sum_{o=0}^{H} \sum_{u,v,w} \frac{q^{\pi,p}_o(u,v)\iota}{n^k_o(u,v)\vee 1}\right) \\
&= \mathcal{O}\left(SH\iota^2 \sum_{k=1}^{K} \sum_{m=0}^{H-1} \sum_{x,y} \frac{\widehat{q}^k_m(x,y)}{n^k_m(x,y)\vee 1} + SH\iota^2 \sum_{k=1}^{K} \sum_{o=0}^{H} \sum_{u,v} \frac{\widehat{q}^k_o(u,v)}{n^k_o(u,v)\vee 1}\right) \\
&= \mathcal{O}\left(S^2 HA \ln K\iota^2\right)
\end{aligned} \tag{14}$$

where the first step comes from the facts that $\sum_{o=m+1}^{H} \sum_{u,v} q^{\pi,p}_{o|m+1}(u,v|z) \leq H$ for any z, and $\sum_{x,y,z} q^{\pi,p}_m(x,y) \cdot p_m(z|x,y) \cdot q^{\pi,p}_{o|m+1}(u,v|z) = q^{\pi,p}(u,v)$ for any (u,v) according to the definitions of conditional occupancy measures; the second step follows from the definition of $\widehat{q}^k$; the last step applies Lemma 10 of Jin et al. (2020a) with probability at least $1-2\delta$.

On the other hand, the first term can be written as $SH\iota$ multiplied by the following (ignoring some constants):

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)\sum_{m=0}^{H-1}\sum_{x,y,z}\sum_{o=m+1}^{H}\sum_{u,v,w}q_m^{\pi,p}(x,y)\cdot\sqrt{\frac{p_m(z|x,y)}{n_m^k(x,y)\vee 1}}\cdot q_{o|m+1}^{\pi,p}(u,v|z)\cdot\sqrt{\frac{p_o(w|u,v)}{n_o^k(u,v)\vee 1}}\\
&=\sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)\sum_{m=0}^{H-1}\sum_{x,y,z}\sum_{o=m+1}^{H}\sum_{u,v,w}\sqrt{\frac{q_m^{\pi,p}(x,y)p_m(z|x,y)q_{o|m+1}^{\pi,p}(u,v|z)}{n_m^k(x,y)\vee 1}}\cdot\sqrt{\frac{q_m^{\pi,p}(x,y)p_o(w|u,v)q_{o|m+1}^{\pi,p}(u,v|z)}{n_o^k(u,v)\vee 1}}\\
&=\sum_{k=1}^{K}\sum_{\pi\in\Omega}\sum_{m=0}^{H-1}\sum_{x,y,z}\sum_{o=m+1}^{H}\sum_{u,v,w}\sqrt{\frac{\omega^k(\pi)q_m^{\pi,p}(x,y)p_m(z|x,y)q_{o|m+1}^{\pi,p}(u,v|z)}{n_o^k(u,v)\vee 1}}\cdot\sqrt{\frac{\omega^k(\pi)q_m^{\pi,p}(x,y)p_o(w|u,v)q_{o|m+1}^{\pi,p}(u,v|z)}{n_m^k(x,y)\vee 1}}\\
&\le\sum_{m=0}^{H-1}\sum_{o=m+1}^{H}\sqrt{\sum_{k=1}^{K}\sum_{\pi\in\Omega}\sum_{x,y,z}\sum_{u,v,w}\frac{\omega^k(\pi)q_m^{\pi,p}(x,y)p_m(z|x,y)q_{o|m+1}^{\pi,p}(u,v|z)}{n_o^k(u,v)\vee 1}}\\
&\quad\cdot\sqrt{\sum_{k=1}^{K}\sum_{\pi\in\Omega}\sum_{x,y,z}\sum_{u,v,w}\frac{\omega^k(\pi)q_m^{\pi,p}(x,y)p_o(w|u,v)q_{o|m+1}^{\pi,p}(u,v|z)}{n_m^k(x,y)\vee 1}}\\
&\le\sum_{m=0}^{H-1}\sum_{o=m+1}^{H}\sqrt{\sum_{k=1}^{K}\sum_{\pi\in\Omega}\sum_{u,v,w}\frac{\omega^k(\pi)q_o^{\pi,p}(u,v)}{n_o^k(u,v)\vee 1}}\cdot\sqrt{\sum_{k=1}^{K}\sum_{\pi\in\Omega}\sum_{x,y,z}\frac{\omega^k(\pi)q_m^{\pi,p}(x,y)}{n_m^k(x,y)\vee 1}}\\
&=S\sum_{m=0}^{H-1}\sum_{o=m+1}^{H}\sqrt{\sum_{k=1}^{K}\sum_{u,v}\frac{\widehat{q}_o^k(u,v)}{n_o^k(u,v)\vee 1}}\cdot\sqrt{\sum_{k=1}^{K}\sum_{x,y,z}\frac{\widehat{q}_m(x,y)}{n_m^k(x,y)\vee 1}}\\
&=\mathcal{O}\left(S^2H^2A\ln K\right)
\end{aligned}\tag{15}$$

where the third step uses Cauchy-Schwarz inequality; the forth step follows from the properties of conditional occupancy measure $\sum_{x,y,z}q_m^{\pi,p}(x,y)p_m(z|x,y)q_{o|m+1}^{\pi,p}(u,v|z)=q_m^{\pi,p}(u,v)$; the last step applies Lemma 10 of Jin et al. (2020a) with probability at least $1-2\delta$.

Combining Equations (12) to (15) into Equation (10), we have the following inequality holds with probability at least $1-4\delta$ under the event $p\in\bigcap_k\mathcal{P}^k$ that

$$\sum_{k=1}^{K}\left\langle q^k,c^k-\widehat{c}^k\right\rangle=\mathcal{O}\left(\gamma HSAK+H^2S\sqrt{AK\log\iota}+S^3H^3A\ln K\iota\right).\tag{16}$$

With slightly abuse of notations, we use $\bar{p}^k(\pi)$ to denote the transition function that yields $b^k(\pi)$ associated with π and confidence set $\mathcal{P}^k$, that is, $\bar{p}^k(\pi)=\arg\max_{p'\in\mathcal{P}^k}\left\|q^{\pi,p'}-q^{\pi,\bar{p}^k}\right\|_1$. Thus, for $\sum_{k=1}^{K}\left\langle\omega^k,b^k\right\rangle$, we have the following inequality holds with probability at least $1-2\delta$ that

$$
\begin{aligned}
\sum_{k=1}^{K}\left\langle \omega^k, b^k\right\rangle &= \sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)\left\|q^{\pi,\bar{p}^k(\pi)} - q^{\pi,\bar{p}^k}\right\|_1 \\
&\le \sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)\left(\left\|q^{\pi,\bar{p}^k(\pi)} - q^{\pi,p}\right\|_1 + \left\|q^{\pi,p} - q^{\pi,\bar{p}^k}\right\|_1\right) \\
&\le H\sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)\sum_{h=1}^{H} q_h^{\pi,p}(s,a)\cdot\left(\left\|\bar{p}_h^k(\cdot|s,a) - p_h(\cdot|s,a)\right\|_1 + \left\|\bar{p}_h^k(\pi)(\cdot|s,a) - p_h(\cdot|s,a)\right\|_1\right) \\
&\le H\sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)\sum_{h=1}^{H} q_h^{\pi,p}(s,a)\cdot\left(\sum_{s'}\epsilon_h^k(s'|s,a)\right) \\
&\le \mathcal{O}\left(H\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{q}_h^k(s,a)\sqrt{\frac{S\iota}{n_h^k(s,a)\vee 1}}\right) \\
&\le \mathcal{O}\left(H^2 S\sqrt{AK\iota}\right)
\end{aligned} \tag{17}
$$

where the second step follows from the triangle inequality for ℓ_1 norms; the third step comes from Lemma B.1 and B.2 of Rosenberg and Mansour (2019a); the forth step uses the property of ϵ^k defined in Equation (11); the fifth step follows from the fact that $\sum_{\pi\in\Omega}\omega^k(\pi)q^{\pi,p} = \widehat{q}^k$; the final step follows from the same argument as in Equation (13).

Combining Equations (16) and (17) into Equation (9) concludes the proof. ■

A.4 Bound on the Regret with respect to the Loss Estimators (Reg in Equation (6))

Lemma 8 (Reg) *With probability at least $1-32\delta$, Algorithm 4 ensures that*

$$
\text{Reg} = \mathcal{O}\left(\frac{HS\ln(A)}{\eta} + \eta H^2\left(SAK + D\right) + \frac{\eta}{\gamma}\cdot H^2\left(d_{max}+1\right)\iota\right).
$$

Proof Let $\{\widetilde{\omega}^{k+1}\}_{k=1}^K$ be the sequence of probability distributions with both received and un-received loss estimators prior to episode $k+1$, that is,

$$
\widetilde{\omega}^{k+1}(\pi) \propto \omega^1(\pi)\cdot\exp\left(-\eta\left(\sum_{j=1}^{k}\widehat{\ell}^j(\pi) - \sum_{j=1}^{k} b^j(\pi)\right)\right), \forall\pi\in\mathcal{A}^{\mathcal{S}\times[H]}.
$$

On the other hand, according to the fact that $b^j(\pi')\le 2H$, we add a constant $2H$ uniformly to the loss vector $\widehat{\ell}^k - b^k$ and construct $m^k(\pi) = \widehat{\ell}^k(\pi) - b^k(\pi) + 2H$ to ensure the positiveness for any π . Clearly, adding the constant uniformly will not change the outcomes of our algorithm.

With the help of these notations, we are able to decompose Reg into two parts as:

$$
\text{Reg} = \underbrace{\sum_{k=1}^{K}\left\langle\widetilde{\omega}^{k+1} - \omega^\star, \widehat{\ell}^k - b^k\right\rangle}_{\text{Cheating Regret}} + \underbrace{\sum_{k=1}^{K}\left\langle\omega^k - \widetilde{\omega}^{k+1}, \widehat{\ell}^k - b^k\right\rangle}_{\text{Drift}}
$$

where Cheating-Regret is bounded in Gyorgy and Joulani (2021) that

$$
\sum_{k=1}^{K}\left\langle\widetilde{\omega}^{k+1} - \omega^\star, m^k\right\rangle \le \frac{\ln|\Omega|}{\eta} = \frac{\ln\left|\mathcal{A}^{\mathcal{S}\times[H]}\right|}{\eta} = \frac{HS\ln(A)}{\eta}. \tag{18}
$$

For DRIFT, we first rewrite it as

$$
\begin{aligned}
\sum_{k=1}^{K}\left\langle \omega^k-\widetilde{\omega}^{k+1},\widehat{\ell}^k-b^k\right\rangle &= \sum_{k=1}^{K}\left\langle \omega^k-\widetilde{\omega}^{k+1},2H+\widehat{\ell}^k-b^k\right\rangle \\
&= \sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)\left(2H+\widehat{\ell}^k(\pi)-b^k(\pi)\right)\cdot\left(1-\frac{\widetilde{\omega}^{k+1}(\pi)}{\omega^k(\pi)}\right) \\
&= \sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)m^k(\pi)\cdot\left(1-\frac{\widetilde{\omega}^{k+1}(\pi)}{\omega^k(\pi)}\right)
\end{aligned} \tag{19}
$$

where the second step follows from the fact that $\sum_{\pi\in\Omega}\widetilde{\omega}^{k+1}(\pi)=\sum_{\pi\in\Omega}\omega^k(\pi)=1$. Then, we consider the ratio between $\omega^k(\pi)$ and $\widetilde{\omega}^{k+1}(\pi)$:

$$
\begin{aligned}
\frac{\widetilde{\omega}^{k+1}(\pi)}{\omega^k(\pi)} &= \frac{\exp\left(-\eta\sum_{j=1}^{k}\left(\widehat{\ell}^j(\pi)-b^j(\pi)\right)\right)}{\sum_{\pi'\in\Omega}\exp\left(-\eta\sum_{j=1}^{k}\left(\widehat{\ell}^j(\pi')-b^j(\pi')\right)\right)}\cdot\frac{\sum_{\pi'\in\Omega}\exp\left(-\eta\sum_{j:j+d^j<k}\widehat{\ell}^j(\pi')+\eta\sum_{j=1}^{k-1}b^j(\pi')\right)}{\exp\left(-\eta\sum_{j:j+d^j<k}\widehat{\ell}^j(\pi)+\eta\sum_{j=1}^{k-1}b^j(\pi)\right)} \\
&= \frac{\exp\left(-\eta\sum_{j=1}^{k}\left(\widehat{\ell}^j(\pi)-b^j(\pi)\right)-\eta 2H\right)}{\sum_{\pi'\in\Omega}\exp\left(-\eta\sum_{j=1}^{k}\left(\widehat{\ell}^j(\pi')-b^j(\pi')\right)-\eta 2H\right)}\cdot\frac{\sum_{\pi'\in\Omega}\exp\left(-\eta\sum_{j:j+d^j<k}\widehat{\ell}^j(\pi')+\eta\sum_{j=1}^{k-1}b^j(\pi')\right)}{\exp\left(-\eta\sum_{j:j+d^j<k}\widehat{\ell}^j(\pi)+\eta\sum_{j=1}^{k-1}b^j(\pi)\right)} \\
&= \frac{\sum_{\pi'\in\Omega}\exp\left(-\eta\sum_{j:j+d^j<k}\widehat{\ell}^j(\pi')+\eta\sum_{j=1}^{k-1}b^j(\pi')\right)}{\sum_{\pi'\in\Omega}\exp\left(-\eta\sum_{j=1}^{k}\left(\widehat{\ell}^j(\pi')-b^j(\pi')\right)-\eta 2H\right)}\cdot\frac{\exp\left(-\eta\sum_{j=1}^{k}\left(\widehat{\ell}^j(\pi)-b^j(\pi)\right)-\eta 2H\right)}{\exp\left(-\eta\sum_{j:j+d^j<k}\widehat{\ell}^j(\pi)+\eta\sum_{j=1}^{k-1}b^j(\pi)\right)}
\end{aligned} \tag{20}
$$

where the second step follows from multiplying denominator and nominator together by $\exp(-\eta 2H)$. Note that $b^k(\pi)\le 2H$ and $\widehat{\ell}^k(\pi)\ge 0$ for any π and k, we thus have the following inequality holds that

$$
\sum_{j=1}^{k}\left(\widehat{\ell}^j(\pi')-b^j(\pi')\right)+2H=\sum_{j=1}^{k}\widehat{\ell}^j(\pi')-\sum_{j=1}^{k-1}b^j(\pi')+2H-b^k(\pi')\ge\sum_{j=1,j+d^j<k}^{k-1}\widehat{\ell}^j(\pi')-\sum_{j=1}^{k-1}b^j(\pi')
$$

which indicates that the first fraction is lower bounded by 1.

Therefore, the ratio $\widetilde{\omega}^{k+1}(\pi)/\omega^k(\pi)$ for any policy $\pi\in\Omega$ can be further bounded by

$$
\begin{aligned}
\widetilde{\omega}^{k+1}(\pi)/\omega^k(\pi) &\ge \exp\left(-\eta\left(\widehat{\ell}^k(\pi)+\sum_{j=1,j+d^j\ge k}^{k-1}\widehat{\ell}^k(\pi)+\left(2H-b^k(\pi)\right)\right)\right) \\
&\ge 1-\eta\left(m^k(\pi)+\sum_{j=1,j+d^j\ge k}^{k-1}\widehat{\ell}^k(\pi)\right),
\end{aligned}
$$

where the last step uses $1+x\le e^x$ for any $x\in\mathbb{R}$.

Plugging this inequality back to Equation (19), we have DRIFT bounded and then decomposed into two parts as

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)m^k(\pi)\left(1-\frac{\widetilde{\omega}^{k+1}(\pi)}{\omega^k(\pi)}\right) &\le \eta\sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)m^k(\pi)\left(m^k(\pi)+\sum_{j=1,j+d^j\ge k}^{k-1}\widehat{\ell}^k(\pi)\right) \\
&= \eta\sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)m^k(\pi)^2+\eta\sum_{k=1}^{K}\sum_{\pi\in\Omega}\omega^k(\pi)m^k(\pi)\sum_{j=1,j+d^j\ge k}^{k-1}\widehat{\ell}^j(\pi)
\end{aligned} \tag{21}
$$

where the first part associates with the regret incurred without the delayed feedback and can be controlled by standard arguments as:

$$\begin{aligned}\eta \sum_{k=1}^{K} \sum_{\pi\in\Omega} \omega^k(\pi) m^k(\pi)^2 &= \eta \sum_{k=1}^{K} \sum_{\pi\in\Omega} \omega^k(\pi) \left(\sum_{h=1}^{H} \sum_{s,a} q_h^{\pi,\bar{p}^k}(s,a) \widehat{c}_h^k(s,a) + 2H - b^k(\pi) \right)^2 \\ &\le 2\eta \sum_{k=1}^{K} \sum_{\pi\in\Omega} \omega^k(\pi) \left[\left(\sum_{h=1}^{H} \sum_{s,a} q_h^{\pi,\bar{p}^k}(s,a) \widehat{c}_h^k(s,a) \right)^2 + 4H^2 \right] \\ &\le 8\eta H^2 K + 2\eta H \sum_{k=1}^{K} \sum_{\pi\in\Omega} \omega^k(\pi) \sum_{h=1}^{H} \left(\sum_{s,a} q_h^{\pi,\bar{p}^k}(s,a) \cdot \widehat{c}_h^k(s,a) \right)^2 \\ &= 8\eta H^2 K + 2\eta H \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s,a} \sum_{\pi\in\Omega} \omega^k(\pi) q_h^{\pi,\bar{p}^k}(s,a)^2 \cdot \widehat{c}_h^k(s,a)^2 \\ &\le 8\eta H^2 K + 2\eta H \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s,a} \widehat{c}_h^k(s,a)^2 \left(\sum_{\pi\in\Omega} \omega^k(\pi) q_h^{\pi,\bar{p}^k}(s,a)^2 \right) \\ &\le 8\eta H^2 K + 2\eta H \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s,a} \widehat{c}_h^k(s,a)\end{aligned}$$

where the second step follows from the fact that $(x+y)^2 \le x^2 + y^2$; the third step uses Cauchy-Schwartz inequality; the forth step follows from the fact $\mathbb{I}\left\{s_h^k = s, a_h^k = a\right\} \mathbb{I}\left\{s_h^k = s', a_h^k = a'\right\} = 0$ for all $(s,a), (s',a') \in \mathcal{S} \times \mathcal{A}$ such that $(s,a) \neq (s',a')$; the final step uses the fact that $u_h^k(s,a) \ge \sum_{\pi\in\Omega} \omega^k(\pi) q_h^{\pi,\bar{p}^k}(s,a)$ and the definition of loss estimator $\widehat{c}^k$.

Moreover, with Lemma 11 of Jin et al. (2020a), we can show that the following inequality hold with probability at least $1 - 9\delta$ that

$$8\eta H^2 K + 2\eta H \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s,a} \widehat{c}_h^k(s,a) = \mathcal{O}\left(\eta H^2 SAK + \frac{\eta H^2}{\gamma} \iota \right). \tag{22}$$

Similarly, for some part of the second term of Equation (21), we have

$$\begin{aligned}\eta \sum_{k=1}^{K} \sum_{\pi\in\Omega} \omega^k(\pi) \left(2H - b^k(\pi)\right) \sum_{j=1, j+d^j \ge k}^{k-1} \widehat{c}^k(\pi) &\le 2\eta H \sum_{k=1}^{K} \sum_{j=1, j+d^j \ge k}^{k-1} \sum_{\pi\in\Omega} \omega^k(\pi) \left(\sum_{h=1}^{H} \sum_{s,a} q_h^{\pi,\bar{p}^j}(s,a) \widehat{c}_h^j(s,a) \right) \\ &= 2\eta H \sum_{k=1}^{K} \sum_{j=1, j+d^j \ge k}^{k-1} \sum_{\pi\in\Omega} \sum_{h=1}^{H} \sum_{s,a} \omega^k(\pi) q_h^{\pi,\bar{p}^j}(s,a) \widehat{c}_h^j(s,a) \\ &\le \mathcal{O}\left(\frac{\eta}{\gamma} H^2 d_{max} \iota \right) + 2\eta H \sum_{k=1}^{K} \sum_{j=1, j+d^j \ge k}^{k-1} \sum_{\pi\in\Omega} \sum_{h=1}^{H} \sum_{s,a} \omega^k(\pi) q_h^{\pi,\bar{p}^j}(s,a) \\ &= \mathcal{O}\left(\frac{\eta}{\gamma} H^2 d_{max} \iota + \eta H^2 D \right)\end{aligned} \tag{23}$$

where the third step uses Lemma 11 of Jin et al. (2020a) under the event that $p \in \cap_k \mathcal{P}^k$, which holds with probability at least $1 - 9\delta$.

On the other hand, the rest of the second part can be be bounded with respect to the conditional independence between loss estimators $\widehat{c}^k$ and $\widehat{c}_j$ for any $j < k$ satisfying $j + d^j \geq k$:

$$
\begin{aligned}
&\eta \sum_{k=1}^{K} \sum_{\pi \in \Omega} \omega^k(\pi) \widehat{c}^k(\pi) \sum_{j=1, j+d^j \geq k}^{k-1} \widehat{c}^k(\pi) \\
&\leq \eta \sum_{k=1}^{K} \sum_{j=1, j+d^j \geq k}^{k-1} \sum_{\pi \in \Omega} \omega^k(\pi) \left(\sum_{h=1}^{H} \sum_{s,a} q_h^{\pi, \bar{p}^k}(s,a) \widehat{c}_h^k(s,a) \right) \left(\sum_{h=1}^{H} \sum_{s,a} q_h^{\pi, \bar{p}^j}(s,a) \widehat{c}_h^j(s,a) \right) \\
&= \eta \sum_{k=1}^{K} \sum_{j=1, j+d^j \geq k}^{k-1} \sum_{h=1}^{H} \sum_{s,a} \sum_{h'=1}^{H} \sum_{s',a'} \widehat{c}_h^k(s,a) \widehat{c}_{h'}^j(s',a') \left(\sum_{\pi \in \Omega} \omega^k(\pi) \cdot q_h^{\pi, \bar{p}^k}(s,a) q_{h'}^{\pi, \bar{p}^j}(s',a') \right)
\end{aligned}
$$

where the first step uses the definition of loss estimators. Similarly, we have the following inequality holds with probability at least $1 - 12\delta$ that

$$
\begin{aligned}
&\eta \sum_{k=1}^{K} \sum_{j=1, j+d^j \geq k}^{k-1} \sum_{h=1}^{H} \sum_{s,a} \sum_{h'=1}^{H} \sum_{s',a'} \widehat{c}_h^k(s,a) \widehat{c}_{h'}^j(s',a') \left(\sum_{\pi \in \Omega} \omega^k(\pi) \cdot q_h^{\pi, \bar{p}^k}(s,a) q_{h'}^{\pi, \bar{p}^j}(s',a') \right) \\
&\leq \eta \sum_{k=1}^{K} \sum_{j=1, j+d^j \geq k}^{k-1} \sum_{h=1}^{H} \sum_{s,a} \sum_{h'=1}^{H} \sum_{s',a'} \widehat{c}_h^k(s,a) \left(\sum_{\pi \in \Omega} \omega^k(\pi) \cdot q_h^{\pi, \bar{p}^k}(s,a) q_{h'}^{\pi, \bar{p}^j}(s',a') \right) + \mathcal{O}\left(\frac{\eta}{\gamma} H^2 d_{max} \iota \right) \\
&\leq \eta \sum_{k=1}^{K} \sum_{j=1, j+d^j \geq k}^{k-1} \sum_{h=1}^{H} \sum_{s,a} \sum_{h'=1}^{H} \sum_{s',a'} \sum_{\pi \in \Omega} \omega^k(\pi) \cdot q_h^{\pi, \bar{p}^k}(s,a) q_{h'}^{\pi, \bar{p}^j}(s',a') + \mathcal{O}\left(\frac{\eta}{\gamma} H^2 d_{max} \iota \right) \\
&= \eta \sum_{k=1}^{K} \sum_{j=1, j+d^j \geq k}^{k-1} \sum_{\pi \in \Omega} \omega^k(\pi) \left(\sum_{h=1}^{H} \sum_{s,a} q_h^{\pi, \bar{p}^k}(s,a) \right) \left(\sum_{h=1}^{H} \sum_{s,a} q_h^{\pi, \bar{p}^j}(s,a) \right) + \mathcal{O}\left(\frac{\eta}{\gamma} H^2 d_{max} \iota \right) \\
&= \mathcal{O}\left(\frac{\eta}{\gamma} H^2 d_{max} \iota \right) + \eta H^2 \sum_{k=1}^{K} \sum_{j=1, j+d^j \geq k}^{k-1} 1 = \mathcal{O}\left(\frac{\eta}{\gamma} H^2 d_{max} \iota \right) + \eta H^2 \sum_{j=1}^{K} \sum_{k=1, k>j, k \leq j+d^j}^{K} 1 \\
&= \mathcal{O}\left(\eta H^2 D + \frac{\eta}{\gamma} H^2 d_{max} \iota \right)
\end{aligned} \tag{24}
$$

where the first and second step apply Lemma 11 of Jin et al. (2020a) twice under the event that $p \in \cap_k \mathcal{P}^k$, based on the fact that $q_{h'}^{\pi, \bar{p}^j}(s',a') \leq 1$ and $\sum_{\pi \in \Omega} \omega^k(\pi) \cdot q_h^{\pi, \bar{p}^k}(s,a) q_{h'}^{\pi, \bar{p}^j}(s',a') \leq \sum_{\pi \in \Omega} \omega^k(\pi) \cdot q_h^{\pi, \bar{p}^k}(s,a) \leq u_h^k(s,a)$.

Combining Equations (22) to (24) yields the following bound of DRIFT with probability at least $1 - 30\delta$ under the event $p \in \cap_{k=1}^K \mathcal{P}^k$:

$$
\text{DRIFT} = \mathcal{O}\left(\eta H^2 \left(D + H^2 SAK \right) + \frac{\eta}{\gamma} H^2 \left(d_{max} + 1 \right) \iota \right). \tag{25}
$$

Finally, combining the bounds for CHEATING-REGRET and DRIFT in Equations (18) and (25) concludes the proof. ■

Algorithm 6 Delayed UOB-FTRL with Normal Loss Estimator

Input: State space $\mathcal{S}$, Action space $\mathcal{A}$, Horizon H, Number of episodes K, Learning rate $\eta > 0$, Exploration parameter $\gamma > 0$, Confidence parameter $\delta > 0$.

Initialization: Set $\pi_h^1(a \mid s) = \frac{1}{A}$, $q_h^1(s,a,s') = \frac{1}{S^2 A}$, $n_h^1(s,a) = 0, n_h^1(s,a,s')$ for every $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$ and $\mathcal{P}^1$ be the set of all transition functions.

for $k = 1, 2, ..., K$ **do**

  Play episode k with policy π^k and observe trajectory $\{(s_h^k, a_h^k)\}_{h=1}^H$.

  Define confidence set $\mathcal{P}^{k+1}$ by Algorithm 5.

  for $j : j + d^j = k$ **do**

    Observe feedback $\{c_h^j(s_h^j, a_h^j)\}_{h=1}^H$.

    Compute upper occupancy bound $u_h^j(s,a) = \max_{p' \in \mathcal{P}^j} q_h^{p',\pi^j}(s,a)$.

    Compute loss estimator $\widehat{c}_h^j(s,a) = \frac{c_h^j(s,a)\mathbb{I}\{s_h^j = s, a_h^j = a\}}{u_h^j(s,a)+\gamma}$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

  end for

  Update occupancy measure:

$$q^{k+1} = \arg\min_{q \in \cap_{j=1}^{k+1} \Delta(\mathcal{M},j)} \left\langle q, \sum_{j: j+d^j \le k} \widehat{c}^j \right\rangle + \phi(q),$$

  where $\phi(q) = \frac{1}{\eta}\sum_{h,s,a,s'} q_h(s,a,s') \log q_h(s,a,s')$ is the Shannon entropy regularizer, and $\Delta(\mathcal{M},k) = \{q^{\pi,p'} \mid \pi \in (\Delta_{\mathcal{A}})^{\mathcal{S}\times[H]}, p' \in \mathcal{P}^k\}$.

  Update policy: $\pi_h^{k+1}(a \mid s) = \frac{\sum_{s'} q_h^{k+1}(s,a,s')}{\sum_{a'}\sum_{s'} q_h^{k+1}(s,a',s')}$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

end for

Appendix B. FTRL with normal loss estimator

In this section, we show that applying the FTRL framework with normal loss estimators and fixed amount Shannon entropy can achieve $\widetilde{\mathcal{O}}\left(\sqrt{K} + \sqrt{D}\right)$ expected regret (ignoring dependence on other parameters). We propose Algorithm 6 which based on this simple idea and Theorem 9 below shows that our algorithm essentially achieves this goal.

As one may noticed that, compared with Algorithm 8 which uses the Online Mirror Descent framework, Algorithm 6 uses $\cap_{j=1}^k \Delta(\mathcal{M},j)$, the set of occupancy measures associated with transition functions that belong to all confidence sets prior to episode k, as the decision space to compute q^k. This setup is necessary to adopt the FTRL framework for ensuring that a shrinking sequence of decision sets, which is critical to analyze the penalty term as in Lemma 15. Please see the proof of Lemma 15 for more details. On the other hand, the unknown underlying transition p belongs to all the confidence sets with high probability, which ensures that the intersection of confidence sets is nonempty with high probability.

Theorem 9 *With confidence parameter* $\delta = \frac{1}{H^2 S^2 A^2 K^5}$, *learning rate* $\eta = \sqrt{\frac{H \log \frac{HSAK}{\delta}}{HSAK + (HSA)^2 D}}$ *and exploration parameter* $\gamma = \sqrt{\frac{\log \frac{HSAK}{\delta}}{SAK}}$, *Algorithm 6 ensures that*

$$\mathbb{E}\left[R_K\right] = O\left(H^2 S \sqrt{AK \log(HSAK)} + HSA\sqrt{HD \log(HSAK)} + H^4 S^2 A^2 \log^2(HSAK)\right).$$

B.1 Proof of the Main Theorem

We first decompose the regret into four terms according to the work of Jin et al. (2020a):

$$R_K = \underbrace{\sum_{k=1}^{K} \left\langle q^{\pi^k} - q^k, c^k \right\rangle}_{\textsc{Est}} + \underbrace{\sum_{k=1}^{K} \left\langle q^k, c^k - \widehat{c}^k \right\rangle}_{\textsc{Bias}_1} + \underbrace{\sum_{k=1}^{K} \left\langle q^k - q^*, \widehat{c}^k \right\rangle}_{\textsc{Reg}} + \underbrace{\sum_{k=1}^{K} \left\langle q^\star, \widehat{c}^k - c^k \right\rangle}_{\textsc{Bias}_2}, \tag{26}$$

where q^k is the computed occupancy measure of episode k; q^{π^k} is the underlying occupancy measure associated with the unknown transition p and policy π^k; $q^\star$ is the occupancy measure of the optimal policy $\pi^\star$ in hindsight .

Then, with the help of Lemma 4, 6 and 14 of Jin et al. (2020a), we have the following lemma for $\textsc{Est}$, $\textsc{Bias}_1$ and $\textsc{Bias}_2$.

Lemma 10 *with probability at least* $1 - 9\delta$*, Algorithm 6 ensures that*

$$\textsc{Est} = \mathcal{O}\left(H^2 S \sqrt{AK \log\left(\frac{HSAK}{\delta}\right)} + H^4 S^2 A^2 \log^2\left(\frac{HSAK}{\delta}\right) \right),$$

$$\textsc{Bias}_1 = \mathcal{O}\left(H^2 S \sqrt{AK \log\left(\frac{HSAK}{\delta}\right)} + \gamma HSAK \right),$$

$$\textsc{Bias}_2 = \mathcal{O}\left(\frac{H}{\gamma} \log\left(\frac{HSA}{\delta}\right) \right).$$

Proof Without loss of generality, we convert our MDP setting to that of Jin et al. (2020a) by setting $\mathcal{X} = \mathcal{S} \times [H]$ and $L = H$. Then, by direct application of Lemma 4, 6 and 14 of Jin et al. (2020a) (which are combined together in the proof of Theorem 3), we arrive at the high-probability bounds of these terms. Note that, the double epoch scheduling and larger confidence sets of transition functions only changes the constant of regret bound, which is hidden in $\mathcal{O}(\cdot)$ operator. ■

Based on the high-probability bound, we have the following corollary for the expected bound of these terms.

Corollary 11 *Algorithm 6 ensures that* $\mathbb{E}\left[\textsc{Est} + \textsc{Bias}_1 + \textsc{Bias}_2\right]$ *is bounded at most* $\mathcal{O}\left(H^4 S^2 A^2 \log^2\left(\frac{HSAK}{\delta}\right)\right)$ *plus:*

$$\mathcal{O}\left(H^2 S \sqrt{AK \log\left(\frac{HSAK}{\delta}\right)} + \gamma HSAK + \frac{H}{\gamma} \log\left(\frac{HSA}{\delta}\right) + HK\delta \right).$$

Then, we prove the following lemma for the expected bound of $\textsc{Reg}$ with the help a unique novel analysis, and defer the complete proof to to Appendix B.2.

Lemma 12 *Algorithm 6 ensures that* $\mathbb{E}\left[\textsc{Reg}\right]$ *is bounded by:*

$$\mathcal{O}\left(\frac{H \ln\left(S^2 A\right)}{\eta} + \eta\left(HSAK + (HSA)^2 D\right) + \frac{H^2 S^2 A^2 K^3}{\gamma^2} \delta \right).$$

With the help of above lemmas, we are ready to prove the Theorem 9.

Proof of Theorem 9 Combining the expected bound of $\text{EST} + \text{BIAS}_1 + \text{BIAS}_2$ in Corollary 11 and that of REG in Lemma 12, we are able to show that the expected regret $\mathbb{E}[R_K]$ is bounded by

$$\mathcal{O}\left(H^2 S\sqrt{AK\log\left(\frac{HSAK}{\delta}\right)} + \gamma HSAK + \frac{H}{\gamma}\log\left(\frac{HSA}{\delta}\right) + \frac{H\ln\left(S^2A\right)}{\eta} + \eta\left(HSAK + (HSA)^2 D\right)\right)$$
$$+ \mathcal{O}\left(\frac{H^2S^2A^2K^3}{\gamma^2}\delta + H^4S^2A^2\log^2\left(\frac{HSAK}{\delta}\right)\right).$$

Finally, selecting a small enough confidence parameter $\delta = \frac{1}{H^2S^2A^2K^5}$ and picking up the learning rate $\eta = \sqrt{\frac{H\log\frac{HSAK}{\delta}}{HSAK+(HSA)^2D}}$ and the exploration parameter $\gamma = \sqrt{\frac{\log\frac{HSAK}{\delta}}{SAK}}$ ensure that

$$\mathbb{E}[R_K] = O\left(H^2S\sqrt{AK\log(HSAK)} + HSA\sqrt{HD\log(HSAK)} + H^4S^2A^2\log^2(HSAK)\right).$$

■

B.2 Bound on the Regret with respect to the Loss Estimators (REG in Equation (26))

In this part, we focus on REG defined in Eq (26) with delayed feedback of losses, and prove Lemma 12 through the introduced key steps in Section 4. To this end, we will use the following decomposition of REG in this section:

$$\begin{aligned}
\text{REG} = \sum_{k=1}^{K}\left\langle q^k - q^\star, \widehat{c}^k\right\rangle &= \sum_{k=1}^{K}\Phi_k(q^k) + \left\langle q^k, \widehat{c}^k\right\rangle - \Phi_k'(\widehat{q}^k) && (\text{STABILITY})\\
&+ \sum_{k=1}^{K}\Phi_k'(\widehat{q}^k) - \Phi_k(q^k) - \left(\Phi_k^C(\widetilde{q}_k') - \Phi_k^B(\widetilde{q}_k)\right) && (\text{DELAY-CAUSED DRIFT})\\
&+ \sum_{k=1}^{K}\Phi_k^C(\widetilde{q}_k') - \Phi_k^B(\widetilde{q}_k) - \left\langle q^\star, \widehat{c}^k\right\rangle && (\text{PENALTY})
\end{aligned}$$

where the functions $\Phi_k, \Phi_k', \Phi_k^B, \Phi_k^C$ and the occupancy measures $q^k, \widehat{q}^k, \widetilde{q}_k, \widetilde{q}_k'$ are defined as

$$\begin{aligned}
\Phi_k(q) &= \left\langle q, \widehat{L}_k^{\text{obs}}\right\rangle + \phi(q), & q^k &= \mathop{\arg\min}_{q\in\cap_{j=1}^k\Delta(\mathcal{M},j)} \Phi_k(q),\\
\Phi_k'(q) &= \left\langle q, \widehat{L}_k^{\text{obs}} + \widehat{c}^k\right\rangle + \phi(q), & \widehat{q}^k &= \mathop{\arg\min}_{q\in\cap_{j=1}^k\Delta(\mathcal{M},j)} \Phi_k'(q),\\
\Phi_k^B(q) &= \left\langle q, \widehat{L}_k\right\rangle + \phi(q), & \widetilde{q}_k &= \mathop{\arg\min}_{q\in\cap_{j=1}^k\Delta(\mathcal{M},j)} \Phi_k^B(q),\\
\Phi_k^C(q) &= \left\langle q, \widehat{L}_k + \widehat{c}^k\right\rangle + \phi(q), & \widetilde{q}_k' &= \mathop{\arg\min}_{q\in\cap_{j=1}^k\Delta(\mathcal{M},j)} \Phi_k^C(q).
\end{aligned}$$

with $\widehat{L}_k = \sum_{j=1}^{k-1}\widehat{c}^j$ being the un-delayed cumulative loss estimator prior to episode k, and $\widehat{L}_k^{\text{obs}} = \sum_{j=1,j+d^j<k}^{k-1}\widehat{c}^j$ being the received cumulative loss estimator.

On the other hand, with the help of $F_k^\star(x) = -\min_{q\in\cap_{j=1}^k\Delta(\mathcal{M},j)}\{\phi(x) - \langle x, q\rangle\}$, the convex conjugate with respect to $\phi(\cdot)$, these functions and occupancy measures ensures that

$$\Phi_k(q^k) = -F_k^\star\left(-\widehat{L}_k^{\text{obs}}\right), \Phi_k'(\widehat{q}^k) = -F_k^\star\left(-\widehat{L}_k^{\text{obs}} - \widehat{c}^k\right), \Phi_k^B(\widetilde{q}_k) = -F_k^\star\left(-\widehat{L}_k\right), \Phi_k^C(\widetilde{q}_k') = -F_k^\star\left(-\widehat{L}_k - \widehat{c}^k\right).$$

In addition, according to the property of convex conjugates, these occupancy measures are able to be presented as the gradient of the convex conjugate with different inputs as

$$q^k = \nabla F_k^\star\left(-\widehat{L}_k^{\text{obs}}\right), \widehat{q}^k = \nabla F_k^\star\left(-\widehat{L}_k^{\text{obs}} - \widehat{c}^k\right), \widetilde{q}_k = \nabla F_k^\star\left(-\widehat{L}_k\right), \widetilde{q}_k' = \nabla F_k^\star\left(-\widehat{L}_k - \widehat{c}^k\right).$$

For notational convenience, we denote $\widehat{\Delta}_k = \widehat{L}_k - \widehat{L}_k^{\text{obs}}$ as the summation of un-received loss estimators prior to episode k, that is, $\widehat{\Delta}_k = \sum_{j=1, j+d^j \geq k}^{k-1} \widehat{c}_j$. Thus, $\Phi_k^B(\widetilde{q}_k)$ and $\Phi_k^C(\widetilde{q}'_k) = -F_k^\star\left(-\widehat{L}_k - \widehat{c}^k\right)$ can be represented as

$$\Phi_k^B(\widetilde{q}'_k) = -F_k^\star\left(-\widehat{L}_k^{\text{obs}} - \widehat{\Delta}_k\right), \Phi_k^C(\widetilde{q}'_k) = -F_k^\star\left(-\widehat{L}_k^{\text{obs}} - \widehat{\Delta}_k - \widehat{c}^k\right).$$

With the help of these definitions, we are now ready to bound the terms STABILTY, DELAY-CAUSED DRFIT and PENALTY in following lemmas.

Lemma 13 *(Stability) With fixed learning rate $\eta > 0$ and exploration $\gamma > 0$, Algorithm 6 ensures that*

$$\sum_{k=1}^{K} \Phi_k(q^k) + \left\langle q^k, \widehat{c}^k \right\rangle - \Phi'_k(\widehat{q}^k) \leq \eta \sum_{k=1}^{K} \sum_{h,s,a} q_h^k(s,a) \widehat{c}_h^k(s,a)^2.$$

Proof Let $D_k(u,v) = \phi(u) - \phi(v) - \langle u - v, \nabla\phi(v)\rangle$ be the Bregman divergence with the convex regularizer ϕ. Then,

$$\begin{aligned}
\Phi_k(q^k) &= \left\langle q^k, \widehat{L}_k^{\text{obs}} \right\rangle + \phi(q^k) = \left\langle \widehat{q}^k, \widehat{L}_k^{\text{obs}} \right\rangle + \phi(\widehat{q}^k) - \left(\left\langle \widehat{q}^k - q^k, \widehat{L}_k^{\text{obs}} \right\rangle + \phi(\widehat{q}^k) - \phi(q^k)\right) \\
&\leq \left\langle \widehat{q}^k, \widehat{L}_k^{\text{obs}} \right\rangle + \phi(\widehat{q}^k) - \left(-\left\langle \widehat{q}^k - q^k, \nabla\phi(q^k)\right\rangle + \phi(\widehat{q}^k) - \phi(q^k)\right) \\
&= \left\langle \widehat{q}^k, \widehat{L}_k^{\text{obs}} \right\rangle + \phi(\widehat{q}^k) - D_k(\widehat{q}^k, q^k) = \Phi'_k(\widehat{q}^k) - \left\langle \widehat{q}^k, \widehat{c}^k \right\rangle - D_k(\widehat{q}^k, q^k),
\end{aligned}$$

where the third step follows from the first order optimality of q^k with respect to Φ_k, in other words, $\left\langle \widehat{q}^k - q^k, \widehat{L}_k^{\text{obs}} + \nabla\phi(q^k) \right\rangle \geq 0$. Rearranging terms and adding $\left\langle q^k, \widehat{c}^k \right\rangle$ on both sides give us the following inequality:

$$\Phi_k(q^k) + \left\langle q^k, \widehat{c}^k \right\rangle - \Phi'_k(\widehat{q}^k) \leq \left\langle q^k - \widehat{q}^k, \widehat{c}^k \right\rangle - D_k(\widehat{q}^k, q^k).$$

To bound the right hand side term, we relax the constraints and taking the maximum as:

$$\left\langle q^k - \widehat{q}^k, \widehat{c}^k \right\rangle - D_k(\widehat{q}^k, q^k) \leq \max_{q \in \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}\times[H]\times S}} \left\langle q^k - q, \widehat{c}^k \right\rangle - D_k(q, q^k) = \left\langle q^k - \xi^k, \widehat{c}^k \right\rangle - D_k(\xi^k, q^k),$$

where ξ_k denotes the maximizer point. Setting the gradient to zero gives the equality that $\nabla\phi(q^k) - \nabla\phi(\xi^k) = \widehat{c}^k$. By direct calculation, one can verify that $\xi_h^k(s,a,s') = q_h^k(s,a,s') \cdot \exp\left(-\eta \widehat{c}_h^k(s,a)\right)$ for all transition tuples. Therefore, we have the following inequality that

$$\begin{aligned}
\left\langle q^k - \xi^k, \widehat{c}^k \right\rangle - D_k(\xi^k, q^k) &= \left\langle q^k - \xi^k, \widehat{c}^k \right\rangle - \phi(\xi^k) + \phi(q^k) - \left\langle q^k - \xi^k, \nabla\phi(q^k)\right\rangle = D_k(q^k, \xi^k) \\
&= \frac{1}{\eta} \sum_{h=1}^{H} \sum_{s,a,s'} \left(q_h^k(s,a,s') \ln\left(\frac{q_h^k(s,a,s')}{\xi_h^k(s,a,s')}\right) - q_h^k(s,a,s') + \xi_h^k(s,a,s') \right) \\
&= \frac{1}{\eta} \sum_{h=1}^{H} \sum_{s,a,s'} q_h^k(s,a,s') \left(\eta \widehat{c}_h^k(s,a) - 1 + \exp\left(-\eta \widehat{c}_h^k(s,a)\right)\right) \\
&\leq \eta \sum_{h=1}^{H} \sum_{s,a,s'} q_h^k(s,a,s') \widehat{c}_h^k(s,a)^2 = \eta \sum_{h=1}^{H} \sum_{s,a} q_h^k(s,a) \widehat{c}_h^k(s,a)^2,
\end{aligned}$$

where the second step uses $\nabla\phi(q^k) - \nabla\phi(\xi_k) = \widehat{c}^k$; the forth step follows from the fact that $e^{-x} \leq 1 - x + x^2$ for any $x \geq 0$. Finally, taking the summation over all episodes finishes the proof. ■

Lemma 14 *(Delay-caused Drift) Algorithm 6 guarantees that*

$$\sum_{k=1}^{K} \Phi'_k(\widehat{q}^k) - \Phi_k(q^k) - \left(\Phi_k^C(\widetilde{q}'_k) - \Phi_k^B(\widetilde{q}_k)\right) \le 2\eta \sum_{k=1}^{K} \left(\sum_{h=1}^{H}\sum_{s,a} \widehat{c}_h^k(s,a)\right) \cdot \left(\sum_{h=1}^{H}\sum_{s,a} \widehat{\Delta}_h^k(s,a)\right).$$

Proof With the help of the convex conjugate $F_k^\star(\cdot)$, we have the following inequality holds for some $\theta \in [0,1]$ that:

$$\begin{aligned}
\Phi'_k(\widehat{q}^k) - \Phi_k(q^k) - \left(\Phi_k^C(\widetilde{q}'_k) - \Phi_k^B(\widetilde{q}_k)\right) &= -F_k^\star(-\widehat{L}_k^{\text{obs}} - \widehat{c}^k) + F_k^\star(-\widehat{L}_k^{\text{obs}}) - \left(-F_k^\star(-\widehat{L}_k - \widehat{c}^k) + F_k^\star(-\widehat{L}_k)\right) \\
&= \int_0^1 \left\langle \widehat{c}^k, \nabla F_k^\star(-\widehat{L}_k^{\text{obs}} - x\widehat{c}^k)\right\rangle dx - \int_0^1 \left\langle \widehat{c}^k, \nabla F_k^\star(-\widehat{L}_k - x\widehat{c}^k)\right\rangle dx \\
&= \int_0^1 \left\langle \widehat{c}^k, \nabla F_k^\star(-\widehat{L}_k^{\text{obs}} - x\widehat{c}^k) - \nabla F_k^\star(-\widehat{L}_k - x\widehat{c}^k)\right\rangle dx \\
&= \left\langle \widehat{c}^k, \nabla F_k^\star(-\widehat{L}_k^{\text{obs}} - \theta\widehat{c}^k) - \nabla F_k^\star(-\widehat{L}_k - \theta\widehat{c}^k)\right\rangle,
\end{aligned}$$

where the second step uses Newton-Leibniz theorem; the forth step uses the mean value theorem. To analyze the right hand side, we define the functions W and W' as

$$W(q) = \left\langle q, \widehat{L}_k^{\text{obs}} + \theta\widehat{c}^k\right\rangle + \phi(q) \quad ; \quad W'(q) = \left\langle q, \widehat{L}_k + \theta\widehat{c}^k\right\rangle + \phi(q),$$

and denote their minimizer occupancy measures within the decision set $\cap_{j=1}^k \Delta(\mathcal{M}, j)$ by u and v. According to the properties of convex conjugate, we have $u = \nabla F_k^\star(-\widehat{L}_k^{\text{obs}} - \theta\widehat{c}^k)$ and $v = \nabla F_k^\star(-\widehat{L}_k - \theta\widehat{c}^k)$.

To analyze $\langle u - v, \widehat{c}^k\rangle$, we first lower bound $W(u) + \left\langle u, \widehat{\Delta}_k\right\rangle - W'(v)$ as

$$W(u) + \left\langle u, \widehat{\Delta}_k\right\rangle - W'(v) = W'(u) - W'(v) = \langle \nabla W'(v), u - v\rangle + \frac{1}{2}\|u - v\|^2_{\nabla^2\phi(\xi)} \ge \frac{1}{2}\|u - v\|^2_{\nabla^2\phi(\xi)},$$

where the second step applies Taylor's expansion with ξ being an intermediate point between u and v; the last step uses the first order optimality condition of v. On the other hand, we can upper $W(u) + \left\langle u, \widehat{\Delta}_k\right\rangle - W'(v)$ as

$$\begin{aligned}
W(u) + \left\langle u, \widehat{L}_k - \widehat{L}_k^{\text{obs}}\right\rangle - W'(v) &= W(u) - W(v) + \left\langle u - v, \widehat{L}_k - \widehat{L}_k^{\text{obs}}\right\rangle \le \left\langle u - v, \widehat{L}_k - \widehat{L}_k^{\text{obs}}\right\rangle \\
&\le \|u - v\|_{\nabla^2\phi(\xi)} \left\|\widehat{L}_k - \widehat{L}_k^{\text{obs}}\right\|_{\nabla^{-2}\phi(\xi)},
\end{aligned}$$

where the second step uses the optimality of u, and the last step comes from Hölder's inequality. Combining the lower bound and upper bound, we arrives at the following inequality

$$\|u - v\|_{\nabla^2\phi(\xi)} \le 2\left\|\widehat{L}_k - \widehat{L}_k^{\text{obs}}\right\|_{\nabla^{-2}\phi(\xi)}.$$

Therefore, we can upper bound the term $\langle \widehat{c}^k, u - v\rangle$ with the help of Hölder's inequality again as

$$\langle \widehat{c}^k, u - v\rangle \le \left\|\widehat{c}^k\right\|_{\nabla^{-2}\phi(\xi)} \|u - v\|_{\nabla^2\phi(\xi)} \le 2\left\|\widehat{c}^k\right\|_{\nabla^{-2}\phi(\xi)} \left\|\widehat{L}_k - \widehat{L}_k^{\text{obs}}\right\|_{\nabla^{-2}\phi(\xi)}.$$

By direct calculation, one can verify the following:

$$\begin{aligned}
2\left\|\widehat{c}^k\right\|_{\nabla^{-2}\phi(\xi)} \cdot \left\|\widehat{\Delta}_k\right\|_{\nabla^{-2}\phi(\xi)} &= 2\sqrt{\eta\sum_{h=1}^{H}\sum_{s,a,s'} \widehat{c}_h^k(s,a)^2\xi(s,a,s')} \cdot \sqrt{\eta\sum_{h=1}^{H}\sum_{s,a,s'} \widehat{\Delta}_h^k(s,a)^2\xi(s,a,s')} \\
&\le 2\eta\sqrt{\sum_{h=1}^{H}\sum_{s,a} \widehat{c}_h^k(s,a)^2} \cdot \sqrt{\sum_{h=1}^{H}\sum_{s,a} \widehat{\Delta}_h^k(s,a)^2} \\
&\le 2\eta\left(\sum_{h=1}^{H}\sum_{s,a} \widehat{c}_h^k(s,a)\right) \cdot \left(\sum_{h=1}^{H}\sum_{s,a} \widehat{\Delta}_h^k(s,a)\right),
\end{aligned}$$

where the second step follows from the fact that ξ is a valid occupancy measure and $\sum_{s'} \xi(s,a,s') = \xi(s,a) \leq 1$ holds for all state-action pairs. Taking the summation over all episodes concludes the proof. ■

Lemma 15 *(Penalty) With the shrinking decision set sequence that* $\cap_{j=1}^{k+1} \Delta(\mathcal{M}, j) \subset \cap_{j=1}^{k} \Delta(\mathcal{M}, j)$ *for* $k = 1, \ldots K-1$*, Algorithm 6 ensures that*

$$\sum_{k=1}^{K} \Phi_k^C(\widetilde{q}'_k) - \Phi_k^B(\widetilde{q}_k) - \left\langle q^\star, \widehat{c}^k \right\rangle \leq \frac{H \ln\left(S^2 A\right)}{\eta}.$$

Proof First, we observe that

$$\begin{aligned}\Phi_k^C(\widetilde{q}'_k) &= \min_{q \in \cap_{j=1}^{k} \Delta(\mathcal{M},j)} \left\langle q, \widehat{L}_k + \widehat{c}^k \right\rangle + \phi(q) \leq \min_{q \in \cap_{j=1}^{k+1} \Delta(\mathcal{M},j)} \left\langle q, \widehat{L}_k + \widehat{c}^k \right\rangle + \phi(q) \\ &= \min_{q \in \cap_{j=1}^{k+1} \Delta(\mathcal{M},j)} \left\langle q, \widehat{L}_{k+1} \right\rangle + \phi(q) = \Phi_{k+1}^B(\widetilde{q}_{k+1}),\end{aligned}$$

where the second step follows from the fact that $\mathcal{P}^{k+1} \subset \mathcal{P}^k$ by the definition. Therefore, we have the following inequality:

$$\begin{aligned}\sum_{k=1}^{K} \Phi_k^C(\widetilde{q}'_k) - \Phi_k^B(\widetilde{q}_k) - \left\langle q^\star, \widehat{c}^k \right\rangle &= \Phi_K^C(\widetilde{q}'_K) - \Phi_1^B(\widetilde{q}_1) - \left\langle q^\star, \widehat{L}_{K+1} \right\rangle + \sum_{k=1}^{K-1} \Phi_k^C(\widetilde{q}'_k) - \Phi_{k+1}^B(\widetilde{q}_{k+1}) \\ &\leq \Phi_K^C(\widetilde{q}'_K) - \Phi_1^B(\widetilde{q}_1) - \left\langle q^\star, \widehat{L}_{K+1} \right\rangle \leq \phi(q^\star) - \phi(\widetilde{q}_1) \leq \frac{H \ln\left(S^2 A\right)}{\eta},\end{aligned}$$

where the third step follows from the optimality of $\widetilde{q}'_K$ and the last steps follows the standard argument of Shannon entropy (such as, Lemma 12 of Jin et al. (2020a)). ■

We are now ready to prove Lemma 12 by combining the results of Lemmas 13 to 15 and taking the expectation.

Proof of Lemma 12 By combining Lemmas 13 to 15, we have Reg bounded as

$$\textsc{Reg} \leq \frac{H \ln\left(S^2 A\right)}{\eta} + \eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s,a} q_h^k(s,a) \widehat{c}_h^k(s,a)^2 + 2\eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s,a} \sum_{h'=1}^{H} \sum_{s',a'} \widehat{c}_h^k(s,a) \widehat{\Delta}_{h'}^k(s',a').$$

To analyze the expectation, we use the indicator $Z_k = \mathbb{I}\{p \notin \mathcal{P}_k\}$ to denote the event that the true transition function p is not included in the confidence set of episode k. Clearly, one can verify that $q_h^k(s,a) \leq Z_k + u_h^k(s,a)$ and $q_h^{\pi^k}(s,a) \leq Z_k + u_h^k(s,a)$ due to the definition of upper occupancy bound u_k and the property of occupancy measures. Therefore, we are able to bound $\mathbb{E}\left[\textsc{Reg}\right]$ by

$$\begin{aligned}&\frac{H \ln\left(S^2 A\right)}{\eta} + \eta \mathbb{E}\left[\sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s,a} q_h^k(s,a) \widehat{c}_h^k(s,a)^2 + 2 \sum_{h=1}^{H} \sum_{s,a} \sum_{h'=1}^{H} \sum_{s',a'} \widehat{c}_h^k(s,a) \widehat{\Delta}_h^k(s,a)\right] \\ &\leq \frac{H \ln\left(S^2 A\right)}{\eta} + \eta \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}_k\left[\sum_{h=1}^{H} \sum_{s,a} \widehat{c}_h^k(s,a) + 2\eta \sum_{h=1}^{H} \sum_{s,a} \sum_{h'=1}^{H} \sum_{s',a'} \widehat{c}_h^k(s,a) \widehat{\Delta}_h^k(s,a)\right]\right] \\ &\leq \frac{H \ln\left(S^2 A\right)}{\eta} + \eta \mathbb{E}\left[\sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s,a} \frac{q_h^{\pi^k}(s,a)}{u_h^k(s,a) + \gamma} + 2 \sum_{j=1, j+d^j \geq k}^{k-1} \sum_{h'=1}^{H} \sum_{s',a'} \frac{q_h^{\pi^k}(s,a)}{u_h^k(s,a) + \gamma} \frac{q_{h'}^{\pi^j}(s',a')}{u_{h'}^j(s',a') + \gamma}\right] \\ &\leq \frac{H \ln\left(S^2 A\right)}{\eta} + \eta \left(HSAK + 2(HSA)^2 D\right) + \frac{HSAK + 4(HSA)^2 D}{\gamma^2} \cdot \mathbb{E}\left[\sum_{k=1}^{K} Z_k\right],\end{aligned}$$

where the first step uses the fact that $q_h^k(s,a) \leq u_h^k(s,a)$ for any state-action pair; the second step uses the definition of loss estimators; the third step follows from the fact that $q_h^{\pi^k}(s,a) \leq Z_k + u_h^k(s,a)$.

According to Lemma 2 of Jin et al. (2020a), we have the expectation of $\mathbb{E}\left[\sum_{k=1}^K Z_k\right]$ bounded by $4K\delta$, and the following upper bound of $\mathbb{E}\left[\textsc{Reg}\right]$:

$$\mathcal{O}\left(\frac{H\ln\left(S^2A\right)}{\eta} + \eta\left(HSAK + (HSA)^2D\right) + \frac{H^2S^2A^2K^3}{\gamma^2}\delta\right).$$

■

Algorithm 7 Delayed O-REPS with delay-adapted estimator and known transition

Input: State space $\mathcal{S}$, Action space $\mathcal{A}$, Horizon H, Number of episodes K, Transition function p, Learning rate $\eta > 0$, Exploration parameter $\gamma > 0$.
Initialization: Set $\pi_h^1(a \mid s) = \frac{1}{A}$, $q_h^1(s,a) = \frac{1}{SA}$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
for $k = 1, 2, ..., K$ **do**
  Play episode k with policy π^k and observe trajectory $\{(s_h^k, a_h^k)\}_{h=1}^H$.
  for $j : j + d^j = k$ **do**
    Observe feedback $\{c_h^j(s_h^j, a_h^j)\}_{h=1}^H$.
    Compute loss estimator $\hat{c}_h^j(s,a) = \frac{c_h^j(s,a)\mathbb{I}\{s_h^j = s, a_h^j = a\}}{\max\{q_h^j(s,a), q_h^k(s,a)\} + \gamma}$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
  end for
  Update occupancy measure:

$$q^{k+1} = \arg\min_{q \in \Delta(\mathcal{M})} \eta \left\langle q, \sum_{j:j+d^j=k} \hat{c}^j \right\rangle + \mathrm{KL}(q \parallel q^k), \tag{27}$$

  where $\mathrm{KL}(q \parallel q') = \sum_{h,s,a} q_h(s,a) \ln \frac{q_h(s,a)}{q'_h(s,a)} + q'_h(s,a) - q_h(s,a)$.
  Update policy: $\pi_h^{k+1}(a \mid s) = \frac{q_h^{k+1}(s,a)}{\sum_{a'} q_h^{k+1}(s,a')}$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
end for

Appendix C. Delayed O-REPS with delay-adapted estimator

Explicitly solving this optimization problem in Equation (27), we get (Zimin and Neu, 2013):

$$q_h^{k+1}(s,a) = \frac{q_h^k(s,a) e^{B_h^k(s,a|v^k)}}{Z_h^k(v^k)},$$

for:

$$B_h^k(s,a \mid v) = v_h(s) - \eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a) - \sum_{s'} p_h(s' \mid s,a) v_{h+1}(s')$$
$$Z_h^k(v) = \sum_{s,a} q_h^k(s,a) e^{B_h^k(s,a|v)}$$
$$v^k = \arg\min_v \sum_h \log Z_h^k(v).$$

These different formulations will be helpful in the regret analysis.

Theorem 16 *Running O-REPS with the delay-adapted estimator,* $\eta = \gamma = \min\{\sqrt{\frac{\log\frac{HSA}{\delta}}{SAK}}, \sqrt{\frac{\log\frac{HSA}{\delta}}{\sqrt{HSAD}}}\}$ *guarantees, with probability* $1 - \delta$*,*

$$R_K = O\left(H\sqrt{SAK \log\frac{HSA}{\delta}} + (HSA)^{1/4} \cdot H\sqrt{D \log\frac{HSA}{\delta}} + H^{3/2} d_{max} \log\frac{H}{\delta} \right).$$

C.1 The good event

Let $\widetilde{\mathcal{H}}^k$ be the history of episodes $\{j : j + d^j < k\}$. Define the following events:

$$E^c = \left\{ \sum_{k=1}^{K} \langle \mathbb{E}[\hat{c}^k \mid \widetilde{\mathcal{H}}^{k+d^k}] - \hat{c}^k, q^k \rangle \le 4H\sqrt{K \log \frac{10}{\delta}} \right\}$$

$$E^{\hat{c}} = \left\{ \sum_{k=1}^{K} \langle |q^k - q^{k+d^k}|, \hat{c}^k \rangle \le 4 \sum_{k=1}^{K} \langle |q^k - q^{k+d^k}|, c^k \rangle + \frac{40H \log \frac{10H}{\delta}}{\gamma} \right\}$$

$$E^d = \left\{ \sum_{k,h,s,a} |\mathcal{F}^{k+d^k}| \hat{c}_h^k(s,a) \le \sum_{k,h,s,a} |\mathcal{F}^{k+d^k}| c_h^k(s,a) + \frac{10H d_{max} \log \frac{10H}{\delta}}{\gamma} \right\}$$

$$E^{sq} = \left\{ \sum_{k=1}^{K} \sum_{i=1}^{K} \mathbb{I}\{k \le i + d^i < k + d^k\} \sum_{h,s,a} \sqrt{q_h^{i+d^i}(s,a)} (\hat{c}_h^i(s,a) - 4c_h^i(s,a)) \le \frac{10H d_{max} \log \frac{10H}{\delta}}{\gamma} \right\}$$

$$E^\star = \left\{ \sum_{k=1}^{K} \langle \hat{c}^k - c^k, q^\star \rangle \le \frac{H \log \frac{10HSA}{\delta}}{\gamma} \right\}$$

The good event is the intersection of the above events. The following lemma establishes that the good event holds with high probability.

Lemma 17 (The Good Event) *Let $\mathbb{G} = E^c \cap E^{\hat{c}} \cap E^d \cap E^{sq} \cap E^\star$ be the good event. It holds that $\Pr[\mathbb{G}] \ge 1 - \delta$.*

Proof We show that each of the events $\neg E^c, \neg E^{\hat{c}}, \neg E^d, \neg E^{sq}, \neg E^\star$ occur with probability at most $\delta/5$. Then, by a union bound we obtain the statement.

- $\Pr[\neg E^c] < \delta/5$ by Azuma inequality since it is a martingale with respect to the filtration $\{\tilde{\mathcal{H}}^{1+d^1}, \tilde{\mathcal{H}}^{2+d^2}, \dots\}$ where the differences are bounded by H.

- $\Pr[\neg E^{\hat{c}}] < \delta/5$ by (Cohen et al., 2021b, Lemma E.2) since $\langle |q^k - q^{k+d^k}|, \hat{c}^k \rangle \le H/\gamma$, and $\mathbb{E}[\langle |q^k - q^{k+d^k}|, \hat{c}^k \rangle \mid \widetilde{\mathcal{H}}^{i+d^i}] \le \langle |q^k - q^{k+d^k}|, c^k \rangle$.

- $\Pr[\neg E^d] < \delta/5$ by (Jin et al., 2020a, Lemma 11).

- $\Pr[\neg E^{sq}] < \delta/5$ by (Cohen et al., 2021b, Lemma E.2) in the following way.
  Denote $Y_i = \sum_k \mathbb{I}\{k \le i + d^i < k + d^k\} \sum_{h,s,a} \sqrt{q_h^{i+d^i}(s,a)} \hat{c}_h^i(s,a)$ and notice that $Y_i \le H d_{max}/\gamma$, and that:

$$\mathbb{E}[Y_i \mid \widetilde{\mathcal{H}}^{i+d^i}] \le \sum_k \mathbb{I}\{k \le i + d^i < k + d^k\} \sum_{h,s,a} \sqrt{q_h^{i+d^i}(s,a)} c_h^i(s,a).$$

- $\Pr[\neg E^\star] < \delta/5$ by (Jin et al., 2020a, Lemma 14).

■

C.2 Proof of the Main Theorem

Proof of Theorem 16 By Lemma 17, the good event holds with probability $1-\delta$. We now analyze the regret under the assumption that the good event holds. We decompose the regret as follows:

$$
\begin{aligned}
R_K &= \sum_{k=1}^{K} \langle q^k - q^\star, c^k \rangle \\
&= \underbrace{\sum_{k=1}^{K} \langle q^k, c^k - \hat{c}^k \rangle}_{\textsc{Bias}_1} + \underbrace{\sum_{k=1}^{K} \langle q^\star, \hat{c}^k - c^k \rangle}_{\textsc{Bias}_2} + \underbrace{\sum_{k=1}^{K} \langle q^k - q^{k+d^k}, \hat{c}^k \rangle}_{\textsc{Drift}} + \underbrace{\sum_{k=1}^{K} \langle q^{k+d^k} - q^\star, \hat{c}^k \rangle}_{\textsc{Reg}} . \qquad (28)
\end{aligned}
$$

$\textsc{Bias}_2$ is bounded under event $E^\star$ by $O(\frac{H \log \frac{HSA}{\delta}}{\gamma})$, $\textsc{Reg}$ is bounded in Lemma 18 by $O(\frac{H\log(HSA)}{\eta} + \eta HSAK + \frac{\eta}{\gamma} d_{max} \log \frac{H}{\delta})$, $\textsc{Drift}$ is bounded in Lemma 19 by $O(\eta\sqrt{H^3SA}(D+K) + \frac{\eta}{\gamma} H^{3/2} d_{max} \log \frac{H}{\delta} + \frac{H \log \frac{H}{\delta}}{\gamma})$, and $\textsc{Bias}_1$ is bounded in Lemma 20 by $O(H\sqrt{K \log \frac{1}{\delta}} + \gamma HSAK + \eta\sqrt{H^3SA}(D+K) + \frac{\eta}{\gamma} H^{3/2} d_{max} \log \frac{H}{\delta})$. Putting everything together:

$$
R_K = O\left(H\sqrt{K \log \frac{1}{\delta}} + (\eta+\gamma) HSAK + (\frac{1}{\eta} + \frac{1}{\gamma}) H \log \frac{HSA}{\delta} + \eta\sqrt{H^3SA}(D+K) + \frac{\eta}{\gamma} H^{3/2} d_{max} \log \frac{H}{\delta} \right),
$$

and plugging in the definitions of η and γ finishes the proof. ■

C.3 Bound on the Regret with respect to the Loss Estimators and Future Policies (REG in Equation (28))

Lemma 18 (REG Term) *Under the good event,*

$$
\sum_{k=1}^{K} \langle q^{k+d^k} - q^\star, \hat{c}^k \rangle = O\left(\frac{H \log(HSA)}{\eta} + \eta HSAK + \frac{\eta}{\gamma} H d_{max} \log \frac{H}{\delta} \right).
$$

Proof Let $\tilde{q}_h^{k+1}(s,a) = q_h^k(s,a) e^{-\eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a)}$. Taking the log,

$$
\eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a) = \log q_h^k(s,a) - \log \tilde{q}_h^{k+1}(s,a).
$$

Hence for any q

$$
\begin{aligned}
\eta \left\langle \sum_{j:j+d^j=k} \hat{c}_h^j, q^k - q^\star \right\rangle &= \langle \log q^k - \log \tilde{q}^{k+1}, q^k - q^\star \rangle = \mathrm{KL}(q^\star \parallel q^k) - \mathrm{KL}(q^\star \parallel \tilde{q}^{k+1}) + \mathrm{KL}(q^k \parallel \tilde{q}^{k+1}) \\
&\le \mathrm{KL}(q^\star \parallel q^k) - \mathrm{KL}(q^\star \parallel q^{k+1}) - \mathrm{KL}(q^{k+1} \parallel \tilde{q}^{k+1}) + \mathrm{KL}(q^k \parallel \tilde{q}^{k+1}) \\
&\le \mathrm{KL}(q^\star \parallel q^k) - \mathrm{KL}(q^\star \parallel q^{k+1}) + \mathrm{KL}(q^k \parallel \tilde{q}^{k+1}),
\end{aligned}
$$

where the second equality follows directly the definition of KL, the first inequality is by (Zimin, 2013, Lemma 1.2), and the second inequality is since the KL is non-negative. Now, the last term is bounded as follows:

$$
\begin{aligned}
\mathrm{KL}(q^k \parallel \tilde{q}^{k+1}) &\le \mathrm{KL}(q^k \parallel \tilde{q}^{k+1}) + \mathrm{KL}(\tilde{q}^{k+1} \parallel q^k) \\
&= \sum_h \sum_{s,a} \tilde{q}_h^{k+1}(s,a) \log \frac{\tilde{q}_h^{k+1}(s,a)}{q_h^k(s,a)} + \sum_h \sum_{s,a} q_h^k(s,a) \log \frac{q_h^k(s,a)}{\tilde{q}_h^{k+1}(s,a)} \\
&= \langle q^k - \tilde{q}^{k+1}, \log q^k - \log \tilde{q}^{k+1} \rangle = \eta \left\langle q^k - \tilde{q}^{k+1}, \sum_{j:j+d^j=k} \hat{c}^j \right\rangle.
\end{aligned}
$$

We get that

$$\eta \left\langle \sum_{j:j+d^j=k} \hat{c}^j, q^k - q^\star \right\rangle \le \mathrm{KL}(q^\star \parallel q^k) - \mathrm{KL}(q^\star \parallel q^{k+1}) + \eta \left\langle q^k - \tilde{q}^{k+1}, \sum_{j:j+d^j=k} \hat{c}^j \right\rangle .$$

Summing over k and dividing by η, we get

$$\begin{aligned}
\underbrace{\sum_{k=1}^K \sum_{j:j+d^j=k} \left\langle \hat{c}^j, q^k - q^\star \right\rangle}_{(*)} &\le \frac{\mathrm{KL}(q^\star \parallel q^1) - \mathrm{KL}(q^\star \parallel q^{K+1})}{\eta} + \sum_{k=1}^K \left\langle q^k - \tilde{q}^{k+1}, \sum_{j:j+d^j=k} \hat{c}^j \right\rangle \\
&\le \frac{\mathrm{KL}(q^\star \parallel q^1)}{\eta} + \sum_{k=1}^K \left\langle q^k - \tilde{q}^{k+1}, \sum_{j:j+d^j=k} \hat{c}^j \right\rangle \\
&\le \frac{2H\log(SA)}{\eta} + \underbrace{\sum_{k=1}^K \left\langle q^k - \tilde{q}^{k+1}, \sum_{j:j+d^j=k} \hat{c}^j \right\rangle}_{(**)},
\end{aligned}$$

where the last inequality is a standard argument (see (Zimin, 2013; Hazan, 2019)). We now rearrange $(*)$ and $(**)$:

$$\begin{aligned}
(*) &= \sum_{k=1}^K \sum_{j=1}^K \mathbb{I}\{j + d^j = k\} \langle \hat{c}^j, q^k - q^\star \rangle = \sum_{j=1}^K \sum_{k=1}^K \mathbb{I}\{j + d^j = k\} \langle \hat{c}^j, q^k - q^\star \rangle \\
&= \sum_{j=1}^K \langle \hat{c}^j, q^{j+d^j} - q^\star \rangle = \sum_{k=1}^K \langle \hat{c}^k, q^{k+d^k} - q^\star \rangle.
\end{aligned}$$

In a similar way,

$$\begin{aligned}
(**) &= \sum_{k=1}^K \sum_{j:j+d^j=k} \langle q^k - \tilde{q}^{k+1}, \hat{c}^j \rangle = \sum_{k=1}^K \sum_{j=1}^K \mathbb{I}\{j + d^j = k\} \langle q^k - \tilde{q}^{k+1}, \hat{c}^j \rangle \\
&= \sum_{j=1}^K \sum_{k=1}^K \mathbb{I}\{j + d^j = k\} \langle q^k - \tilde{q}^{k+1}, \hat{c}^j \rangle = \sum_{k=1}^K \langle q^{k+d^k} - \tilde{q}^{k+d^k+1}, \hat{c}^k \rangle.
\end{aligned}$$

This gives us,

$$\sum_{k=1}^K \langle \hat{c}^k, q^{k+d^k} - q^\star \rangle \le \frac{2H\log(SA)}{\eta} + \sum_{k=1}^K \langle q^{k+d^k} - \tilde{q}^{k+d^k+1}, \hat{c}^k \rangle.$$

It remains to bound the second term on the right hand side:

$$
\begin{aligned}
\sum_k \langle q^{k+d^k} - \tilde{q}^{k+d^k+1}, \hat{c}^k \rangle &= \sum_{k,h,s,a} \hat{c}_h^k(s,a)(q_h^{k+d^k}(s,a) - \tilde{q}_h^{k+d^k+1}(s,a)) \\
&= \sum_{k,h,s,a} \hat{c}_h^k(s,a) \left(q_h^{k+d^k}(s,a) - q_h^{k+d^k}(s,a) e^{-\eta \sum_{j:j+d^j=k+d^k} \hat{c}_h^j(s,a)} \right) \\
&= \sum_{k,h,s,a} q_h^{k+d^k}(s,a)\hat{c}_h^k(s,a) \left(1 - e^{-\eta \sum_{j:j+d^j=k+d^k} \hat{c}_h^j(s,a)} \right) \\
&\le \eta \sum_{k,h,s,a} q_h^{k+d^k}(s,a)\hat{c}_h^k(s,a) \left(\sum_{j:j+d^j=k+d^k} \hat{c}_h^j(s,a) \right) && (1-e^{-x} \le x) \\
&= \eta \sum_{k,h,s,a} q_h^{k+d^k}(s,a) \frac{\mathbb{I}\{s_h^k = s, a_h^k = a\} c_h^k(s,a)}{\max\{q_h^k(s,a), q_h^{k+d^k}(s,a)\} + \gamma} \left(\sum_{j:j+d^j=k+d^k} \hat{c}_h^j(s,a) \right) \\
&\le \eta \sum_{k,h,s,a} \sum_{j:j+d^j=k+d^k} \hat{c}_h^j(s,a) = \eta \sum_{k,h,s,a} \sum_j \mathbb{I}\{j+d^j = k+d^k\} \hat{c}_h^j(s,a) \\
&= \eta \sum_{j,h,s,a} \hat{c}_h^j(s,a) \sum_k \mathbb{I}\{j+d^j = k+d^k\} \le \eta \sum_{k,h,s,a} |\mathcal{F}^{k+d^k}| \hat{c}_h^k(s,a).
\end{aligned}
$$

Finally, by event E^d,

$$
\sum_{k,h,s,a} |\mathcal{F}^{k+d^k}| \hat{c}_h^k(s,a) = O\left(\sum_{k,h,s,a} |\mathcal{F}^{k+d^k}| c_h^k(s,a) + \frac{H d_{max} \log \frac{H}{\delta}}{\gamma} \right) = O\left(\eta HSAK + \frac{H d_{max} \log \frac{H}{\delta}}{\gamma} \right).
$$

■

C.4 Bound on the Delay-caused Drift (DRIFT in Equation (28))

Lemma 19 (DRIFT term) *Under the good event,*

$$
\sum_{k=1}^K \langle q^k - q^{k+d^k}, \hat{c}^k \rangle = O\left(\eta \sqrt{H^3 SA}(D+K) + \frac{\eta}{\gamma} H^{3/2} d_{max} \log \frac{H}{\delta} + \frac{H \log \frac{H}{\delta}}{\gamma} \right).
$$

Proof By event $E^{\hat{c}}$ we have:

$$
\sum_{k=1}^K \langle \hat{c}^k, q^k - q^{k+d^k} \rangle \le \sum_{k=1}^K \langle \hat{c}^k, |q^k - q^{k+d^k}| \rangle = O\left(\sum_{k=1}^K \langle c^k, |q^k - q^{k+d^k}| \rangle + \frac{H \log \frac{H}{\delta}}{\gamma} \right).
$$

Now, by Pinsker inequality and Jensen inequality:

$$\begin{aligned}
\sum_{k=1}^{K} \langle c^k, |q^k - q^{k+d^k}|\rangle &\le \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \sum_{h,s,a} |q_h^j(s,a) - q_h^{j+1}(s,a)| = \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \sum_{h} \|q_h^j - q_h^{j+1}\|_1 \\
&\le \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \sum_{h} \sqrt{2\mathrm{KL}(q_h^j \parallel q_h^{j+1})} \le \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \sqrt{2H \sum_{h} \mathrm{KL}(q_h^j \parallel q_h^{j+1})} \\
&\le \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \sqrt{H \sum_{h} \sum_{s,a} q_h^j(s,a) \Big(\eta \sum_{i:i+d^i=j} \hat{c}_h^i(s,a)\Big)^2} \\
&\le \eta\sqrt{H} \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \sum_{i:i+d^i=j} \sum_{h,s,a} \sqrt{q_h^j(s,a)}\hat{c}_h^i(s,a),
\end{aligned}$$

where the last inequality is by $\|x\|_2 \le \|x\|_1$, and the one before is by Lemma 21. Finally, we rearrange as follows:

$$\begin{aligned}
\sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \sum_{i:i+d^i=j} \sum_{h,s,a} \sqrt{q_h^j(s,a)}\hat{c}_h^i(s,a) &= \sum_{k,j,i} \mathbb{I}\{k \le j < k + d^k, i + d^i = j\} \sum_{h,s,a} \sqrt{q_h^j(s,a)}\hat{c}_h^i(s,a) \\
&= \sum_{k,j,i} \mathbb{I}\{k \le j < k + d^k, i + d^i = j\} \sum_{h,s,a} \sqrt{q_h^{i+d^i}(s,a)}\hat{c}_h^i(s,a) \\
&= \sum_{k,i} \mathbb{I}\{k \le i + d^i < k + d^k\} \sum_{h,s,a} \sqrt{q_h^{i+d^i}(s,a)}\hat{c}_h^i(s,a) \\
&= O\left(\sum_{k,i} \mathbb{I}\{k \le i + d^i < k + d^k\} \sum_{h,s,a} \sqrt{q_h^{i+d^i}(s,a)}c_h^i(s,a) + \frac{H d_{max} \log \frac{H}{\delta}}{\gamma}\right),
\end{aligned}$$

where the last relation is by event E^{sq}. To finish the proof we use Lemma 22:

$$\begin{aligned}
\sum_{k,i} \mathbb{I}\{k \le i + d^i < k + d^k\} \sum_{h,s,a} \sqrt{q_h^{i+d^i}(s,a)}c_h^i(s,a) &\le \sqrt{HSA} \sum_{k,i} \mathbb{I}\{k \le i + d^i < k + d^k\} \sqrt{\sum_{h,s,a} q_h^{i+d^i}(s,a)} \\
&= H\sqrt{SA} \sum_{k,i} \mathbb{I}\{k \le i + d^i < k + d^k\} \le H\sqrt{SA}(D + K).
\end{aligned}$$

■

C.5 Bound on the Bias of the Delay-adapted Estimator (Bias$_1$ in Equation (28))

Lemma 20 (Bias$_1$) *Under the good event,*

$$\sum_{k=1}^{K} \langle c^k - \hat{c}^k, q^k \rangle = O\left(H\sqrt{K \log \frac{1}{\delta}} + \gamma HSAK + \eta\sqrt{H^3SA}(D + K) + \frac{\eta}{\gamma} H^{3/2} d_{max} \log \frac{H}{\delta}\right).$$

Proof Decompose Bias$_1$ as follows:

$$\sum_{k=1}^{K} \langle c^k - \hat{c}^k, q^k \rangle = \sum_{k=1}^{K} \langle c^k - \mathbb{E}\Big[\hat{c}^k \mid \tilde{\mathcal{H}}^{k+d^k}\Big], q^k \rangle + \sum_{k=1}^{K} \langle \mathbb{E}\Big[\hat{c}^k \mid \tilde{\mathcal{H}}^{k+d^k}\Big] - \hat{c}^k, q^k \rangle.$$

The second term is bounded by $O(H\sqrt{K\log\frac{1}{\delta}})$ under event E^c. The first term is bounded as follows:

$$\begin{aligned}
\sum_{k=1}^{K}\langle c^k - \mathbb{E}[\hat{c}^k \mid \tilde{\mathcal{H}}^{k+d^k}], q^k\rangle &= \sum_{k,h,s,a} q_h^k(s,a)c_h^k(s,a)\left(1 - \frac{\mathbb{E}\left[\mathbb{I}\{s_h^k = s, a_h^k = a\} \mid \tilde{\mathcal{H}}^{k+d^k}\right]}{\max\{q_h^{k+d^k}(s,a), q_h^k(s,a)\} + \gamma}\right) \\
&= \sum_{k,h,s,a} q_h^k(s,a)c_h^k(s,a)\left(1 - \frac{q_h^k(s,a)}{\max\{q_h^{k+d^k}(s,a), q_h^k(s,a)\} + \gamma}\right) \\
&= \sum_{k,h,s,a} \frac{q_h^k(s,a)}{\max\{q_h^{k+d^k}(s,a), q_h^k(s,a)\} + \gamma}(\max\{q_h^{k+d^k}(s,a), q_h^k(s,a)\} - q_h^k(s,a) + \gamma) \\
&\le \sum_{k,h,s,a} (\max\{q_h^{k+d^k}(s,a), q_h^k(s,a)\} - q_h^k(s,a)) + \gamma HSAK \\
&\le \sum_{k,h,s,a} |q_h^{k+d^k}(s,a) - q_h^k(s,a)| + \gamma HSAK \\
&\le \eta\sqrt{H^3SA}(D+K) + \frac{\eta}{\gamma}H^{3/2}d_{\max} + \gamma HSAK.
\end{aligned}$$

where the first equality uses the fact that q^k and q^{k+d^k} are determined by the history $\tilde{\mathcal{H}}^{k+d^k}$, the second equality is since the k-th episode is not part of the history $\tilde{\mathcal{H}}^{k+d^k}$ as $k \notin \{j : j + d^j < k + d^k\}$, and the last inequality is as in the proof of Lemma 19. ■

C.6 Auxiliary lemmas

Lemma 21 $\sum_h KL(q_h^k \parallel q_h^{k+1}) \le \frac{\eta^2}{2}\sum_{h,s,a} q_h^k(s,a)(\sum_{j:j+d^j=k}\hat{c}_h^j(s,a))^2$.

Proof We start with expanding $\mathrm{KL}(q_h^k \parallel q_h^{k+1})$ as follows:

$$\begin{aligned}
\sum_h \mathrm{KL}(q_h^k \parallel q_h^{k+1}) &= \sum_{h,s,a} q_h^k(s,a)\log\frac{q_h^k(s,a)}{q_h^{k+1}(s,a)} = \sum_{h,s,a} q_h^k(s,a)\log\frac{Z_h^k(v^k)q_h^k(s,a)}{q_h^k(s,a)e^{B_h^k(s,a\mid v^k)}} \\
&= \sum_{h,s,a} q_h^k(s,a)\log Z_h^k(v^k) - \sum_{h,s,a} q_h^k(s,a)B_h^k(s,a \mid v^k) \\
&= \sum_h \log Z_h^k(v^k) - \sum_{h,s,a} q_h^k(s,a)B_h^k(s,a \mid v^k). \qquad (29)
\end{aligned}$$

For the first term in Equation (29), by definition of v^k and Z_h^k:

$$\begin{aligned}
\sum_h \log Z_h^k(v^k) &\le \sum_h \log Z_h^k(0) = \sum_h \log\left(\sum_{s,a} q_h^k(s,a) e^{B_h^k(s,a|0)}\right) = \sum_h \log\left(\sum_{s,a} q_h^k(s,a) e^{-\eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a)}\right) \\
&\le \sum_h \log\left(\sum_{s,a} q_h^k(s,a)\left(1 - \eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a) + \frac{1}{2}\left(\eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a)\right)^2\right)\right) \\
&= \sum_h \log\left(1 - \eta \sum_{s,a} \sum_{j:j+d^j=k} q_h^k(s,a)\hat{c}_h^j(s,a) + \frac{\eta^2}{2}\sum_{s,a} q_h^k(s,a)\left(\sum_{j:j+d^j=k} \hat{c}_h^j(s,a)\right)^2\right) \\
&\le \sum_h \left(-\eta \sum_{s,a} \sum_{j:j+d^j=k} q_h^k(s,a)\hat{c}_h^j(s,a) + \frac{\eta^2}{2}\sum_{s,a} q_h^k(s,a)\left(\sum_{j:j+d^j=k} \hat{c}_h^j(s,a)\right)^2\right) \\
&= -\eta \sum_{h,s,a} \sum_{j:j+d^j=k} q_h^k(s,a)\hat{c}_h^j(s,a) + \frac{\eta^2}{2}\sum_{h,s,a} q_h^k(s,a)\left(\sum_{j:j+d^j=k} \hat{c}_h^j(s,a)\right)^2,
\end{aligned}$$

where the second inequality is by $e^s \le 1 + s + s^2/2$ for $s \le 0$, and the third inequality is by $\log(1+s) \le s$ for all s. The second term in Equation (29) can be written as follows:

$$\begin{aligned}
\sum_{h,s,a} q_h^k(s,a) B_h^k(s,a \mid v^k) = \sum_{h,s,a} q_h^k(s,a) v_h^k(s) - \eta \sum_{h,s,a} \sum_{j:j+d^j=k} q_h^k(s,a)\hat{c}_h^j(s,a) \\
- \sum_{h,s,a,s'} q_h^k(s,a) p_h(s' \mid s,a) v_{h+1}^k(s').
\end{aligned}$$

So now, by occupancy measure constraints:

$$\sum_{h,s,a,s'} q_h^k(s,a) p_h(s' \mid s,a) v_{h+1}^k(s') = \sum_{h,s'} v_{h+1}^k(s') \sum_{s,a} q_h^k(s,a) p_h(s' \mid s,a) = \sum_{h,s',a'} q_{h+1}^k(s',a') v_{h+1}^k(s'),$$

which forms a telescopic sum, so by $v_0^k(s) = v_{H+1}^k(s) = 0$, we have:

$$\sum_{h,s,a} q_h^k(s,a) B_h^k(s,a \mid v^k) = -\eta \sum_{h,s,a} \sum_{j:j+d^j=k} q_h^k(s,a)\hat{c}_h^j(s,a).$$

■

Lemma 22 (Thune et al. (2019)) $\sum_{k=1}^K \sum_{i=1}^K \mathbb{I}\{k \le i + d^i < k + d^k\} \le D + K.$

Proof

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{i=1}^{K}\mathbb{I}\{k \le i+d^i < k+d^k\} &= \sum_{k=1}^{K}\sum_{i=1}^{K}\mathbb{I}\{k \le i+d^i < k+d^k\} \\
&= \sum_{k=1}^{K}\sum_{i=1}^{k}\mathbb{I}\{k \le i+d^i < k+d^k\} + \sum_{k=1}^{K}\sum_{i=k+1}^{K}\mathbb{I}\{k \le i+d^i < k+d^k\} \\
&= \sum_{k=1}^{K}\sum_{i=1}^{k}\mathbb{I}\{k \le i+d^i\} - \sum_{k=1}^{K}\sum_{i=1}^{k}\mathbb{I}\{k \le i+d^i, i+d^i \ge k+d^k\} + \sum_{k=1}^{K}\sum_{i=k+1}^{K}\mathbb{I}\{k \le i+d^i < k+d^k\} \\
&= \sum_{k=1}^{K}\sum_{i=1}^{K}\mathbb{I}\{i \le k \le i+d^i\} - \sum_{k=1}^{K}\sum_{i=1}^{k}\mathbb{I}\{k+d^k \le i+d^i\} + \sum_{k=1}^{K}\sum_{i=1}^{K}\mathbb{I}\{i \ge k+1, k \le i+d^i < k+d^k\} \\
&= \sum_{i=1}^{K}\sum_{k=1}^{K}\underbrace{\mathbb{I}\{i \le k \le i+d^i\}}_{=d^i+1} - \sum_{k=1}^{K}\sum_{i=1}^{K}\mathbb{I}\{i \le k, k+d^k \le i+d^i\} + \sum_{k=1}^{K}\sum_{i=1}^{K}\mathbb{I}\{i \ge k+1, k \le i+d^i < k+d^k\} \\
&\le D+K - \sum_{k=1}^{K}\sum_{i=1}^{K}\mathbb{I}\{i \le k, k+d^k \le i+d^i\} + \sum_{k=1}^{K}\sum_{i=1}^{K}\mathbb{I}\{k \le i, i+d^i \le k+d^k\} \le D+K.
\end{aligned}
$$

■

Algorithm 8 Delayed UOB-REPS with delay-adapted estimator

Input: State space $\mathcal{S}$, Action space $\mathcal{A}$, Horizon H, Number of episodes K, Learning rate $\eta > 0$, Exploration parameter $\gamma > 0$, Confidence parameter $\delta > 0$.
Initialization: Set $\pi_h^1(a \mid s) = \frac{1}{A}$, $q_h^1(s,a,s') = \frac{1}{S^2 A}$, $n_h^1(s,a) = 0$, $n_h^1(s,a,s')$ for every $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$.
for $k = 1, 2, ..., K$ **do**
  Play episode k with policy π^k and observe trajectory $\{(s_h^k, a_h^k)\}_{h=1}^H$.
  Update confidence set $\mathcal{P}^{k+1}$ by Algorithm 5.
  for $j : j + d^j = k$ **do**
    Observe feedback $\{c_h^j(s_h^j, a_h^j)\}_{h=1}^H$.
    Compute $u_h^j(s,a) = \max_{p' \in \mathcal{P}^j} q_h^{p',\pi^j}(s,a)$ and $u_h^k(s,a) = \max_{p' \in \mathcal{P}^k} q_h^{p',\pi^k}(s,a)$.
    Compute loss estimator $\hat{c}_h^j(s,a) = \frac{c_h^j(s,a)\mathbb{I}\{s_h^j = s, a_h^j = a\}}{\max\{u_h^j(s,a), u_h^k(s,a)\} + \gamma}$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
  end for
  Update occupancy measure:

$$q^{k+1} = \arg \min_{q \in \Delta(\mathcal{M},k+1)} \eta \left\langle q, \sum_{j:j+d^j=k} \hat{c}^j \right\rangle + \mathrm{KL}(q \parallel q^k), \tag{30}$$

  where $\mathrm{KL}(q \parallel q') = \sum_{h,s,a,s'} q_h(s,a,s') \ln \frac{q_h(s,a,s')}{q'_h(s,a,s')} + q'_h(s,a,s') - q_h(s,a,s')$ and $\Delta(\mathcal{M}, k+1) = \{q^{\pi,p'} \mid \pi \in (\Delta_\mathcal{A})^{\mathcal{S}\times[H]}, p' \in \mathcal{P}^{k+1}\}$.
  Update policy: $\pi_h^{k+1}(a \mid s) = \frac{\sum_{s'} q_h^{k+1}(s,a,s')}{\sum_{a'}\sum_{s'} q_h^{k+1}(s,a',s')}$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
end for

Appendix D. Delayed UOB-REPS with delay-adapted estimator

Explicitly solving this optimization problem in Equation (30), we get (Rosenberg and Mansour, 2019a):

$$q_h^{k+1}(s,a,s') = \frac{q_h^k(s,a,s') e^{B_h^k(s,a,s' \mid v^{\mu^k}, e^{\mu^k,\beta^k})}}{Z_h^k(v^{\mu^k}, e^{\mu^k,\beta^k})},$$

for:

$$\begin{aligned}
B_h^k(s,a,s' \mid v,e) &= e_h(s,a,s') + v_h(s,a,s') - \eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a) - \sum_{s''} \bar{p}_h^k(s'' \mid s,a) v_{h+1}(s,a,s'') \\
v_h^\mu(s,a,s') &= \mu_h^-(s,a,s') - \mu_h^+(s,a,s') \\
e_h^{\mu,\beta}(s,a,s') &= \beta_{h+1}(s') - \beta_h(s) + \sum_{s''} (\mu_h^-(s,a,s'') + \mu_h^+(s,a,s''))\epsilon_h^k(s'' \mid s,a) \\
\epsilon_h^k(s' \mid s,a) &= \sqrt{\frac{16\bar{p}_h^k(s'|s,a)\log\frac{10HSAK}{\delta}}{n_h^k(s,a) \vee 1}} + \frac{10\log\frac{10HSAK}{\delta}}{n_h^k(s,a) \vee 1} \\
Z_h^k(v,e) &= \sum_{s,a,s'} q_h^k(s,a,s') e^{B_h^k(s,a,s' \mid v,e)} \\
\mu^k, \beta^k &= \arg\min_{\beta,\mu \geq 0} \sum_{h=1}^H \log Z_h^k(v^\mu, e^{\mu,\beta}).
\end{aligned}$$

Theorem 23 *Running UOB-REPS with the delay-adapted estimator,* $\eta = \gamma = \min\{\sqrt{\frac{\log\frac{KHSA}{\delta}}{SAK}}, \sqrt{\frac{\log\frac{KHSA}{\delta}}{\sqrt{HSAD}}}\}$ *guarantees, with probability* $1-\delta$,

$$R_K = O\biggl(H^2 S\sqrt{AK\log\frac{KHSA}{\delta}} + (HSA)^{1/4}\cdot H\sqrt{D\log\frac{KHSA}{\delta}} + H^{3/2}d_{max}\log\frac{KHSA}{\delta} + H^3 S^3 A\log^3\frac{KHSA}{\delta}\biggr).$$

D.1 The good event

Let $\tilde{\mathcal{H}}^k$ be the history of episodes $\{j : j + d^j < k\}$ and $\iota = \log\frac{HSAK}{\delta}$. Define the following events:

$$E^p = \left\{\exists k, s', s, a, h : \left|p_h(s' \mid s,a) - \bar{p}_h^k(s' \mid s,a)\right| \le 4\sqrt{\frac{\bar{p}_h^k(s' \mid s,a)\log\frac{10HSAK}{\delta}}{n_h^k(s,a)\vee 1}} + 10\frac{\log\frac{10HSAK}{\delta}}{n_h^k(s,a)\vee 1}\right\}$$

$$E^c = \left\{\sum_{k=1}^K \langle \mathbb{E}[\hat{c}^k \mid \widetilde{\mathcal{H}}^{k+d^k}] - \hat{c}^k, q^k\rangle \le 4H\sqrt{K\log\frac{10}{\delta}}\right\}$$

$$E^{\hat{c}} = \left\{\sum_{k=1}^K \langle |q^k - q^{k+d^k}|, \hat{c}^k\rangle \le 4\sum_{k=1}^K \langle |q^k - q^{k+d^k}|, c^k\rangle + \frac{40H\log\frac{10H}{\delta}}{\gamma}\right\}$$

$$E^d = \left\{\sum_{k,h,s,a} |\mathcal{F}^{k+d^k}|\hat{c}_h^k(s,a) \le \sum_{k,h,s,a} |\mathcal{F}^{k+d^k}| c_h^k(s,a) + \frac{10Hd_{max}\log\frac{10H}{\delta}}{\gamma}\right\}$$

$$E^{sq} = \left\{\sum_{k=1}^K\sum_{i=1}^K \mathbb{I}\{k \le i + d^i < k + d^k\}\sum_{h,s,a}\sqrt{q_h^{i+d^i}(s,a)}(\hat{c}_h^i(s,a) - 4c_h^i(s,a)) \le \frac{10Hd_{max}\log\frac{10H}{\delta}}{\gamma}\right\}$$

$$E^\star = \left\{\sum_{k=1}^K \langle \hat{c}^k - c^k, q^\star\rangle \le \frac{H\log\frac{10HSA}{\delta}}{\gamma}\right\}$$

The good event is the intersection of the above events. The following lemma establishes that the good event holds with high probability.

Lemma 24 (The Good Event) *Let* $\mathbb{G} = E^p \cap E^c \cap E^{\hat{c}} \cap E^d \cap E^{sq} \cap E^\star$ *be the good event. It holds that* $\Pr[\mathbb{G}] \ge 1-\delta$.

Proof Similar to the proof of Lemma 17. Event E^p is standard (see, e.g., Jin et al. (2020a)). ■

D.2 Proof of the Main Theorem

Proof of Theorem 23 By Lemma 24, the good event holds with probability $1-\delta$. We now analyze the regret under the assumption that the good event holds. We decompose the regret as follows:

$$\begin{aligned} R_K &= \sum_{k=1}^K \langle q^{\pi^k} - q, c^k\rangle \\ &= \underbrace{\sum_{k=1}^K \langle q^{\pi^k} - q^k, c^k\rangle}_{\textsc{Est}} + \underbrace{\sum_{k=1}^K \langle q^k, c^k - \hat{c}^k\rangle}_{\textsc{Bias}_1} + \underbrace{\sum_{k=1}^K \langle q^\star, \hat{c}^k - c^k\rangle}_{\textsc{Bias}_2} + \underbrace{\sum_{k=1}^K \langle q^k - q^{k+d^k}, \hat{c}^k\rangle}_{\textsc{Drift}} + \underbrace{\sum_{k=1}^K \langle q^{k+d^k} - q^\star, \hat{c}^k\rangle}_{\textsc{Reg}}. \end{aligned} \quad (31)$$

$\textsc{Bias}_2$ is bounded under event $E^\star$ by $O(\frac{H\iota}{\gamma})$, $\textsc{Est}$ is bounded in Lemma 25 by $O(H^2S\sqrt{AK\iota}+H^2S^2A\iota^2)$, $\textsc{Reg}$ is bounded in Lemma 26 by $O(\frac{H\iota}{\eta}+\eta HSAK+\frac{\eta}{\gamma}Hd_{max}\iota)$, $\textsc{Drift}$ is bounded in Lemma 27 by $O(\eta\sqrt{H^3SA}(D+K)+\frac{\eta}{\gamma}H^{3/2}d_{max}\iota+\frac{H\iota}{\gamma})$, and $\textsc{Bias}_1$ is bounded in Lemma 28 by $O(H^2S\sqrt{AK\iota}+H^3S^3A\iota^3+\gamma HSAK+\eta\sqrt{H^3SA}(D+K)+\frac{\eta}{\gamma}H^{3/2}d_{max}\iota)$. Putting everything together:

$$R_K = O\left(H^2S\sqrt{AK\iota}+H^3S^3A\iota^3+(\eta+\gamma)HSAK+(\frac{1}{\eta}+\frac{1}{\gamma})H\iota+\eta\sqrt{H^3SA}(D+K)+\frac{\eta}{\gamma}H^{3/2}d_{max}\iota\right),$$

and plugging in the definitions of η and γ finishes the proof. ■

D.3 Bound on the Transition Estimation Error ($\textsc{Est}$ in Equation (31))

Lemma 25 ($\textsc{Est}$ Term) *Under the good event,*

$$\sum_{k=1}^K \langle q^{\pi^k}-q^k,c^k\rangle = O\left(H^2S\sqrt{AK\iota}+H^2S^2A\iota^2\right).$$

Proof Let $q^k=q^{\pi^k,p^k}$. By the value difference lemma (Shani et al., 2020):

$$\begin{aligned}
\sum_{k=1}^K \langle q^{\pi^k}-q^k,c^k\rangle &= \sum_{k,h,s,a} q_h^{\pi^k}(s,a)\sum_{s'}\left(p_h^k(s'\mid s,a)-p_h(s'\mid s,a)\right)V_{h+1}^{\pi^k,p^k}(s') \\
&\le H\sum_{k,h,s,a} q_h^{\pi^k}(s,a)\|p_h^k(\cdot\mid s,a)-p_h(\cdot\mid s,a)\|_1 \\
&= O\left(H\sqrt{S\iota}\sum_{k,h,s,a}\frac{q_h^{\pi^k}(s,a)}{\sqrt{n_h^k(s,a)\vee 1}}+HS\iota\sum_{k,h,s,a}\frac{q_h^{\pi^k}(s,a)}{n_h^k(s,a)\vee 1}\right) = O(H^2S\sqrt{AK\iota}+H^2S^2A\iota^2),
\end{aligned}$$

where the seconed inequality is by event E^p and the last relation is by standard analysis (Jin et al., 2020a, Lemma 10).
■

D.4 Bound on the Regret with respect to the Loss Estimators and Future Policies ($\textsc{Reg}$ in Equation (31))

Lemma 26 ($\textsc{Reg}$ Term) *Under the good event,*

$$\sum_{k=1}^K \langle q^{k+d^k}-q^\star,\hat{c}^k\rangle = O\left(\frac{H\iota}{\eta}+\eta HSAK+\frac{\eta}{\gamma}Hd_{max}\iota\right).$$

Proof Let $\tilde{q}_h^{k+1}(s,a,s')=q_h^k(s,a,s')e^{-\eta\sum_{j:j+d^j=k}\hat{c}_h^j(s,a)}$. Taking the log,

$$\eta\sum_{j:j+d^j=k}\hat{c}_h^j(s,a)=\log q_h^k(s,a,s')-\log\tilde{q}_h^{k+1}(s,a,s').$$

Hence,

$$\begin{aligned}
\eta\left\langle\sum_{j:j+d^j=k}\hat{c}_h^j,q^k-q\right\rangle &= \left\langle\log q^k-\log\tilde{q}^{k+1},q^k-q^\star\right\rangle = \mathrm{KL}(q^\star\parallel q^k)-\mathrm{KL}(q^\star\parallel\tilde{q}^{k+1})+\mathrm{KL}(q^k\parallel\tilde{q}^{k+1}) \\
&\le \mathrm{KL}(q^\star\parallel q^k)-\mathrm{KL}(q^\star\parallel q^{k+1})-\mathrm{KL}(q^{k+1}\parallel\tilde{q}^{k+1})+\mathrm{KL}(q^k\parallel\tilde{q}^{k+1}) \\
&\le \mathrm{KL}(q^\star\parallel q^k)-\mathrm{KL}(q^\star\parallel q^{k+1})+\mathrm{KL}(q^k\parallel\tilde{q}^{k+1}),
\end{aligned}$$

where the second equality follows directly the definition of KL, the first inequality is by (Zimin, 2013, Lemma 1.2), and the second inequality is since the KL is non-negative. Now, the last term is bounded as follows:

$$
\begin{aligned}
\mathrm{KL}(q^k \parallel \tilde{q}^{k+1}) &\le \mathrm{KL}(q^k \parallel \tilde{q}^{k+1}) + \mathrm{KL}(\tilde{q}^{k+1} \parallel q^k) \\
&= \sum_h \sum_{s,a,s'} \tilde{q}_h^{k+1}(s,a,s') \log \frac{\tilde{q}_h^{k+1}(s,a,s')}{q_h^k(s,a,s')} + \sum_h \sum_{s,a,s'} q_h^k(s,a,s') \log \frac{q_h^k(s,a,s')}{\tilde{q}_h^{k+1}(s,a,s')} \\
&= \langle q^k - \tilde{q}^{k+1}, \log q^k - \log \tilde{q}^{k+1} \rangle = \eta \Big\langle q^k - \tilde{q}^{k+1}, \sum_{j: j+d^j = k} \hat{c}^j \Big\rangle.
\end{aligned}
$$

We get that

$$
\eta \left\langle \sum_{j:j+d^j=k} \hat{c}^j, q^k - q^\star \right\rangle \le \mathrm{KL}(q^\star \parallel q^k) - \mathrm{KL}(q^\star \parallel q^{k+1}) + \eta \left\langle q^k - \tilde{q}^{k+1}, \sum_{j:j+d^j=k} \hat{c}^j \right\rangle.
$$

Summing over k and dividing by η, we get

$$
\begin{aligned}
\underbrace{\sum_{k=1}^K \sum_{j:j+d^j=k} \langle \hat{c}^j, q^k - q^\star \rangle}_{(*)} &\le \frac{\mathrm{KL}(q^\star \parallel q^1) - \mathrm{KL}(q^\star \parallel q^{K+1})}{\eta} + \sum_{k=1}^K \left\langle q^k - \tilde{q}^{k+1}, \sum_{j:j+d^j=k} \hat{c}^j \right\rangle \\
&\le \frac{\mathrm{KL}(q^\star \parallel q^1)}{\eta} + \sum_{k=1}^K \left\langle q^k - \tilde{q}^{k+1}, \sum_{j:j+d^j=k} \hat{c}^j \right\rangle \\
&\le \frac{4H \log(SA)}{\eta} + \underbrace{\sum_{k=1}^K \left\langle q^k - \tilde{q}^{k+1}, \sum_{j:j+d^j=k} \hat{c}^j \right\rangle}_{(**)},
\end{aligned}
$$

where the last inequality is a standard argument (see (Zimin, 2013; Hazan, 2019)). We now rearrange $(*)$ and $(**)$:

$$
\begin{aligned}
(*) &= \sum_{k=1}^K \sum_{j=1}^K \mathbb{I}\{j + d^j = k\} \langle \hat{c}^j, q^k - q^\star \rangle = \sum_{j=1}^K \sum_{k=1}^K \mathbb{I}\{j + d^j = k\} \langle \hat{c}^j, q^k - q^\star \rangle \\
&= \sum_{j=1}^K \langle \hat{c}^j, q^{j+d^j} - q^\star \rangle = \sum_{k=1}^K \langle \hat{c}^k, q^{k+d^k} - q^\star \rangle.
\end{aligned}
$$

In a similar way,

$$
\begin{aligned}
(**) &= \sum_{k=1}^K \sum_{j:j+d^j=k} \langle q^k - \tilde{q}^{k+1}, \hat{c}^j \rangle = \sum_{k=1}^K \sum_{j=1}^K \mathbb{I}\{j + d^j = k\} \langle q^k - \tilde{q}^{k+1}, \hat{c}^j \rangle \\
&= \sum_{j=1}^K \sum_{k=1}^K \mathbb{I}\{j + d^j = k\} \langle q^k - \tilde{q}^{k+1}, \hat{c}^j \rangle = \sum_{k=1}^K \langle q^{k+d^k} - \tilde{q}^{k+d^k+1}, \hat{c}^k \rangle.
\end{aligned}
$$

This gives us,

$$
\sum_{k=1}^K \langle \hat{c}^k, q^{k+d^k} - q^\star \rangle \le \frac{4H \log(SA)}{\eta} + \sum_{k=1}^K \langle q^{k+d^k} - \tilde{q}^{k+d^k+1}, \hat{c}^k \rangle.
$$

It remains to bound the second term on the right hand side:

$$
\begin{aligned}
\sum_k \langle q^{k+d^k} - \tilde{q}^{k+d^k+1}, \hat{c}^k \rangle &= \sum_{k,h,s,a,s'} \hat{c}_h^k(s,a)(q_h^{k+d^k}(s,a,s') - \tilde{q}_h^{k+d^k+1}(s,a,s')) \\
&= \sum_{k,h,s,a,s'} \hat{c}_h^k(s,a) \left(q_h^{k+d^k}(s,a,s') - q_h^{k+d^k}(s,a,s') e^{-\eta \sum_{j:j+d^j=k+d^k} \hat{c}_h^j(s,a)} \right) \\
&= \sum_{k,h,s,a,s'} q_h^{k+d^k}(s,a,s') \hat{c}_h^k(s,a) \left(1 - e^{-\eta \sum_{j:j+d^j=k+d^k} \hat{c}_h^j(s,a)} \right) \\
&\le \eta \sum_{k,h,s,a} q_h^{k+d^k}(s,a) \hat{c}_h^k(s,a) \left(\sum_{j:j+d^j=k+d^k} \hat{c}_h^j(s,a) \right) \qquad (1 - e^{-x} \le x) \\
&= \eta \sum_{k,h,s,a} q_h^{k+d^k}(s,a) \frac{\mathbb{I}\{s_h^k = s, a_h^k = a\} c_h^k(s,a)}{\max\{u_h^k(s,a), u_h^{k+d^k}(s,a)\} + \gamma} \left(\sum_{j:j+d^j=k+d^k} \hat{c}_h^j(s,a) \right) \\
&\le \eta \sum_{k,h,s,a} \sum_{j:j+d^j=k+d^k} \hat{c}_h^j(s,a) = \eta \sum_{k,h,s,a} \sum_j \mathbb{I}\{j + d^j = k + d^k\} \hat{c}_h^j(s,a) \\
&= \eta \sum_{j,h,s,a} \hat{c}_h^j(s,a) \sum_k \mathbb{I}\{j + d^j = k + d^k\} \le \eta \sum_{k,h,s,a} |\mathcal{F}^{k+d^k}| \hat{c}_h^k(s,a),
\end{aligned}
$$

where the second inequality is since $u_h^{k+d^k}(s,a) \ge q_h^{k+d^k}(s,a)$ under the good event. Finally, by event E^d,

$$
\sum_{k,h,s,a} |\mathcal{F}^{k+d^k}| \hat{c}_h^k(s,a) = O\left(\sum_{k,h,s,a} |\mathcal{F}^{k+d^k}| c_h^k(s,a) + \frac{H d_{max} \iota}{\gamma} \right) = O\left(\eta HSAK + \frac{H d_{max} \iota}{\gamma} \right).
$$

■

D.5 Bound on the Delay-caused Drift (Drift in Equation (31))

Lemma 27 (Drift term) *Under the good event,*

$$
\sum_{k=1}^K \langle q^k - q^{k+d^k}, \hat{c}^k \rangle = O\left(\eta \sqrt{H^3 SA}(D + K) + \frac{\eta}{\gamma} H^{3/2} d_{max} \iota + \frac{H\iota}{\gamma} \right).
$$

Proof By event $E^{\hat{c}}$ we have:

$$
\sum_{k=1}^K \langle \hat{c}^k, q^k - q^{k+d^k} \rangle \le \sum_{k=1}^K \langle \hat{c}^k, |q^k - q^{k+d^k}| \rangle = O\left(\sum_{k=1}^K \langle c^k, |q^k - q^{k+d^k}| \rangle + \frac{H\iota}{\gamma} \right).
$$

Now, by Pinsker inequality and Jensen inequality:

$$
\begin{aligned}
\sum_{k=1}^{K}\langle c^k, |q^k - q^{k+d^k}|\rangle &\le \sum_{k=1}^{K}\sum_{j=k}^{k+d^k-1}\sum_{h,s,a,s'}|q_h^j(s,a,s') - q_h^{j+1}(s,a,s')| = \sum_{k=1}^{K}\sum_{j=k}^{k+d^k-1}\sum_{h}\|q_h^j - q_h^{j+1}\|_1 \\
&\le \sum_{k=1}^{K}\sum_{j=k}^{k+d^k-1}\sum_{h}\sqrt{2\mathrm{KL}(q_h^j \parallel q_h^{j+1})} \le \sum_{k=1}^{K}\sum_{j=k}^{k+d^k-1}\sqrt{2H\sum_{h}\mathrm{KL}(q_h^j \parallel q_h^{j+1})} \\
&\le \sum_{k=1}^{K}\sum_{j=k}^{k+d^k-1}\sqrt{H\sum_{h}\sum_{s,a,s'}q_h^j(s,a,s')\Big(\eta\sum_{i:i+d^i=j}\hat{c}_h^i(s,a)\Big)^2} \\
&\le \eta\sqrt{H}\sum_{k=1}^{K}\sum_{j=k}^{k+d^k-1}\sum_{i:i+d^i=j}\sum_{h,s,a}\sqrt{q_h^j(s,a)}\hat{c}_h^i(s,a),
\end{aligned}
$$

where the last inequality is by $\|x\|_2 \le \|x\|_1$, and the one before is by Lemma 29. Finally, we rearrange as follows:

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{j=k}^{k+d^k-1}\sum_{i:i+d^i=j}\sum_{h,s,a}\sqrt{q_h^j(s,a)}\hat{c}_h^i(s,a) &= \sum_{k,j,i}\mathbb{I}\{k \le j < k+d^k, i+d^i = j\}\sum_{h,s,a}\sqrt{q_h^j(s,a)}\hat{c}_h^i(s,a) \\
&= \sum_{k,j,i}\mathbb{I}\{k \le j < k+d^k, i+d^i = j\}\sum_{h,s,a}\sqrt{q_h^{i+d^i}(s,a)}\hat{c}_h^i(s,a) \\
&= \sum_{k,i}\mathbb{I}\{k \le i+d^i < k+d^k\}\sum_{h,s,a}\sqrt{q_h^{i+d^i}(s,a)}\hat{c}_h^i(s,a) \\
&= O\left(\sum_{k,i}\mathbb{I}\{k \le i+d^i < k+d^k\}\sum_{h,s,a}\sqrt{q_h^{i+d^i}(s,a)}c_h^i(s,a) + \frac{Hd_{max}\iota}{\gamma}\right),
\end{aligned}
$$

where the last relation is by event E^{sq}. To finish the proof we use Lemma 22:

$$
\begin{aligned}
\sum_{k,i}\mathbb{I}\{k \le i+d^i < k+d^k\}\sum_{h,s,a}\sqrt{q_h^{i+d^i}(s,a)}c_h^i(s,a) &\le \sqrt{HSA}\sum_{k,i}\mathbb{I}\{k \le i+d^i < k+d^k\}\sqrt{\sum_{h,s,a}q_h^{i+d^i}(s,a)} \\
&= H\sqrt{SA}\sum_{k,i}\mathbb{I}\{k \le i+d^i < k+d^k\} \le H\sqrt{SA}(D+K).
\end{aligned}
$$

■

D.6 Bound on the Bias of the Delay-adapted Estimator (BIAS$_1$ in Equation (31))

Lemma 28 (BIAS$_1$ Term) *Under the good event,*

$$
\sum_{k=1}^{K}\langle c^k - \hat{c}^k, q^k\rangle = O\left(H^2S\sqrt{AK\iota} + H^3S^3A\iota^3 + \gamma HSAK + \eta\sqrt{H^3SA}(D+K) + \frac{\eta}{\gamma}H^{3/2}d_{max}\iota\right).
$$

Proof Decompose BIAS$_1$ as follows:

$$
\sum_{k=1}^{K}\langle c^k - \hat{c}^k, q^k\rangle = \sum_{k=1}^{K}\langle c^k - \mathbb{E}\Big[\hat{c}^k \mid \tilde{\mathcal{H}}^{k+d^k}\Big], q^k\rangle + \sum_{k=1}^{K}\langle \mathbb{E}\Big[\hat{c}^k \mid \tilde{\mathcal{H}}^{k+d^k}\Big] - \hat{c}^k, q^k\rangle.
$$

The second term is bounded by $O(H\sqrt{K\iota})$ under event E^c. The first term is bounded as follows:

$$\sum_{k=1}^{K}\langle c^k - \mathbb{E}[\hat{c}^k \mid \tilde{\mathcal{H}}^{k+d^k}], q^k\rangle = \sum_{k,h,s,a,s'} q_h^k(s,a,s')c_h^k(s,a)\left(1 - \frac{\mathbb{E}\left[\mathbb{I}\{s_h^k = s, a_h^k = a\} \mid \tilde{\mathcal{H}}^{k+d^k}\right]}{\max\{u_h^k(s,a), u_h^{k+d^k}(s,a)\} + \gamma}\right)$$

$$= \sum_{k,h,s,a} q_h^k(s,a)c_h^k(s,a)\left(1 - \frac{q_h^{\pi^k}(s,a)}{\max\{u_h^k(s,a), u_h^{k+d^k}(s,a)\} + \gamma}\right)$$

$$= \sum_{k,h,s,a} \frac{q_h^k(s,a)}{\max\{u_h^k(s,a), u_h^{k+d^k}(s,a)\} + \gamma}(\max\{u_h^k(s,a), u_h^{k+d^k}(s,a)\} - q_h^{\pi^k}(s,a) + \gamma)$$

$$\le \sum_{k,h,s,a} (\max\{u_h^k(s,a), u_h^{k+d^k}(s,a)\} - q_h^{\pi^k}(s,a)) + \gamma HSAK$$

$$\le \sum_{k,h,s,a} |\max\{u_h^k(s,a), u_h^{k+d^k}(s,a)\} - q_h^{\pi^k}(s,a)| + \gamma HSAK.$$

where the first equality uses the fact that u^k and u^{k+d^k} is determined by the history $\tilde{\mathcal{H}}^{k+d^k}$, the second equality is since the k-th episode is not part of the history $\tilde{\mathcal{H}}^{k+d^k}$ as $k \notin \{j : j + d^j < k + d^k\}$, and the first inequality is since $u_h^k(s,a) \ge q_h^k(s,a)$ under the good event. Finally, we bound:

$$\sum_{k,h,s,a} |\max\{u_h^k(s,a), u_h^{k+d^k}(s,a)\} - q_h^{\pi^k}(s,a)| \le \sum_{k,h,s,a} |u_h^k(s,a) - q_h^{\pi^k}(s,a)| + \sum_{k,h,s,a} |u_h^{k+d^k}(s,a) - q_h^{\pi^k}(s,a)|.$$

The first term is bounded in (Jin et al., 2020a, Lemma 4) by $O(H^2S\sqrt{AK\iota} + H^3S^3A\iota^3)$, and for the second term:

$$\sum_{k,h,s,a} |u_h^{k+d^k}(s,a) - q_h^{\pi^k}(s,a)| \le \sum_{k,h,s,a} |u_h^{k+d^k}(s,a) - q_h^{\pi^{k+d^k}}(s,a)| + \sum_{k,h,s,a} |q_h^{\pi^{k+d^k}}(s,a) - q_h^{\pi^k}(s,a)|,$$

where again the first term is bounded in (Jin et al., 2020a, Lemma 4). Finally,

$$\sum_{k,h,s,a} |q_h^{\pi^{k+d^k}}(s,a) - q_h^{\pi^k}(s,a)| \le \sum_{k,h,s,a} |q_h^{\pi^{k+d^k}}(s,a) - q_h^{k+d^k}(s,a)| + \sum_{k,h,s,a} |q_h^k(s,a) - q_h^{\pi^k}(s,a)| + \sum_{k,h,s,a} |q_h^{k+d^k}(s,a) - q_h^k(s,a)|,$$

where the first two terms are bounded similarly to Lemma 25 and the last term is bounded similarly to Lemma 27. ■

D.7 Auxiliary lemmas

Lemma 29 $\sum_h KL(q_h^k \parallel q_h^{k+1}) \le \frac{\eta^2}{2}\sum_{h,s,a,s'} q_h^k(s,a,s')(\sum_{j:j+d^j=k}\hat{c}_h^j(s,a))^2.$

Proof We start with expanding $\mathrm{KL}(q_h^k \parallel q_h^{k+1})$ as follows:

$$\sum_h \mathrm{KL}(q_h^k \parallel q_h^{k+1}) = \sum_h\sum_{s,a,s'} q_h^k(s,a,s')\log\frac{q_h^k(s,a,s')}{q_h^{k+1}(s,a,s')} = \sum_h\sum_{s,a,s'} q_h^k(s,a,s')\log\frac{Z_h^k(v^{\mu^k}, e^{\mu^k,\beta^k})}{e^{B_h^k(s,a,s'|v^{\mu^k}, e^{\mu^k,\beta^k})}}$$

$$= \underbrace{\sum_h \log Z_h^k(v^{\mu^k}, e^{\mu^k,\beta^k})}_{(A)} - \underbrace{\sum_h\sum_{s,a,s'} q_h^k(s,a,s')B_h^k(s,a,s' \mid v^{\mu^k}, e^{\mu^k,\beta^k})}_{(B)}.$$

By definition of μ^k, β^k, term (A) can be bounded by

$$
\begin{aligned}
(A) &\leq \sum_h \log Z_h^k(0,0) = \sum_h \log(\sum_{s,a,s'} q_h^k(s,a,s') e^{B_h^k(s,a,s'|0,0)}) = \sum_h \log(\sum_{s,a,s'} q_h^k(s,a,s') e^{-\eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a)}) \\
&\leq \sum_h \log \left(\sum_{s,a,s'} q_h^k(s,a,s') \left(1 - \eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a) + \frac{(\eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a))^2}{2} \right) \right) \\
&= \sum_h \log \left(1 - \eta \sum_{s,a,s'} \sum_{j:j+d^j=k} q_h^k(s,a,s') \hat{c}_h^j(s,a) + \sum_{s,a,s'} q_h^k(s,a,s') \frac{(\eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a))^2}{2} \right) \\
&\leq -\eta \sum_h \sum_{s,a,s'} \sum_{j:j+d^j=k} q_h^k(s,a,s') \hat{c}_h^j(s,a) + \sum_h \sum_{s,a,s'} q_h^k(s,a,s') \frac{(\eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a))^2}{2},
\end{aligned}
$$

where the second inequality is by $e^s \leq 1 + s + s^2/2$ for $s \leq 0$, and the third inequality is by $\log(1+s) \leq s$ for all s. Term (B) can be rewritten as

$$
\begin{aligned}
(B) &= \sum_h \sum_{s,a,s'} q_h^k(s,a,s') (e_h^{\mu^k,\beta^k}(s,a,s') + v_h^{\mu^k}(s,a,s') - \eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a) - \sum_{s''} \bar{p}_h^k(s'' \mid s,a) v_{h+1}^{\mu^k}(s,a,s'')) \\
&= \sum_h \sum_{s,a,s'} q_h^k(s,a,s') e_h^{\mu^k,\beta^k}(s,a,s') + \sum_h \sum_{s,a,s'} q_h^k(s,a,s') v_h^{\mu^k}(s,a,s') \\
&\quad - \eta \sum_h \sum_{s,a,s'} \sum_{j:j+d^j=k} q_h^k(s,a,s') \hat{c}_h^j(s,a) - \sum_h \sum_{s,a,s'} \sum_{s''} q_h^k(s,a,s') \bar{p}_h^k(s'' \mid s,a) v_{h+1}^{\mu^k}(s,a,s'') \\
&= \sum_h \sum_{s,a,s'} q_h^k(s,a,s') e_h^{\mu^k,\beta^k}(s,a,s') + \sum_h \sum_{s,a,s'} q_h^k(s,a,s') v_h^{\mu^k}(s,a,s') \\
&\quad - \eta \sum_h \sum_{s,a,s'} \sum_{j:j+d^j=k} q_h^k(s,a,s') \hat{c}_h^j(s,a) - \sum_h \sum_{s,a} \sum_{s''} q_h^k(s,a) \bar{p}_h^k(s'' \mid s,a) v_{h+1}^{\mu^k}(s,a,s'').
\end{aligned}
$$

Notice that:

$$
\begin{aligned}
&\sum_{h,s,a,s''} q_h^k(s,a) \bar{p}_h^k(s'' \mid s,a) v_{h+1}^{\mu^k}(s,a,s'') \\
&\qquad = \sum_{h,s,a,s''} q_h^k(s,a) p_h^k(s'' \mid s,a) v_{h+1}^{\mu^k}(s,a,s'') + \sum_{h,s,a,s''} q_h^k(s,a) (\bar{p}_h^k(s'' \mid s,a) - p_h^k(s'' \mid s,a)) v_{h+1}^{\mu^k}(s,a,s'') \\
&\qquad = \sum_{h,s,a,s''} q_{h+1}^k(s,a,s'') v_{h+1}^{\mu^k}(s,a,s'') + \sum_{h,s,a,s''} q_h^k(s,a) (\bar{p}_h^k(s'' \mid s,a) - p_h^k(s'' \mid s,a)) v_{h+1}^{\mu^k}(s,a,s''),
\end{aligned}
$$

and therefore:

$$
\begin{aligned}
(B) = \sum_h \sum_{s,a,s'} q_h^k(s,a,s') e_h^{\mu^k,\beta^k}(s,a,s') &- \eta \sum_h \sum_{s,a,s'} \sum_{j:j+d^j=k} q_h^k(s,a,s') \hat{c}_h^j(s,a) \\
&- \sum_{h,s,a,s''} q_h^k(s,a) (\bar{p}_h^k(s'' \mid s,a) - p_h^k(s'' \mid s,a)) v_{h+1}^{\mu^k}(s,a,s'').
\end{aligned}
$$

Overall we get:

$$
\begin{aligned}
\sum_h \mathrm{KL}(q_h^k \parallel q_h^{k+1}) \leq \sum_{h,s,a,s'} q_h^k(s,a,s') \frac{(\eta \sum_{j:j+d^j=k} \hat{c}_h^j(s,a))^2}{2} &- \sum_{h,s,a,s'} q_h^k(s,a,s') e_h^{\mu^k,\beta^k}(s,a,s') \\
&+ \sum_{h,s,a,s'} q_h^k(s,a) (\bar{p}_h^k(s' \mid s,a) - p_h^k(s' \mid s,a)) v_{h+1}^{\mu^k}(s,a,s').
\end{aligned}
$$

To finish the proof we show that:

$$\sum_{h,s,a,s'} q_h^k(s,a)(\bar{p}_h^k(s' \mid s,a) - p_h^k(s' \mid s,a))v_{h+1}^{\mu^k}(s,a,s'') \leq \sum_{h,s,a,s'} q_h^k(s,a,s')e_h^{\mu^k,\beta^k}(s,a,s').$$

By definition of v^{μ^k} and ϵ^k, and since $\mu \geq 0$, we have:

$$\begin{aligned}
&\sum_{h,s,a,s'} q_h^k(s,a)(\bar{p}_h^k(s' \mid s,a) - p_h^k(s' \mid s,a))v_{h+1}^{\mu^k}(s,a,s'') \\
&\qquad = \sum_{h,s,a,s'} q_h^k(s,a)(\bar{p}_h^k(s' \mid s,a) - p_h^k(s' \mid s,a))(\mu_{h+1}^{k,-}(s,a,s') - \mu_{h+1}^{k,+}(s,a,s')) \\
&\qquad \leq \sum_{h,s,a,s'} q_h^k(s,a)|\bar{p}_h^k(s' \mid s,a) - p_h^k(s' \mid s,a)|(\mu_{h+1}^{k,-}(s,a,s') + \mu_{h+1}^{k,+}(s,a,s')) \\
&\qquad \leq \sum_{h,s,a,s'} q_h^k(s,a)\epsilon_h^k(s' \mid s,a)(\mu_{h+1}^{k,-}(s,a,s') + \mu_{h+1}^{k,+}(s,a,s')),
\end{aligned}$$

so to finish the proof it suffices to show that $\sum_{h,s,a,s'} \beta_{h+1}^k(s') - \beta_h^k(s) = 0$. Indeed, this follows as the sum is telescopic and $\beta_{H+1}^k = \beta_0^k = 0$. ■