Learning Adversarial Markov Decision Processes with Delayed Feedback

Tal Lancewicki [*1] **Aviv Rosenberg** [*1] **Yishay Mansour** [1 2]

Abstract

Reinforcement learning typically assumes that the agent observes feedback for its actions immediately, but in many real-world applications (like recommendation systems) the feedback is observed in delay. In this paper, we study online learning in episodic Markov decision processes (MDPs) with unknown transitions, adversarially changing costs and unrestricted delayed feedback. That is, the costs and trajectory of episode k are revealed to the learner only in the end of episode $k + d^k$, where the delays d^k are neither identical nor bounded, and are chosen by an oblivious adversary. We present novel algorithms based on policy optimization that achieve near-optimal high-probability regret of $\sqrt{K + D}$ under full-information feedback, where K is the number of episodes and $D = \sum_k d^k$ is the total delay. Under bandit feedback, we prove similar $\sqrt{K + D}$ regret assuming the costs are stochastic, and $(K + D)^{2/3}$ regret in the general case. We are the first to consider regret minimization in the important setting of MDPs with delayed feedback.

1. Introduction

Delayed feedback is a fundamental challenge in sequential decision making that arises in almost every practical application. For example, recommendation systems learn the utility of a recommendation by detecting the occurrence of certain events (e.g., user conversions), which may happen with a variable delay after the recommendation was issued. Other examples include display advertising, autonomous vehicles, delays in video streaming (Changuel et al., 2012), communication delays experienced by interacting learning agents (Chen et al., 2020a) and system delays in robotics (Mahmood et al., 2018).

[*]Equal contribution [1]Tel-Aviv University [2]Google Research, Tel Aviv. Correspondence to: Tal Lancewicki <lancewicki@mail.tau.ac.il>, Aviv Rosenberg <avivros007@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Although handling delayed feedback is crucial for applying reinforcement Learning (RL) algorithms in practice, it was only barely studied from a theoretical perspective, as most of the RL literature focuses on the Markov decision process (MDP) model in which the agent observes feedback regarding her immediate reward and transition to the next state right after performing an action.

In this paper we make a substantial step towards closing the major gap of RL with delayed feedback in the literature. Specifically, we consider the challenging setting of adversarial episodic MDPs in which the cost function changes arbitrarily between episodes while the transition function remains stationary over time (but unknown to the agent). We present the model of *adversarial MDPs with delayed feedback* in which the agent observes feedback for episode k only in the end of episode $k + d^k$, where the delays d^k are unknown and not restricted in any way. This model generalizes adversarial MDPs without delays (where $d^k = 0$ for all k), and it encompasses great challenges that do not arise in standard RL models, e.g., latency in policy updates and exploration without feedback.

We develop novel algorithms based on policy optimization (PO) that perform their updates whenever feedback is available and ignore feedback with large delay, and prove that they obtain high-probability regret bounds of $\widetilde{O}(\sqrt{K} + \sqrt{D})$ under full-information feedback and $\widetilde{O}(K^{2/3} + D^{2/3})$ under bandit feedback, where K is the number of episodes and D is the sum of delays. Unlike simple reductions that can only handle fixed delay d, our algorithms are robust to any kind of variable delays and do not require any prior knowledge. Furthermore, we show that a naive adaptation of existing algorithms suffers from sub-optimal dependence in the number of actions, and present a novel technique that forces exploration in order to achieve tight bounds. To complement our results, we present nearly matching lower bounds of $\Omega(\sqrt{K} + \sqrt{D})$. See detailed bounds in Table 1.

As delays were extensively studied in multi-arm bandits (MAB) (Auer et al., 2002), our algorithms naturally build on ideas from that literature. However, combining them with methodologies for regret minimization in RL is a complex challenge that requires novel techniques and careful analysis. MDPs are highly challenging compared to MAB because the underlying dynamics must be learned while handling

Table 1. Comparison between the regret bounds of our algorithms Delayed OPPO (D-OPPO) and Delayed O-REPS (D-O-REPS), and our lower bound under both full-information feedback (full) and bandit feedback (bandit), for number of episodes K, total delay D, horizon H, number of states S and number of actions A. "Known Transition" assumes dynamics are known to the learner in advance, and "Unknown Transition" means that the learner needs to learn the dynamics. "Delayed Cost" assumes only costs are observed in delay, while in "Delayed Trajectory" the trajectory is also observed in delay, together with the costs. All bounds ignore constant and poly-logarithmic factors.

	Known Transition + Delayed Trajectory	Unknown Transition + Delayed Cost	Unknown Transition + Delayed Trajectory
D-O-REPS (full)	$H\sqrt{K+D}$	$H^{3/2}S\sqrt{AK}+H\sqrt{D}$	$H^2S\sqrt{AK}+H^{3/2}\sqrt{SD}$
D-OPPO (full)	$H^2\sqrt{K+D}$	$H^{3/2}S\sqrt{AK}+H^2\sqrt{D}$	$H^2S\sqrt{AK}+H^{3/2}\sqrt{SD}$
Lower Bound (full)	$H\sqrt{K+D}$	$H^{3/2}\sqrt{SAK}+H\sqrt{D}$	$H^{3/2}\sqrt{SAK}+H\sqrt{D}$
D-OPPO (bandit)	$HS\sqrt{A}K^{2/3}+H^2D^{2/3}$	$HS\sqrt{A}K^{2/3}+H^2D^{2/3}$	$HS\sqrt{A}K^{2/3}+H^2D^{2/3}$
Lower Bound (bandit)	$H\sqrt{SAK}+H\sqrt{D}$	$H^{3/2}\sqrt{SAK}+H\sqrt{D}$	$H^{3/2}\sqrt{SAK}+H\sqrt{D}$

delays, and exploration must be performed with limited feedback. This work belongs to the important line of papers that generalize MAB methods to the challenging MDP environment. Starting with the seminal UCRL algorithm (Jaksch et al., 2010) that adapts the UCB algorithm from MAB to MDP, this line of work includes numerous important contributions, like best-of-both-worlds (Jin & Luo, 2020), corruptions (Lykouris et al., 2019), incentives (Simchowitz & Slivkins, 2021), etc.

1.1. Related work

Delays in RL. Although delay is a common challenge that RL algorithms need to face in practice (Schuitema et al., 2010; Liu et al., 2014; Changuel et al., 2012; Mahmood et al., 2018), the theoretical literature on the subject is very limited. Katsikopoulos & Engelbrecht (2003) considered delayed state observability, in which the state is observable in delay and the agent has to pick an action without full knowledge of its current state. They showed that, for fixed delay d, this problem can be reduced to a non-delayed MDP of size exponential in d. Walsh et al. (2009) further studied this direction, and proved that this exponential dependence is necessary since the planning problem in this setting is computationally hard. The focus of this paper is the problem of delayed feedback rather than delayed state observability, and thus these results are only partially related as the challenges are different.

Delays in MAB. Delays where extensively studied in recent years in MAB since this is an extremely fundamental issue that arises in many real applications (Vernade et al., 2017; Pike-Burke et al., 2018; Cesa-Bianchi et al., 2018; Zhou et al., 2019; Gael et al., 2020; Lancewicki et al., 2021). Our work is most related to the literature on delays in adversarial MAB: Cesa-Bianchi et al. (2016) showed that the optimal regret for MAB with fixed delay d is $\widetilde{\Theta}(\sqrt{(A+d)K})$, where A is the number of actions. Even earlier, variable delays were studied by Quanrud & Khashabi (2015) in the setting of online learning with full-information feedback, where they showed optimal regret of $\widetilde{\Theta}(\sqrt{K+D})$. More recently, Thune et al. (2019); Bistritz et al. (2019); Zimmert & Seldin (2020); György & Joulani (2020) studied variable delays in MAB, proving optimal $\widetilde{\Theta}(\sqrt{AK+D})$ regret. Note that unlike MDPs, in MAB there is no underlying dynamics, and the only challenge is feedback about the cost arriving in delay.

Regret minimization in stochastic MDPs. There is a vast literature on regret minimization in RL that mostly builds on the optimism in face of uncertainty principle. Most literature focuses on the tabular setting, where the number of states is small (Bartlett & Tewari, 2009; Jaksch et al., 2010; Osband et al., 2016; Azar et al., 2017; Dann et al., 2017; Jin et al., 2018; Zanette & Brunskill, 2019; Efroni et al., 2019; Shani et al., 2020; Fruit et al., 2018; Simchowitz & Jamieson, 2019; Tarbouriech et al., 2020; Rosenberg et al., 2020). Recently it was extended to function approximation under various assumptions (Jin et al., 2020b; Cai et al., 2020; Yang & Wang, 2019; Zanette et al., 2020a;b).

Adversarial MDPs. Early works on adversarial MDPs (Even-Dar et al., 2009; Neu et al., 2014; 2010; 2012) focused on known transitions and used various reductions to MAB. Zimin & Neu (2013) presented a reduction to online linear optimization, known as O-REPS, that achieves optimal regret bounds with known dynamics. Later, this was extended to unknown dynamics by Rosenberg & Mansour (2019b;a); Jin et al. (2020a) obtaining near-optimal regret bounds. The O-REPS method was also extended to stochastic shortest path problems with adversarial costs (Rosenberg & Mansour, 2020; Chen et al., 2020b; Chen & Luo, 2021). Recently, Cai et al. (2020); Shani et al. (2020) showed that PO methods (that are widely used in practice) also achieve near-optimal regret bounds.

2. Setting

A finite-horizon adversarial MDP is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, p, \{c^k\}_{k=1}^K)$, where $\mathcal{S}$ is the state space with cardinality S, $\mathcal{A}$ is the action space with cardinality A, H is the horizon (i.e., episode length), $p = \{p_h : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0,1]\}_{h=1}^H$ is the transition function, and $c^k = \{c_h^k : \mathcal{S}\times\mathcal{A} \to [0,1]\}_{h=1}^H$ is the cost function for episode k. To simplify presentation we assume $S \geq \max\{A, H^2\}$.

The interaction between the learner and the environment proceeds as follows. At the beginning of episode k, the learner starts in a fixed initial state[1] $s_1^k = s_{\text{init}} \in S$ and picks a policy $\pi^k = \{\pi_h^k : \mathcal{A}\times\mathcal{S} \to [0,1]\}_{h=1}^H$ where $\pi_h^k(a \mid s)$ gives the probability that the agent takes action a at time h given that the current state is s. Then, the policy is executed in the MDP generating a trajectory $U^k = \{(s_h^k, a_h^k)\}_{h=1}^H$, where $a_h^k \sim \pi_h^k(\cdot|s_h^k)$ and $s_{h+1}^k \sim p_h(\cdot|s_h^k, a_h^k)$. With no delays, the learner observes the feedback in the end of the episode, that is, the trajectory U^k and either the entire cost function c^k under *full-information feedback* or the suffered costs $\{c_h^k(s_h^k, a_h^k)\}_{h=1}^H$ under *bandit feedback*. In contrast, with delayed feedback, these are revealed to the learner only in the end of episode $k + d^k$, where the delays $\{d^k\}_{k=1}^K$ are unknown and chosen by an oblivious adversary before the interaction starts. Denote the maximal delay by $d_{max} = \max_{k=1,\ldots,K} d^k$ and the total delay by $D = \sum_{k=1}^K d^k$. Notice that adversarial MDPs without delays are a special case in which $d^k = 0\ \forall k$.

For a given policy π, we define its expected cost with respect to cost function c, when starting from state s at time h, as $V_h^\pi(s) = \mathbb{E}\big[\sum_{h'=h}^H c_{h'}(s_{h'}, a_{h'}) \mid s_h = s, \pi, p\big]$, where the expectation is taken over the randomness of the transition function p and the policy π. This is known as the *value function* of π, and we also define the *Q-function* by $Q_h^\pi(s,a) = \mathbb{E}\big[\sum_{h'=h}^H c_{h'}(s_{h'}, a_{h'})|s_h = s, a_h = a, \pi, p\big]$. The value function and Q-function satisfy the Bellman equations (see Sutton & Barto (2018)):

$$\begin{aligned} Q_h^\pi(s,a) &= c_h(s,a) + \langle p_h(\cdot \mid s,a), V_{h+1}^\pi\rangle \\ V_h^\pi(s) &= \langle \pi_h(\cdot \mid s), Q_h^\pi(s,\cdot)\rangle, \end{aligned} \tag{1}$$

where $\langle\cdot,\cdot\rangle$ is the dot product. Let $V^{k,\pi}$ be the value function of π with respect to c^k. We measure the performance of the learner by the *regret*, which is the cumulative difference between the value of the learner's policies and the value of the best fixed policy in hindsight, i.e.,

$$\mathcal{R}_K = \sum_{k=1}^K V_1^{k,\pi^k}(s_1^k) - \min_\pi \sum_{k=1}^K V_1^{k,\pi}(s_1^k).$$

Additional Notations. Generally indices of episodes appear as superscripts and indices of in-episode time steps as subscripts. $\mathcal{F}^k = \{j : j + d^j = k\}$ denotes the set of episodes that their feedback arrives in the end of episode k, and the number visits to state-action pair (s,a) at time h by the end of episode $k-1$ is denoted by $m_h^k(s,a)$. Similarly, $n_h^k(s,a)$ denotes the number of these visits for which feedback was observed until the end of episode $k-1$. $\mathbb{E}^\pi[\cdot] = \mathbb{E}[\cdot \mid s_1^k = s_{\text{init}}, \pi, p]$ denotes the expectation given a policy π, the notation $\widetilde{O}(\cdot)$ ignores constant and poly-logarithmic factors and $x \vee y = \max\{x, y\}$. We denote the set $\{1,\ldots,n\}$ by $[n]$, and the indicator of an event E by $\mathbb{I}\{E\}$.

3. Warm-up: a black-box reduction

One simple way to deal with delays (adopted in several MAB and online optimization works, e.g., Weinberger & Ordentlich (2002); Joulani et al. (2013)) is to simulate a non-delayed algorithm and use its regret guarantees. Specifically, we can maintain $d_{max}+1$ instances of the non-delayed algorithm, running the i-th instance on the episodes k such that $k = i \mod (d_{max}+1)$. That is, at the first $d_{max}+1$ episodes, the learner plays the first policy that each instance outputs. By the end of episode $d_{max}+1$, the feedback for the first episode is observed, allowing the learner to feed it to the first instance. The learner would then play the second output of that instance, and so on. Effectively, each instance plays $K/(d_{max}+1)$ episodes, so we can use the regret of the non-delayed algorithm $\widetilde{\mathcal{R}}_K$ in order to bound the learner's regret by $\mathcal{R}_K \leq (d_{max}+1)\widetilde{\mathcal{R}}_{K/(d_{max}+1)}$. Plugging in the regret bounds of Rosenberg & Mansour (2019b); Jin et al. (2020a) for adversarial MDPs without delays, we obtain the following regret for both full-information and bandit feedback,

$$\mathcal{R}_K = \widetilde{O}\big(H^2 S\sqrt{AK(d_{max}+1)} + H^2S^2A(d_{max}+1)\big).$$

Although simple, the black-box reduction approach suffers from many evident shortcomings. First, it is highly non-robust to variable delays as its regret scales with the *worst-case delay* Kd_{max} which becomes very large even if the feedback from just one episode is missing. One of the major challenges that we tackle in the rest of the paper is to achieve regret bounds that are independent of d_{max} and scale with the *average delay*, i.e., the total delay D which is usually much smaller than worst-case. Second, even if we ignore the problematic dependence in the worst-case delay, this regret bound is still sub-optimal as it suggests a multiplicative relation between d_{max} and A (and S^2) which does not appear in the MAB setting. Our analysis focuses on eliminating this sub-optimal dependence through a clever algorithmic feature that forces exploration and ensures tight near-optimal regret. Finally, the reduction is highly inefficient as it requires running $d_{max}+1$ different algorithms in parallel. Moreover, the $\sqrt{Kd_{max}}$ regret

[1] The algorithm readily extends to a fixed initial distribution.

under bandit feedback is only achievable using O-REPS algorithms that are extremely inefficient to implement in practice. In contrast, our algorithm is based on efficient and practical PO methods. Its running time is independent of the delays and it does not require any prior knowledge or parameter tuning (unlike the reduction that needs d_{max}).

4. Delayed OPPO: a policy optimization algorithm for delayed feedback

In this section we present *Delayed OPPO* (Algorithm 1 and with more details in Appendix A) – the first algorithm for regret minimization in adversarial MDPs with delayed feedback. Delayed OPPO is a policy optimization algorithm, and therefore implements a smoother version of Policy Iteration (Sutton & Barto, 2018). That is, it alternates between a policy evaluation step – in which an optimistic estimate for the Q-function of the learner's policy is computed, and a policy improvement step – where the learner's policy is improved in a "soft" manner regularized by the KL-divergence.

Delayed OPPO is based on the optimistic proximal policy optimization (OPPO) algorithm of Cai et al. (2020); Shani et al. (2020). As a policy optimization algorithm, it enjoys many merits of practical PO algorithms that have had great empirical success in recent years, e.g., TRPO (Schulman et al., 2015), PPO (Schulman et al., 2017) and SAC (Haarnoja et al., 2018). It is computationally efficient, easy to implement and readily extends to function approximation.

The main difference between Delayed OPPO and previous algorithms is that it performs its updates using all available feedback at the current time step. Moreover, in Sections 4.1 and 4.2 we equip our algorithm with novel mechanisms that make it robust to all kinds of variable delays without any prior knowledge and enable us to prove tight regret bounds. Even with these algorithmic mechanisms, proving our regret bounds requires careful analysis and new ideas that do not appear in the MAB with delays literature, as we tackle the much more complex MDP environment.

In the beginning of episode k, the algorithm computes an optimistic estimate Q^j of Q^{π^j} for all the episodes j that their feedback just arrived. To that end, we maintain confidence sets that contain the true transition function p with high probability, and are built using all the trajectories available at the moment. That is, for every (s,a,h), we compute the empirical transition function $\bar{p}_h^k(s' \mid s,a)$ and define the confidence set $\mathcal{P}_h^k(s,a)$ as the set of transition functions $p'_h(\cdot \mid s,a)$ such that, for every $s' \in \mathcal{S}$,

$$|p'_h(s' \mid s,a) - \bar{p}_h^k(s' \mid s,a)| \le \epsilon_h^k(s' \mid s,a)$$
$$\stackrel{\text{def}}{=} 4\sqrt{\frac{\bar{p}_h^k(s' \mid s,a)\ln\frac{HSAK}{4\delta}}{1 \vee n_h^k(s,a)}} + \frac{10\ln\frac{HSAK}{4\delta}}{1 \vee n_h^k(s,a)}.$$

Algorithm 1 Delayed OPPO

Input: $\mathcal{S}, \mathcal{A}, H, \eta > 0, \gamma > 0, \delta > 0$.
Initialization: Set $\pi_h^1(a \mid s) = 1/A$ for every (s,a,h).
for $k = 1, 2, \ldots, K$ **do**
  Play episode k with policy π^k, and observe feedback from all episodes $j \in \mathcal{F}^k$.
  Compute cost estimators $\hat{c}^j$ and confidence set $\mathcal{P}^k$.
  # Policy Evaluation
  for $j \in \mathcal{F}^k$ **do**
    Set $V_{H+1}^j(s) = 0$ for every $s \in \mathcal{S}$.
    for $h = H, \ldots, 1$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
      Compute:

$$\hat{p}_h^j(\cdot|s,a) \in \operatorname*{arg\,min}_{p'_h(\cdot|s,a) \in \mathcal{P}_h^k(s,a)} \langle p'_h(\cdot \mid s,a), V_{h+1}^j \rangle.$$
$$Q_h^j(s,a) = \hat{c}_h^j(s,a) + \langle \hat{p}_h^j(\cdot \mid s,a), V_{h+1}^j \rangle$$
$$V_h^j(s) = \langle Q_h^j(s,\cdot), \pi_h^j(\cdot \mid s) \rangle.$$

    end for
  end for
  # Policy Improvement
  Set for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$:

$$\pi_h^{k+1}(a \mid s) = \frac{\pi_h^k(a \mid s) e^{-\eta \sum_{j \in \mathcal{F}^k} Q_h^j(s,a)}}{\sum_{a' \in \mathcal{A}} \pi_h^k(a' \mid s) e^{-\eta \sum_{j \in \mathcal{F}^k} Q_h^j(s,a')}}.$$

end for

Then, the confidence set for episode k is defined by $\mathcal{P}^k = \{\mathcal{P}_h^k(s,a)\}_{s,a,h}$. Under bandit feedback, the computation of Q^j also requires estimating the cost function c^j in state-action pairs that were not visited in that episode. For building these estimates, we utilize the optimistic importance-sampling estimator of Jin et al. (2020a) that first optimistically estimates the probability to visit each state s in each time h of episode j by

$$u_h^j(s) = \max_{p' \in \mathcal{P}^j} \Pr[s_h = s \mid s_1 = s_{\text{init}}, \pi^j, p']$$

and then (for exploration parameter $\gamma > 0$) sets the estimator to be $\hat{c}_h^j(s,a) = \frac{c_h^j(s,a)\mathbb{I}\{s_h^j = s, a_h^j = a\}}{u_h^j(s)\pi_h^j(a|s) + \gamma}$.

After the optimistic Q-functions are computed, we use them to improve the policy via a softmax update, i.e., for a learning rate $\eta > 0$ we update $\pi_h^{k+1}(a \mid s) \propto \pi_h^k(a \mid s)\exp(-\eta \sum_{j \in \mathcal{F}^k} Q_h^j(s,a))$. This update form, which may be characterized as an online mirror descent (Beck & Teboulle, 2003) step with KL-regularization, stands in the heart of the following regret analysis (full proofs are found in Appendix B). We note that the following theorem handles only delayed feedback regarding the costs, while assuming that feedback regarding the learner's trajectory arrives without delay.

Theorem 1. *Running Delayed OPPO with delayed cost feedback and non-delayed trajectory feedback guarantees the following regret bounds with probability* $1-\delta$*. Under full-information feedback*

$$\mathcal{R}_K = \widetilde{O}(H^{3/2}S\sqrt{AK} + H^2\sqrt{D}).$$

Under bandit feedback

$$\mathcal{R}_K = \widetilde{O}(HS\sqrt{A}K^{2/3} + H^2D^{2/3} + H^2 d_{max}).$$

Notice that the regret bound in Theorem 1 overcomes the major problems that we had with the black-box reduction approach. Namely, the regret scales with the total delay D and not the worst-case delay Kd_{max} (the extra additive dependence in d_{max} is avoided altogether in Section 4.2), and D is not multiplied by neither S nor A. Finally, as a direct corollary of Theorem 1, we deduce the regret bound for the known transitions case, in which term (A) does not appear (at least under full-information feedback). Notice that with known transitions, there is no need to handle delays in the trajectory feedback since the algorithm has full knowledge of the transition function in advance.

Theorem 2. *Running Delayed OPPO with known transition function guarantees the following regret bounds with probability at least* $1-\delta$*. Under full-information feedback* $\mathcal{R}_K = \widetilde{O}(H^2\sqrt{K+D})$*. Uder bandit feedback* $\mathcal{R}_K = \widetilde{O}(HS\sqrt{A}K^{2/3} + H^2D^{2/3} + H^2 d_{max})$.

Proof sketch of Theorem 1. By a standard regret decomposition (based on the value difference lemma of Shani et al. (2020)), we can show that the regret scales with two main terms: $(A) = \sum_{k=1}^{K} V_1^{\pi^k}(s_1^k) - V_1^k(s_1^k)$ is the bias between the estimated and true value of π^k; and $(B) = \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}^{\pi}[\langle Q_h^k(s_h^k,\cdot), \pi_h^k(\cdot \mid s_h^k) - \pi_h(\cdot \mid s_h^k)\rangle]$ which, when fixing $(s,h) \in \mathcal{S}\times[H]$, can be viewed as the regret of a delayed MAB algorithm with full-information feedback, where the losses are the estimated Q-functions.

Since the trajectories are not observed in delay, we can bound term (A) similarly to Shani et al. (2020) using our confidence sets that shrink over time. To bound term (B), we fix (s,h) and follow a "cheating" algorithm technique (György & Joulani, 2020). To that end, we define the "cheating" algorithm that does not experience delay and sees one step into the future, i.e., in episode k it plays policy $\bar{\pi}_h^{k+1}(a|s) \propto e^{-\eta\sum_{j=1}^{k} Q_h^j(s,a)}$. Then, we can break term (B) into two terms: (i) The regret of the "cheating" algorithm which is bounded by $\frac{\log A}{\eta}$ using a Be-The-Leader argument (see, e.g., Joulani et al. (2020)), and (ii) The difference between $\bar{\pi}^{k+1}$ and π^k which we can bound by looking at the exponential weights update form. Specifically, we bound the ratio $\bar{\pi}_h^{k+1}(a|s)/\pi_h^k(a|s)$ from below by $1-\eta\sum_{j\le k, j+d^j\ge k} Q_h^j(s,a)$ and this bounds term (ii) in terms of the missing feedback, i.e.,

$$\begin{aligned}
(ii) &= \sum_{k=1}^{K}\sum_{a\in A} Q_h^k(s,a)\big(\pi_h^k(a \mid s) - \bar{\pi}_h^{k+1}(a \mid s)\big) \\
&= \sum_{k=1}^{K}\sum_{a\in A} \pi_h^k(a|s) Q_h^k(s,a)\Big(1 - \frac{\bar{\pi}_h^{k+1}(a \mid s)}{\pi_h^k(a \mid s)}\Big) \\
&\le \eta \sum_{k=1}^{K}\sum_{a\in A} \pi_h^k(a \mid s) Q_h^k(s,a) \sum_{j\le k, j+d^j\ge k} Q_h^j(s,a).
\end{aligned}$$

Under full-information feedback, all the Q-function estimates are bounded by H, which leads to

$$\begin{aligned}
(ii) &\le \eta H^2 \sum_{k=1}^{K} |\{j\in[k] : j + d^j \ge k\}| \\
&= \eta H^2 \sum_{k=1}^{K}\sum_{j=1}^{K} \mathbb{I}\{j\le k, j+d^j\ge k\} \\
&= \eta H^2 \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbb{I}\{j\le k\le j+d^j\} \\
&\le \eta H^2 \sum_{j=1}^{K}(1+d^j) = \eta H^2(K+D).
\end{aligned}$$

To finish the proof we set $\eta = 1/H\sqrt{K+D}$. Under bandit feedback, this argument becomes a lot more delicate because the Q-function estimates are naively bounded only by H/γ. Thus, we need to prove concentration of $\sum_k V_h^k(s)$ around $\sum_k V_h^{\pi^k}(s)$ (which is indeed bounded by HK). □

4.1. Handling delayed trajectories

Previously, we analyzed the Delayed OPPO algorithm in the setting where only cost is observed in delay. In this section, we face the *delayed trajectory feedback* setting in which the trajectory of episode k is observed only in the end of episode $k+d^k$ together with the cost. Delayed trajectory feedback is a unique challenge in MDPs that does not arise in MAB, as no underlying dynamics exist. Next, we provide the first analysis for delays of this kind and shed light on the difficulties involved.

Let us focus on full-information feedback to convey the main ideas. Consider the natural analysis that simply investigates our confidence sets which are now updated in delay (since they are constructed using the observed trajectories). Specifically, we carefully follow the stochastic process that describes the way that the confidence sets shrink over time, and use concentration arguments to prove that term (A) from the proof of Theorem 1 can now be bounded by $H^2S\sqrt{A(K+D)}$. As discussed before, this is far

from optimal since the total delay should not scale with the number of states and actions.

To get a better regret bound, we must understand the new challenges to cause this sub-optimality. The main issue here is *wasted exploration* due to unobserved feedback. To first tackle this issue, we leverage the key observation that the importance of unobserved exploration becomes less significant as time progresses, since our understanding of the underlying dynamics is already substantial. With this in mind we propose a new technique to analyze term (A) – isolate the first $\approx d_{max}$ visits to each state-action pair, and for other visits use the fact that some knowledge of the transition function is already evident. With this technique we are able to get the improved bound $(A) \lesssim H^2 S\sqrt{AK} + H^2 SAd_{max}$.

Note that this is a major improvement since d_{max} is much smaller than D, and avoided altogether in Section 4.2. However, we still see the undesirable multiplicative relation with S and A. To tighten the regret bound even further we propose a novel algorithmic mechanism to specifically direct wasted exploration. The mechanism, that we call *explicit exploration*, forces the agent to explore until it observes sufficient amount of feedback. Specifically, until we observe feedback for $2d_{max}\log\frac{HSA}{\delta}$ visits to state s at time h, we pick actions uniformly at random in this state. The explicit exploration mechanism directly improves the bound on term (A) by a factor of A (as shown in the following theorem), and is in fact necessary for optimistic algorithms in the presence of delays (see Section 5).

Theorem 3. *Running Delayed OPPO with explicit exploration, with delayed cost feedback and delayed trajectory feedback guarantees the following regret bounds with probability at least $1-\delta$. Under full-information feedback*

$$\mathcal{R}_K = \widetilde{O}(H^2 S\sqrt{AK} + H^2\sqrt{D} + H^2 Sd_{max}).$$

Under bandit feedback

$$\mathcal{R}_K = \widetilde{O}(HS\sqrt{A}K^{2/3} + H^2 D^{2/3} + H^2 SAd_{max}).$$

Proof sketch. We start by isolating episodes in which we visit some state for which we observed less than $2d_{max}\log\frac{HSA}{\delta}$ visits. Since d_{max} is the maximal delay, there are only $\widetilde{O}(HSd_{max})$ such episodes (and the cost in each episode is at most H). For the rest of the episodes, by virtue of explicit exploration, we now have that the number of observed visits to each (s,a,h) is at least d_{max}/A.

Term (A) that measures the Q-functions estimation error is controlled by the rate at which the confidence sets shrink. Let $\mathbb{I}_h^k(s,a) = \mathbb{I}\{s_h^k = s, a_h^k = a\}$, we can bound it as follows with standard analysis, $(A) \lesssim H\sqrt{S}\sum_{s\in S}\sum_{a\in A}\sum_{h=1}^{H}\sum_k {}^{\mathbb{I}_h^k(s,a)}/\sqrt{n_h^k(s,a)}$. Now we address the delays. Fix (s,a,h) and denote the number of unobserved visits by $N_h^k(s,a) = \max\{m_h^k(s,a) - n_h^k(s,a), 0\}$. Next, we decouple the statistical estimation error and the effect of the delays in the following way,

$$\begin{aligned}\sum_k \frac{\mathbb{I}_h^k(s,a)}{\sqrt{n_h^k(s,a)}} &= \sum_k \frac{\mathbb{I}_h^k(s,a)}{\sqrt{m_h^k(s,a)}}\sqrt{\frac{m_h^k(s,a)}{n_h^k(s,a)}} \\ &\le \sum_k \frac{\mathbb{I}_h^k(s,a)}{\sqrt{m_h^k(s,a)}}\sqrt{1+\frac{N_h^k(s,a)}{n_h^k(s,a)}} \\ &\le \sum_k \frac{\mathbb{I}_h^k(s,a)}{\sqrt{m_h^k(s,a)}} + \sum_k \frac{\mathbb{I}_h^k(s,a)}{\sqrt{m_h^k(s,a)}}\sqrt{\frac{N_h^k(s,a)}{n_h^k(s,a)}}.\end{aligned} \tag{2}$$

The first term is unaffected by delays and bounded by $H^2 S\sqrt{AK}$. For the second term, we utilize explicit exploration in the sense that $n_h^k(s,a) \ge d_{max}/A$. Combine this with the observation that $N_h^k(s,a) \le d_{max}$ (since d_{max} is the maximal delay), to obtain the bound $H^2 SA\sqrt{K}$. Finally, to get the tight bound (i.e., eliminate the extra $\sqrt{A}$), we split the second sum into: (1) episodes with $n_h^k(s,a) \ge d_{max}$ where ${}^{N_h^k(s,a)}/_{n_h^k(s,a)}$ is tightly bounded by 1 (and not A), and (2) episodes with $n_h^k(s,a) < d_{max}$ in which the regret scales as $\sqrt{d_{max}}$ (which is at most $\sqrt{K}$). □

4.2. Handling large delays and unknown total delay

In this section we address the two final issues with our Delayed OPPO algorithm. First, we eliminate the dependence in the maximal delay d_{max} that may be as large as K even when the total delay is relatively small. Second, we avoid the need for any prior knowledge regarding the delays which is hardly ever available, making the algorithm completely parameter-free.

To handle large delays, we use a simple *skipping* technique (Thune et al., 2019). That is, if some feedback arrives in delay larger than β (where $\beta > 0$ is a skipping parameter), we just ignore it. Thus, effectively, the maximal delay experienced by the algorithm is β, but we also need to bound the number of skipped episodes. To that end, denote the set of skipped episodes by $\mathcal{K}_\beta$ and note that $D = \sum_{k=1}^K d^k \ge |\mathcal{K}_\beta|\beta$, implying that the number of skipped episodes is bounded by $|\mathcal{K}_\beta| \le D/\beta$. In Appendix C we apply the skipping technique to all the settings considered in the paper to obtain the final regret bounds in Table 1. Here, we take the unknown transitions case with delayed trajectory feedback and under full-information feedback as an example. Setting $\beta = \sqrt{D/SH}$ yields the following bound that is independent of the maximal delay d_{max},

$$\begin{aligned}\mathcal{R}_K &= \widetilde{O}(H^2 S\sqrt{AK} + H^2\sqrt{D} + H^2 S\beta + HD/\beta) \\ &= \widetilde{O}(H^2 S\sqrt{AK} + H^{3/2}\sqrt{SD}).\end{aligned}$$

To address unknown number of episodes K and total delay D, we adopt a *doubling* technique (Bistritz et al., 2019) that is widely used in MAB and RL to handle unknown number of episodes. Note that K and D are the only parameters that the algorithm requires, since the skipping scheme replaces the need to know d_{max} with the parameter β (which is tuned using D). The doubling scheme manages the tuning of the algorithm's parameters η, γ, β, making it completely parameter-free and eliminating the need for any prior knowledge regarding the delays.

The doubling scheme maintains an optimistic estimate of $K+D$ and uses it to tune the algorithm's parameters. Every time that the estimate doubles, the algorithm is restarted with the new doubled estimate. This ensures that our optimistic estimate is always relatively close to the true value of $K+D$ and that the number of restarts is only logarithmic, allowing to keep the same regret bounds.

The optimistic estimate of $K+D$ is computed as follows. Denote the number of episodes with missing feedback at the end of episode k by $M^k = \sum_{j=1}^{k} \mathbb{I}\{j + d^j > k\}$. Now notice that $\sum_{k=1}^{K} M^k \leq D$ because the feedback from episode j was missing in exactly d^j episodes. Thus, at the end of episode k our optimistic estimate is $k + \widetilde{D} = k + \sum_{j=1}^{k} M^j$. So for every episode j with observed feedback, its delay is estimated exactly by d^j, and if its feedback was not observed, we estimate it as if we will observe its feedback in the next episode.

In Appendix D, we give the full pseudo-code of Delayed OPPO when combined with doubling, and formally prove that the doubling technique does not damage our regret. As a final note, we mention that the doubling scheme (which may be undesirable in practice since it restarts the algorithm) can be easily replaced with adaptive tuning of the parameters η, γ, β based on the same optimsitic estimate of $K+D$ (Zimmert & Seldin, 2020; György & Joulani, 2020).

5. Additional results and empirical evaluation

Lower bound. For regret minimization in episodic finite-horizon stochastic MDPs, the optimal minimax regret bound is $\widetilde{\Theta}(H^{3/2}\sqrt{SAK})$ (Jaksch et al., 2010; Osband & Van Roy, 2016; Azar et al., 2017; Jin et al., 2018; Zanette & Brunskill, 2019). As adversarial MDPs generalize the stochastic MDP model, this lower bound also applies to our setting. The lower bound for multi-arm bandits with delays is based on a simple reduction to non-delayed MAB with full-information feedback. Namely, we can construct a non-delayed algorithm for full-information feedback using an algorithm $\mathcal{A}$ for fixed delay d by simply feeding $\mathcal{A}$ with the same cost function for d consecutive rounds. Using the known lower bound for full-information MAB, this yields a lower bound of $\Omega(\sqrt{dK}) = \Omega(\sqrt{D})$ which easily translates to a $\Omega(H\sqrt{D})$ lower bound in adversarial MDPs.

Combining these two bounds gives a lower bound of $\Omega(H^{3/2}\sqrt{SAK} + H\sqrt{D})$ for all settings, except for full-information feedback with known dynamics where the lower bound is $\Omega(H\sqrt{K+D})$. In light of this lower bound, we discuss the regret bounds of Delayed OPPO and present open problems.

Under bandit feedback, our $\widetilde{O}(K^{2/3} + D^{2/3})$ regret bounds are still far from the lower bound. However, it is important to emphasize that we cannot expect more from PO methods. Our bounds match the state-of-the-art regret bounds for policy optimization under bandit feedback (Shani et al., 2020). It is an open problem whether PO methods can obtain $\sqrt{K}$ regret bounds in adversarial MDPs under bandit feedback (with either known or unknown dynamics). Currently, the only algorithm with $\sqrt{K}$ regret for this setting is the O-REPS algorithm of Jin et al. (2020a), and it remains an important and interesting open problem to extend it to delayed feedback in the bandit case (see next paragraph).

Under full-information feedback, our regret bounds match the lower bound up to a factor of $\sqrt{S}$ (there is also sub-optimal dependence in H but it can be avoided with Delayed O-REPS as discussed in the next paragraph). However, this extra $\sqrt{S}$ factor already appears in the regret bounds of Rosenberg & Mansour (2019b); Jin et al. (2020a) for adversarial MDPs without delays. Determining the correct dependence in the number of states for adversarial MDPs is an important open problem that must be solved without delays first. We note that if only cost feedback is delayed (and not trajectory feedback), then the delays are not entangled in the estimation of the transition function, and therefore the $\sqrt{D}$ term in our regret bounds is optimal.

Another important note is that, even with delayed trajectory feedback, our $\sqrt{D}$ term is still optimal for a wide class of delays – *monotonic delays*. That is, if the sequence of delays is monotonic, i.e., $d^j \leq d^k$ for $j < k$, then the $\sqrt{D}$ term of our regret bound for delayed trajectory feedback is not multiplied by $\sqrt{S}$. This follows because term (A) in the proof of Theorem 3, that handles the delays in the estimation of the transition function p, can be bounded in this case by $H\sqrt{S}\sum_{h,s,a,k} \mathbb{I}_h^k(s,a)/\sqrt{n_h^{k+d^k}(s,a)}$.[2] Now, under monotonic delays we have $n_h^{k+d^k}(s,a) \geq m_h^k(s,a)$ which allows us to bound term (A) as in the non-delayed case. We note that monotonic delays include the fundamental setting of a fixed delay d.

Delayed O-REPS vs Delayed OPPO. PO methods directly optimize the policy. Practically, this translates to computing

[2] For that, we utilize the fact that our algorithm estimates Q^k at time $k + d^k$, but also uses the monotonic delays assumption to validate a concentration bound that is done in the analysis. See Remark 3 in Appendix B for more details.

an estimate of the Q-function and then applying a closed-form update to the policy in each state. Alternatively, O-REPS methods (Zimin & Neu, 2013) optimize over the state-action occupancy measures instead of directly on policies. This requires solving a global convex optimization problem of size HS^2A (Rosenberg & Mansour, 2019b) in the beginning of each episode, which has no closed-form solution and is extremely inefficient computationally. Another significant shortcoming of O-REPS is the difficulty to scale it to function approximation, since the constrained optimization problem becomes non-convex. In contrast, PO methods extend naturally to function approximation and enjoy great empirical success (e.g., Haarnoja et al. (2018)).

Other than their practical merits, this paper reveals an important theoretical advantage of PO methods over O-REPS – simple update form. We utilize the exponential weights update form of Delayed OPPO in order to investigate the propagation of delayed feedback through the episodes. This results in an intuitive analysis that achieves the best available PO regret bounds even when feedback is delayed. On the other hand, there is very limited understanding regarding the solution for the O-REPS optimization problem, making it very hard to extend beyond its current scope. Specifically, studying the effect of delays on this optimization problem is extremely challenging and takes involved analysis. While we were able to analyze Delayed O-REPS under full-information feedback in Appendix E and give tight regret bounds in Theorem 4, we were not able to extend our analysis to bandit feedback because it involves a complicated in-depth investigation of the difference between any two consecutive occupancy measures chosen by the algorithm. Our analysis bounds this difference under full-information feedback, but in order to bound the regret under bandit feedback its ratio (and the high variance of importance-sampling estimators) must also be bounded. Extending Delayed O-REPS to bandit feedback remains an important open problem, for which our analysis lays the foundations, and is currently the only way that can achieve $\sqrt{K}$ regret in the presence of delays.

Theorem 4. *Running Delayed O-REPS under full-information feedback guarantees the following regret with probability* $1-\delta$. *With known transitions* $\mathcal{R}_K = \widetilde{O}(H\sqrt{K+D})$. *With unknown dynamics, delayed cost feedback and non-delayed trajectory feedback* $\mathcal{R}_K = \widetilde{O}(H^{3/2}S\sqrt{AK}+H\sqrt{D})$.

Stochastic MDPs with delayed feedback. Most RL algorithms focus on stochastic MDPs, that are a special case of adversarial MDPs in which the cost $c_h^k(s,a)$ of episode k is sampled i.i.d from a fixed distribution $C_h(s,a)$ with support $[0,1]$. Thus, studying the effects of delayed feedback on stochastic MDPs is a natural question. With stochastic costs, the OPPO algorithm achieves $\sqrt{K}$ regret even under bandit feedback, since we can replace the importance-sampling estimator with the empirical average. This means that with stochastic costs and bandit feedback, our Delayed OPPO algorithm obtains the same near-optimal regret bounds as under full-information feedback. However, with stochastic costs our lower bound does not hold, suggesting that a $\sqrt{D}$ dependence is not necessary.

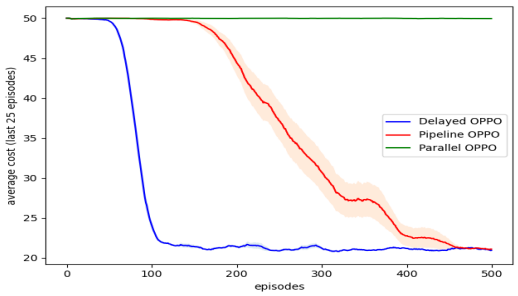

Figure 1. Average cost of delayed algorithms in grid world with geometrically distributed delays.

Indeed, delayed versions of optimistic algorithms (e.g., Zanette & Brunskill (2019)) can utilize our analysis in Section 4.1 to prove regret that does not scale with $\sqrt{D}$ but only with H^2SAd_{max} which can again be improved to H^2Sd_{max} using explicit exploration. This highlights the significance of our analysis and explicit exploration in the presence of delays, as even the UCB algorithm for MAB suffers sub-optimal Ad_{max} regret (Lancewicki et al., 2021), that becomes optimal d_{max} only with explicit exploration.

6. Empirical evaluation

We conducted synthetic experiments to compare the performance of *Delayed OPPO* to two other generic approaches for handling delays: *Parallel-OPPO* running in parallel d_{max} online algorithms, as described in Section 3, and *Pipeline-OPPO* - another simple approach for turning a non-delayed to an algorithm that handles delays by simply waiting for the first d_{max} episodes and then feeding the feedback always with delay d_{max} episodes to the algorithm. We used a simple 10×10 grid world where the agent starts one corner and need to reach the opposite corner, which is the goal state. The cost is 1 in all states except for 0 cost in the goal state. The horizon is $H = 50$ and the delays are drawn i.i.d from a geometric distribution with mean 10. The number of episodes is $K = 500$. The maximum episode delay, d_{max}, is computed on the sequence of realized delays, and it is roughly $10 \log K \approx 60$.

Fig. 1 shows that Delayed OPPO significantly outperforms the other approaches, thus highlighting the importance of handling variable delays and not simply considering the worst-case delay d_{max}. An important note is that, apart from its very high cost, the Parallel-OPPO also requires much more memory (factor d_{max}). For more implementation details and additional experiments, see Appendix F.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), and by the Israel Science Foundation (grant number 993/17).

References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272. JMLR. org, 2017.

Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 35–42. AUAI Press, 2009.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Bistritz, I., Zhou, Z., Chen, X., Bambos, N., and Blanchet, J. Online exp3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, pp. 11349–11358, 2019.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Cesa-Bianchi, N., Gentile, C., Mansour, Y., and Minora, A. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pp. 605–622, 2016.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. Nonstochastic bandits with composite anonymous feedback. In *Conference On Learning Theory*, pp. 750–773, 2018.

Changuel, N., Sayadi, B., and Kieffer, M. Online learning for qoe-based video streaming to mobile receivers. In *2012 IEEE Globecom Workshops*, pp. 1319–1324. IEEE, 2012.

Chen, B., Xu, M., Liu, Z., Li, L., and Zhao, D. Delay-aware multi-agent reinforcement learning. *arXiv preprint arXiv:2005.05441*, 2020a.

Chen, L. and Luo, H. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. *arXiv preprint arXiv:2102.05284*, 2021.

Chen, L., Luo, H., and Wei, C.-Y. Minimax regret for stochastic shortest path with adversarial costs and known transition. *arXiv preprint arXiv:2012.04053*, 2020b.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 5713–5723, 2017.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. Tight regret bounds for model-based reinforcement learning with greedy policies. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 12203–12213, 2019.

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

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *ICML 2018-The 35th International Conference on Machine Learning*, volume 80, pp. 1578–1586, 2018.

Gael, M. A., Vernade, C., Carpentier, A., and Valko, M. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pp. 3348–3356. PMLR, 2020.

György, A. and Joulani, P. Adapting to delays and data in adversarial multi-armed bandits. *arXiv preprint arXiv:2010.06022*, 2020.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870, 2018.

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

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

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pp. 4860–4869. PMLR, 2020a.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143, 2020b.

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

Joulani, P., Gyorgy, A., and Szepesvári, C. Online learning under delayed feedback. In *International Conference on Machine Learning*, pp. 1453–1461, 2013.

Joulani, P., György, A., and Szepesvári, C. A modular analysis of adaptive (non-) convex optimization: Optimism, composite objectives, variance reduction, and variational bounds. *Theoretical Computer Science*, 808: 108–138, 2020.

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

Lancewicki, T., Segal, S., Koren, T., and Mansour, Y. Stochastic multi-armed bandits with unrestricted delay distributions. *arXiv preprint arXiv:2106.02436*, 2021.

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

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. Corruption robust exploration in episodic reinforcement learning. *arXiv preprint arXiv:1911.08689*, 2019.

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

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *stat*, 1050:21, 2009.

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

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

Neu, G., György, A., Szepesvári, C., and Antos, A. Online Markov Decision Processes under bandit feedback. *IEEE Trans. Automat. Contr.*, 59(3):676–691, 2014.

Osband, I. and Van Roy, B. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Osband, I., Van Roy, B., and Wen, Z. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pp. 2377–2386, 2016.

Pike-Burke, C., Agrawal, S., Szepesvari, C., and Grunewalder, S. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pp. 4105–4113. PMLR, 2018.

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

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

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pp. 5478–5486, 2019b.

Rosenberg, A. and Mansour, Y. Adversarial stochastic shortest path. *arXiv preprint arXiv:2006.11561*, 2020.

Rosenberg, A., Cohen, A., Mansour, Y., and Kaplan, H. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pp. 8210–8219. PMLR, 2020.

Schuitema, E., Buşoniu, L., Babuška, R., and Jonker, P. Control delay in reinforcement learning for real-time dynamic systems: a memoryless approach. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3226–3231. IEEE, 2010.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613. PMLR, 2020.

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems*, pp. 1153–1162, 2019.

Simchowitz, M. and Slivkins, A. Exploration and incentives in reinforcement learning. *arXiv preprint arXiv:2103.00360*, 2021.

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

Tarbouriech, J., Garcelon, E., Valko, M., Pirotta, M., and Lazaric, A. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, pp. 9428–9437. PMLR, 2020.

Thune, T. S., Cesa-Bianchi, N., and Seldin, Y. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, pp. 6541–6550, 2019.

Vernade, C., Cappé, O., and Perchet, V. Stochastic bandit models for delayed conversions. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

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

Weinberger, M. J. and Ordentlich, E. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002.

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

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312, 2019.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1964, 2020a.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020b.

Zhou, Z., Xu, R., and Blanchet, J. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*, pp. 5197–5208, 2019.

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

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

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

A. The Delayed OPPO Algorithm

Algorithm 2 Delayed OPPO with known transition function

Input: State space $\mathcal{S}$, Action space $\mathcal{A}$, Horizon H, Transition function p, Learning rate $\eta > 0$, Exploration parameter $\gamma > 0$.

Initialization: Set $\pi_h^1(a \mid s) = 1/A$ for every $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

for $k = 1, 2, \ldots, K$ **do**

Play episode k with policy π^k.

Observe feedback from all episodes $j \in \mathcal{F}^k$.

Policy Evaluation

for $j \in \mathcal{F}^k$ **do**

$\forall s \in \mathcal{S} : V_{H+1}^j(s) = 0$.

for $h = H, \ldots, 1$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**

if bandit feedback **then**

$\hat{c}_h^j(s, a) = \frac{c_h^j(s,a) \cdot \mathbb{I}\{s_h^j = s, a_h^j = a\}}{q_h^{p,\pi^j}(s)\pi_h^j(a|s) + \gamma}$.

else if full-information feedback **then**

$\hat{c}_h^j(s, a) = c_h^j(s, a)$.

end if

$Q_h^j(s, a) = \hat{c}_h^j(s, a) + \langle p_h(\cdot \mid s, a), V_{h+1}^j \rangle$.

$V_h^j(s) = \langle Q_h^j(s, \cdot), \pi_h^j(\cdot \mid s) \rangle$.

end for

end for

Policy Improvement

for $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ **do**

$\pi_h^{k+1}(a|s) = \frac{\pi_h^k(a|s) \exp(-\eta \sum_{j \in \mathcal{F}^k} Q_h^j(s,a))}{\sum_{a' \in \mathcal{A}} \pi_h^k(a'|s) \exp(-\eta \sum_{j \in \mathcal{F}^k} Q_h^j(s,a'))}$.

end for

end for

Algorithm 3 Delayed OPPO with unknown transition function

Input: State space $\mathcal{S}$, Action space $\mathcal{A}$, Horizon H, Learning rate $\eta > 0$, Exploration parameter $\gamma > 0$, Confidence parameter $\delta > 0$, Explicit exploration parameter $\texttt{UseExplicitExploration} \in \{\texttt{true}, \texttt{false}\}$, maximal delay d_{max}.

Initialization: Set $\pi_h^1(a \mid s) = 1/A$ for every $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $\mathcal{K}_{exp} = \emptyset$.

for $k = 1, 2, \ldots, K$ **do**

  for $s \in \mathcal{S}$ **do**

    if $\texttt{UseExplicitExploration} = \texttt{true}$ and $n_h^k(s) \leq 2d_{max} \log \frac{HSA}{\delta}$ **then**

      $\forall a \in \mathcal{A} : \tilde{\pi}_h^k(a \mid s) = 1/A$.

    else

      $\forall a \in \mathcal{A} : \tilde{\pi}_h^k(a \mid s) = \pi_h^k(a \mid s)$.

    end if

  end for

  Play episode k with policy $\tilde{\pi}^k$.

  if trajectory feedback is not delayed **then**

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

  end if

  Observe feedback from all episodes $j \in \mathcal{F}^k$ and update n^{k+1} and $\bar{p}^{k+1}$.

  If explicit exploration was used in some $j \in \mathcal{F}^k$, then add j to $\mathcal{K}_{exp}$.

  Compute confidence set $\mathcal{P}^k = \{\mathcal{P}_h^k(s, a)\}_{s,a,h}$, where $\mathcal{P}_h^k(s, a)$ contains all transition functions $p'_h(\cdot \mid s, a)$ such that for every $s' \in S$,

$$|p'_h(s'|s,a) - \bar{p}_h^k(s'|s,a)| \leq \epsilon_h^k(s'|s,a) = 4\sqrt{\frac{\bar{p}_h^k(s'|s,a)(1-\bar{p}_h^k(s'|s,a))\ln\frac{HSAK}{4\delta}}{n_h^k(s,a)\vee 1}} + 10\frac{\ln\frac{HSAK}{4\delta}}{n_h^k(s,a)\vee 1}.$$

  # Policy Evaluation

  for $j \in \mathcal{F}^k \setminus \mathcal{K}_{exp}$ **do**

    $\forall s \in \mathcal{S} : V_{H+1}^j(s) = 0$.

    for $h = H, \ldots, 1$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**

      if bandit feedback **then**

        $u_h^j(s) = \max_{p' \in \mathcal{P}^j} q_h^{p', \pi^j}(s) = \max_{p' \in \mathcal{P}^j} \Pr[s_h = s \mid s_1 = s_{\text{init}}, \pi^j, p']$.

        $\hat{c}_h^j(s,a) = \frac{c_h^j(s,a) \cdot \mathbb{I}\{s_h^j = s, a_h^j = a\}}{u_h^j(s)\pi_h^j(a|s) + \gamma}$.

        $\hat{p}_h^j(\cdot|s,a) \in \underset{p'_h(\cdot|s,a) \in \mathcal{P}_h^j(\cdot|s,a)}{\arg\min} \langle p'_h(\cdot|s,a), V_{h+1}^j \rangle$.

      else if full-information feedback **then**

        $\hat{c}_h^j(s,a) = c_h^j(s,a)$.

        $\hat{p}_h^j(\cdot|s,a) \in \underset{p'_h(\cdot|s,a) \in \mathcal{P}_h^k(\cdot|s,a)}{\arg\min} \langle p'_h(\cdot|s,a), V_{h+1}^j \rangle$.

      end if

      $Q_h^j(s,a) = \hat{c}_h^j(s,a) + \langle \hat{p}_h^j(\cdot \mid s,a), V_{h+1}^j \rangle$.

      $V_h^j(s) = \langle Q_h^j(s,\cdot), \pi_h^j(\cdot \mid s) \rangle$.

    end for

  end for

  # Policy Improvement

  for $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ **do**

    $\pi_h^{k+1}(a|s) = \frac{\pi_h^k(a|s)\exp(-\eta \sum_{j \in \mathcal{F}^k \setminus \mathcal{K}_{exp}} Q_h^j(s,a))}{\sum_{a' \in \mathcal{A}} \pi_h^k(a'|s)\exp(-\eta \sum_{j \in \mathcal{F}^k \setminus \mathcal{K}_{exp}} Q_h^j(s,a'))}$.

  end for

end for

B. Full proofs of main theorems

We start by defining failure events. The rest of the analysis focuses on the good event: the event in which non of the failure events occur. We show that in order to guarantee a regret bound that holds with probability of $1-\delta$, we only need to pay a factor which is logarithmic in $1/\delta$. In Appendix B.1 we define the failure events and bound their probability in Lemmas 1 and 3.

In Appendix B.2 we prove the regret bound for the most challenging case: unknown transition function + delayed trajectory feedback + bandit feedback, that is, Theorem 3 with bandit feedback. Then, the rest of the proofs follow easily as corollaries.

In Appendix B.3 we prove the regret for the case of unknown transition function + delayed trajectory feedback + full-information feedback, that is, Theorem 3 with full-information feedback.

In Appendix B.4 we prove the regret for the case of unknown transition function but with non-delayed trajectory feedback, that is, Theorem 1. Finally, in Appendix B.5 we prove the regret for the case of known transition function, that is, Theorem 2.

Remark 1. *The analysis for bandit feedback uses estimated Q-functions Q^j that were computed with the confidence set $\mathcal{P}^j$ and not $\mathcal{P}^{j+d^j}$, as in the full-information case. This simplifies concentration arguments but ignores a lot of data. It also means that under bandit feedback the algorithm has worse space complexity, since we may need to keep up to d_{max} empirical transition functions. However, the space complexity can be easily reduced by re-computing the confidence sets only when the number of visits to some state-action pair is doubled, and not in the end of every episode.*

B.1. Failure events

Fix some probability δ'. We now define basic failure events:

- $F_1^{basic} = \left\{\exists k, s', s, a, h : \left|p_h(s' \mid s,a) - \bar{p}_h^k(s' \mid s,a)\right| > \epsilon_h^k(s' \mid s,a)\right\}$
  - ▶ where $\epsilon_h^k(s' \mid s,a) = 4\sqrt{\frac{\bar{p}_h^k(s'|s,a)(1-\bar{p}_h^k(s'|s,a))\ln\frac{HSAK}{4\delta}}{n_h^k(s,a)\vee 1}} + 10\frac{\ln\frac{HSAK}{4\delta}}{n_h^k(s,a)\vee 1}$.
- $F_2^{basic} = \left\{\exists k, s, a, h : \left\|p_h(\cdot \mid s,a) - \bar{p}_h^k(\cdot \mid s,a)\right\|_1 > \sqrt{\frac{14S\ln\left(\frac{HSAK}{\delta'}\right)}{n_h^k(s,a)\vee 1}}\right\}$
- $F_3^{basic} = \left\{\sum_{k,s,a,h}\left(q_h^k(s,a) - \mathbb{I}\{s_h^k = s, a_h^k = a\}\right)\min\{2, r_h^k(s,a)\} > 6\sqrt{K\ln\frac{1}{\delta}}\right\}$
  - ▶ where $r_h^k(s,a) = 8\sqrt{\frac{S\ln\frac{HSAK}{4\delta'}}{n_h^k(s,a)\vee 1}} + 200S\ln\frac{\ln\frac{HSAK}{4\delta'}}{n_h^k(s,a)\vee 1}$, $q_h^k(s,a) = q_h^{p,\pi^k}(s)\pi_h^k(a|s)$, and $q_h^{p,\pi}(s) = \Pr[s_h = s \mid s_1 = s_{\text{init}}, \pi, p]$.
- $F_4^{basic} = \left\{\exists k, s, a, h : \sum_{k'=1}^k \hat{c}_h^{k'}(s,a) - \frac{q_h^{k'}(s)}{u_h^{k'}(s)}c_h^{k'}(s,a) > \frac{\ln\frac{SAHK}{\delta'}}{2\gamma}\right\}$
- $F_5^{basic} = \{\exists k, s, a, h : n_h^k(s) \geq d_{max}\log\frac{HSA}{\delta} \text{ and } n_h^k(s,a) \leq \frac{d_{max}}{2A}\} \cap \{UseExplicitExploration = \texttt{true}\}$
  - ▶ where $n_h^k(s) = \sum_{a'} n_h^k(s,a')$.

We define the basic good by $G^{basic} = \bigcap_{i=1}^5 \neg F_i^{basic}$.

Lemma 1. *The basic good event G^{basic}, occurs with probability of at least $1-5\delta'$.*

Proof.

- By (Maurer & Pontil, 2009, Theorem 4), $\Pr\left(F_1^{basic}\right) \leq \delta'$.
- By (Jaksch et al., 2010, Lemma 17) and union bounds, $\Pr\left(F_2^{basic}\right) \leq \delta'$.

- Let $Y_k = \sum_{s,a,h}\left(q_h^k(s,a) - \mathbb{I}\{s_h^k = s, a_h^k = a\}\right)\min\left(2, r_h^k(s,a)\right)$. Note that r_h^k depends on the history up to the end of episode $k-1$, $\mathcal{H}_{k-1}$. Therefore, $\mathbb{E}[Y_k \mid \mathcal{H}_{k-1}] = 0$ (as $\mathbb{E}\left[\mathbb{I}\{s_h^k = s, a_h^k = a\} \mid H_{k-1}\right] = q_h^k(s,a)$). That is, $\sum_k Y_k$ is a martingale. Also, $|Y_k| \le 4$. By Azuma-Hoeffding inequality,

$$\Pr\left(F_3^{basic}\right) = \Pr\left(\sum_k Y_k > 6\sqrt{K \ln 1/\delta'}\right) \le \delta'.$$

- By (Shani et al., 2020, Lemma 6), $\Pr\left(F_4^{basic}\right) \le \delta'$.[3]

- Fix h, s, a and k such that $n_h^k(s) \ge d_{max} \log \frac{HSA}{\delta'}$, and let k_0 be the first episode such that $n_h^{k_0}(s) \ge d_{max} \log \frac{HSA}{\delta'}$. Since, for at least the first $d_{max} \log \frac{HSA}{\delta'}$ visits in s, we play a uniform policy in s, $\mathbb{E}\left[n_h^{k_0}(s,a)\right] \ge \frac{d_{max}}{A} \log \frac{HSA}{\delta'}$. By Chernoff bound,

$$\Pr\left(n_h^{k_0}(s,a) \ge \frac{1}{2}\frac{d_{max}}{A}\right) \le e^{-\frac{1}{8}\frac{d_{max}}{A}\log\frac{HSA}{\delta'}} \le \frac{\delta'}{HSA},$$

  where the last holds whenever $d_{max} \ge 8A$ (if not, we can actually get better regret bounds). Taking the union bound over h, s, a and noting that it is sufficient to show the claim for the first k that satisfies $n_h^k(s) \ge d_{max} \log \frac{HSA}{\delta'}$, gives us

$$\Pr(\neg F_5^{basic}) \le \delta'.$$

Using the union bound on the above failure events and taking the complement gives us the desired result.

□

Define $\tilde{\epsilon}_h^k(s' \mid s,a) = 8\sqrt{\frac{p_h(s'|s,a)(1-p_h(s'|s,a))\ln\frac{HSAK}{4\delta}}{n_h^k(s,a)\vee 1}} + 100\ln\frac{\ln\frac{HSAK}{4\delta}}{n_h^k(s,a)\vee 1}$.

Lemma 2. *Given the basic good event, the following relations holds:*

1. $\left|p_h(s' \mid s,a) - \hat{p}_h^k(s' \mid s,a)\right| \le \tilde{\epsilon}_h^k(s' \mid s,a)$.
2. $\sum_{s'}\left|p_h(s' \mid s,a) - \hat{p}_h^k(s' \mid s,a)\right| \le r_h^k(s,a)$.

Proof.

1. Under bandit feedback, the first inequality now holds by $\neg F_1^{basic}$ and (Rosenberg et al., 2020, Lemma B.13). Under full information, recall that $\hat{p}^k$ is computed after episode $k + d^k$. That is, $\hat{p}^k \in \mathcal{P}^{k+d^k}$. Similar to the bandit case,

$$\left|p_h(s' \mid s,a) - \hat{p}_h^k(s' \mid s,a)\right| \le \tilde{\epsilon}_h^{k+d^k}(s' \mid s,a) \le \tilde{\epsilon}_h^k(s' \mid s,a),$$

   where the second inequality is since $\tilde{\epsilon}_h^k(s' \mid s,a)$ is decreasing in k.

2. The second relation simply holds by the first relation and Jensen's inequality.

□

We define the following bad events that will not occur with high probability, given that the basic good event occurs:

- $F_1^{cond} = \left\{\sum_{k,s,a,h} q_h^k(s)\pi_h^k(s \mid a)\left(E\left[\hat{c}_h^k(s,a) \mid \mathcal{H}^{k-1}\right] - \hat{c}_h^k(s,a)\right) > H\sqrt{K\frac{\ln H}{2\delta'}}\right\}$
  - ▶ where $\mathcal{H}^{k-1}$ denotes the history up to episode k.

[3] We have invoked Lemma 6 of (Shani et al., 2020) with $\alpha_h^k(s',a') = \mathbb{I}\{s' = s, a' = a\}$ and then take union bound over all s, a and h.

- $F_2^{cond} = \left\{\exists h, s : \sum_{k=1}^{K} V_h^k(s) - V_h^{\pi^k}(s) > \frac{H}{\gamma} \ln \frac{HSK}{\delta'}\right\}$
- $F_3^{cond} = \left\{\exists s, a : \sum_{k=1}^{K} \sum_h \mathbb{E}^{\pi^k, s}\left[\tilde{\epsilon}_h^k(\cdot \mid s, a)(V_{h+1}^k - V_{h+1}^{\pi^k})\right] > \frac{H^2 S}{2\gamma} \ln \frac{H^2 SK}{\delta'}\right\}$
- $F_4^{cond} = \left\{\exists h, s : \sum_k \left|\{j : j \le k, j + d^j \ge k\}\right| \left(V_h^k(s) - V_h^{\pi^k}(s)\right) > d_{max} \frac{H}{\gamma} \ln \frac{HSK}{\delta'}\right\}$

We define the conditioned good event by $G^{cond} = \neg F_1^{cond} \cap \neg F_2^{cond} \cap \neg F_3^{cond} \cap \neg F_4^{cond}$.

Lemma 3. *Conditioned on G^{basic}, the conditioned good event G^{cond}, occurs with probability of at least $1 - 4\delta'$. That is,*

$$\Pr\left(G^{cond} \mid G^{basic}\right) \ge 1 - 4\delta'.$$

Proof.

- Following the proof in (Shani et al., 2020, appendix C.1.3), $\Pr\left(F_i^{cond} \mid G^{basic}\right) \le \delta'$ for $i = 1, 2, 3$.
- By (Shani et al., 2020, Lemmas 7 and 10) and the fact that $\left|\{j : j \le k, j + d^j \ge k\}\right| \le d_{max}$, $\Pr\left(F_4^{cond} \mid G^{basic}\right) \le \delta'$.

Using the union bound on the above failure events and taking the complement gives us the desired result. □

Finally, we define the global good event, $G := G^{basic} \cap G^{cond}$. Using the union bound, for $\delta > 0$ and $\delta' = \delta/9$, $P(G) \ge 1 - \delta$.

B.2. Proof of Theorem 3 with Bandit Feedback

With probability at least $1 - \delta$ the good event holds. For now on, we assume we are outside the failure event, and then our regret bound holds with probability at least $1 - \delta$.

According to the value difference lemma of Shani et al. (2020),

$$\begin{aligned}
\mathcal{R}_K &= \sum_{k=1}^{K} V_1^{\pi^k}(s_1^k) - V_1^{\pi}(s_1^k) \\
&= \sum_{k=1}^{K} V_1^{\pi^k}(s_1^k) - V_1^k(s_1^k) + \sum_{k=1}^{K} V_1^k(s_1^k) - V_1^{\pi}(s_1^k) \\
&= \underbrace{\sum_{k=1}^{K} V_1^{\pi^k}(s_1^k) - V_1^k(s_1^k)}_{(A)} \\
&\quad + \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}^{\pi}\left[\langle Q_h^k(s_h^k, \cdot), \pi_h^k(\cdot \mid s_h^k) - \pi_h(\cdot \mid s_h^k)\rangle\right]}_{(B)} \\
&\quad + \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}^{\pi}\left[Q_h^k(s_h^k, a_h^k) - c_h^k(s_h^k, a_h^k) - \langle p_h(\cdot \mid s_h^k, a_h^k), V_{h+1}^k\rangle\right]}_{(C)}.
\end{aligned} \tag{3}$$

We continue bounding each of these terms separately. Term (A) is bounded in Appendix B.2.1 by $\widetilde{O}\left(H^2 S\sqrt{AK} + H^2 SAd_{max} + \gamma KHSA + \frac{H^2 S}{\gamma} + H^2 S^2 A\right)$, Term (B) is bounded in Appendix B.2.2 by $\widetilde{O}\left(\frac{H}{\eta} + \frac{\eta}{\gamma} H^3(K + D) + \frac{\eta}{\gamma^2} d_{max} H^3\right)$

and Term (C) is bounded in Appendix B.2.3 by $\widetilde{O}\left(\frac{H}{\gamma}\right)$. This gives a total regret bound of

$$\mathcal{R}_K = \widetilde{O}\Big(H^2SAd_{max} + H^2S\sqrt{AK} + \frac{H^2S}{\gamma} + \gamma KHSA + \frac{H}{\eta} + \frac{\eta}{\gamma}H^3(K+D) + \frac{\eta}{\gamma^2}d_{max}H^3 + \frac{H}{\gamma} + H^2S^2A\Big).$$

Choosing $\eta = \frac{1}{H(A^{3/2}K+D)^{2/3}}$ and $\gamma = \frac{1}{(A^{3/2}K+D)^{1/3}}$ gives the theorem's statement.

Remark 2 (Delayed OPPO with stochastic costs under bandit feedback)**.** *As shown by Shani et al. (2020), when the costs are stochastic, we can replace the importance-sampling estimator with a simple empirical average. This means that our estimator is now bounded by H and eliminates all the terms that depend on γ. Thus, we obtain a regret of $\widetilde{O}(H^2S\sqrt{AK} + H^2\sqrt{D} + H^2S^2A + H^2SAd_{max})$ that is similar to the full-information feedback case. Moreover, in this case we can again use explicit exploration to reduce the last term to $\widetilde{O}(H^2Sd_{max})$*

B.2.1. Bounding Term (A)

Lemma 4. *Conditioned on the good event G,*

$$\sum_{k=1}^{K} V_1^{\pi^k}(s_1^k) - V_1^k(s_1^k) = \widetilde{O}\Big(H^2S\sqrt{AK} + H^2S^2A + \frac{H^2S}{\gamma} + \gamma KHSA + H^2SAd_{max}\Big).$$

Proof. We start with a value difference lemma (Shani et al., 2020),

$$\begin{aligned}
\sum_{k=1}^{K} V_1^{\pi^k}(s_1^k) - V_1^k(s_1^k) &= \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}^{\pi^k}\left[c_h^k(s_h^k,a_h^k) - \hat{c}_h^k(s_h^k,a_h^k)\right] \\
&\quad + \mathbb{E}^{\pi^k}\left[\langle p_h(\cdot \mid s_h^k,a_h^k) - \hat{p}_h^k(\cdot \mid s_h^k,a_h^k), V_{h+1}^k\rangle\right] \\
&\le \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}^{\pi^k}\left[c_h^k(s_h^k,a_h^k) - \hat{c}_h^k(s_h^k,a_h^k)\right]}_{(A.1)} \\
&\quad + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}} \mathbb{E}^{\pi^k}\left[|p_h(s' \mid s_h^k,a_h^k) - \hat{p}_h^k(s' \mid s_h^k,a_h^k)|V_{h+1}^{\pi^k}(s')\right]}_{(A.2)} \\
&\quad + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}^{\pi^k}\left[\langle \bar{\epsilon}_h^k(\cdot \mid s_h,a_h), V_{h+1}^k - V_{h+1}^{\pi^k}\rangle\right]}_{(A.3)},
\end{aligned}$$

where we have used Lemma 2 for the inequality. For any k, h, s and a. we have

$$\begin{aligned}
c_h^k(s,a) - \hat{c}_h^k(s,a) &= c_h^k(s,a) - \mathbb{E}\left[\hat{c}_h^k(s,a) \mid \mathcal{H}^{k-1}\right] + \mathbb{E}\left[\hat{c}_h^k(s,a) \mid \mathcal{H}^{k-1}\right] - \hat{c}_h^k(s,a) \\
&= c_h^k(s,a)\left(1 - \frac{q_h^k(s)\pi_h^k(a \mid s)}{u_h^k(s)\pi_h^k(a \mid s) + \gamma}\right) + \mathbb{E}\left[\hat{c}_h^k(s,a) \mid \mathcal{H}^{k-1}\right] - \hat{c}_h^k(s,a) \\
&= c_h^k(s,a)\left(\frac{(u_h^k(s) - q_h^k(s))\pi_h^k(a \mid s) + \gamma}{u_h^k(s)\pi_h^k(a \mid s) + \gamma}\right) \\
&\quad + \mathbb{E}\left[\hat{c}_h^k(s,a) \mid \mathcal{H}^{k-1}\right] - \hat{c}_h^k(s,a).
\end{aligned}$$

Therefore (denoting $q_h^k(s) = q_h^{p,\pi^k}(s)$),

$$
\begin{aligned}
(A.1) &= \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}^{\pi^k}\left[c_h^k(s_h,a_h)\left(\frac{(u_h^k(s_h)-q_h^k(s_h))\pi_h^k(a_h\mid s_h)+\gamma}{u_h^k(s_h)\pi_h^k(a_h\mid s_h)+\gamma}\right)\right] \\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}^{\pi^k}\left[\mathbb{E}\left[\hat{c}_h^k(s_h,a_h)\mid\mathcal{H}^{k-1}\right]-\hat{c}_h^k(s_h,a_h)\right] \\
&= \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a}q_h^k(s)\pi_h^k(s\mid a)c_h^k(s,a)\left(\frac{(u_h^k(s)-q_h^k(s))\pi_h^k(a\mid s)+\gamma}{u_h^k(s)\pi_h^k(a\mid s)+\gamma}\right)}_{(A.1.1)} \\
&\quad + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a}q_h^k(s)\pi_h^k(s\mid a)\left(\mathbb{E}[\hat{c}_h^k(s,a)\mid\mathcal{H}^{k-1}]-\hat{c}_h^k(s,a)\right)}_{(A.1.2)}.
\end{aligned}
$$

Under the good event $p\in\mathcal{P}^{k-1}$. Hence by definition $u_h^k(s)\geq q_h^k(s)$. Therefore,

$$
\begin{aligned}
(A.1.1) &\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a}q_h^k(s)\pi^k(s\mid a)c_h^k(s,a)\left(\frac{(u_h^k(s)-q_h^k(s))\pi_h^k(a\mid s)+\gamma}{q_h^k(s)\pi_h^k(a\mid s)}\right) \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a}c_h^k(s,a)\left((u_h^k(s)-q_h^k(s))\pi_h^k(a\mid s)+\gamma\right) \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s}\left(u_h^k(s)-q_h^k(s)\right)+\gamma KHSA \\
&\leq HS\sqrt{AK}+HSAd_{max}+\gamma KHSA,
\end{aligned}
$$

where the last inequality follows similarly to the bound of Term (A.2) when combined with Lemma 20 of Shani et al. (2020).

Under the good event G (in particular, $\neg F_1^{cond}$),

$$
(A.1.2)\leq H\sqrt{K\frac{\ln H}{2\delta'}}.
$$

By Lemma 5 and the fact that $V_{h+1}^{\pi^k}\leq H$,

$$
(A.2)\leq H^2S\sqrt{AK}+H^2SAd_{max}+H^2S^2A.
$$

Finally, conditioned on the good event ($\neg F_3^{cond}$),

$$
(A.3)\leq\frac{H^2S}{2\gamma}\ln\frac{H^2SK}{\delta'}.
$$

We get that term (A) can be bounded by,

$$
\begin{aligned}
\sum_{k=1}^{K}V_1^{\pi^k}(s_1^k)-V_1^k(s_1^k) &\leq (A.1.1)+(A.1.2)+(A.2)+(A.3) \\
&\lesssim H^2S\sqrt{AK}+H^2SAd_{max}+\gamma KHSA+\frac{H^2S}{\gamma},
\end{aligned}
$$

where $\lesssim$ ignores poly-logarithmic factors. □

Lemma 5. *Under the good event,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}}\mathbb{E}^{\pi^k}\left[\left|p_h(s'\mid s_h^k,a_h^k)-\hat{p}_h^k(s'\mid s_h^k,a_h^k)\right|\right]\lesssim HS\sqrt{AK}+HSAd_{max}+HS^2A.$$

Proof. Define,

$$\mathcal{K}(s,a,h)=\left\{k\in[K]: s_h^k=s, a_h^k=a, \sum_{j:j+d^j\le k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}\le d_{max}\right\},$$

and note that $\mathcal{K}(s,a,h)\le 2d_{max}$. Thus,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}}\mathbb{E}^{\pi^k}\left[\left|p_h(s'\mid s_h^k,a_h^k)-\hat{p}_h^k(s'\mid s_h^k,a_h^k)\right|\right]=\\
&\quad=\sum_{k=1}^{K}\sum_{s,a,h}q_h^k(s,a)\sum_{s'\in\mathcal{S}}\left|p_h(s'\mid s,a)-\hat{p}_h^k(s'\mid s,a)\right|\\
&\quad\le\sum_{k=1}^{K}\sum_{s,a,h}q_h^k(s,a)\min\{2,r_h^k(s,a)\}\\
&\quad\le\sum_{k=1}^{K}\sum_{s,a,h}\left(q_h^k(s,a)-\mathbb{I}\{s_h^k=s,a_h^k=a\}\right)\min\{2,r_h^k(s,a)\}\\
&\qquad+2\sum_{s,a,h}\sum_{k\in\mathcal{K}(s,a,h)}\mathbb{I}\{s_h^k=s,a_h^k=a\}\\
&\qquad+\sum_{s,a,h}\sum_{k\notin\mathcal{K}(s,a,h)}\mathbb{I}\{s_h^k=s,a_h^k=a\}r_h^k(s,a)\\
&\quad\lesssim\sqrt{K}+HSAd_{max}+\sum_{s,a,h}\sum_{k\notin\mathcal{K}(s,a,h)}\mathbb{I}\{s_h^k=s,a_h^k=a\}r_h^k(s,a)\\
&\quad\lesssim\sqrt{K}+HSAd_{max}+\sqrt{S}\sum_{s,a,h}\sum_{k\notin\mathcal{K}(s,a,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{n_h^k(s,a)\vee 1}}\\
&\qquad+S\sum_{s,a,h}\sum_{k\notin\mathcal{K}(s,a,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{n_h^k(s,a)\vee 1}.
\end{aligned}\tag{4}$$

The first inequality follows the fact that $\|p_h(\cdot\mid s,a)-\hat{p}_h^k(\cdot\mid s,a)\|_1\le 2$ and Lemma 2. And the third inequality is by $\neg F_3^{basic}$, and $|\mathcal{K}(s,a,h)|\le d_{max}$. Now,

$$\begin{aligned}
&\sum_{k\notin\mathcal{K}(s,a,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{n_h^k(s,a)\vee 1}}\le\\
&\le\sum_{k\notin\mathcal{K}(s,a,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\cdot\sqrt{\frac{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}{1\vee\sum_{j:j+d^j\le k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\\
&\le\sum_{k\notin\mathcal{K}(s,a,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\cdot\sqrt{1+\frac{1\vee\sum_{j:j<k,j+d^j\ge k}\mathbb{I}\{s_h^j=s,a_h^j=a\}}{1\vee\sum_{j:j+d^j\le k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}.
\end{aligned}$$

Since for any $k\notin\mathcal{K}(s,a,h)$,

$$\sum_{j:j<k,j+d^j\ge k}\mathbb{I}\{s_h^j=s,a_h^j=a\}\le d_{max}\le\sum_{j:j+d^j\le k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}$$

we have

$$\sum_{k\notin\mathcal{K}(s,a,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{n_h^k(s,a)\vee 1}}\lesssim\sum_{k\notin\mathcal{K}(s,a,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}$$
$$\leq\sum_{k=1}^{K}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}. \tag{5}$$

In the same way,

$$\sum_{k\notin\mathcal{K}(s,a,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{n_h^k(s,a)\vee 1}\lesssim\sum_{k=1}^{K}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}. \tag{6}$$

Finally, (5) and (6) appear in the non-delayed setting and can be bounded when summing over (s,a,h) by $\tilde{O}(H\sqrt{SAK})$ and $\tilde{O}(HSA)$, respectively (see for example ,(Jin et al., 2020a, Lemma 4)). Plugging back in (4) compleats the proof. □

B.2.2. Bounding Term (B)

Lemma 6. *Let $s\in S$. Conditioned on the good event G,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left\langle Q_h^k(s,\cdot),\pi_h^k(\cdot\mid s)-\pi_h(\cdot\mid s)\right\rangle\leq\frac{H\log(A)}{\eta}+\eta\frac{H^3}{\gamma}(K+D)+\frac{\eta}{\gamma^2}d_{max}H^3\ln\frac{H}{\delta}.$$

Proof. We adopt the technique presented in (György & Joulani, 2020) for MAB, in order to bound the term $\sum_k\langle Q_h^k(s,\cdot),\pi_h^k(\cdot\mid s)-\pi_h(\cdot\mid s)\rangle$, for each s and h separately.

The proof uses a comparison to a "cheating" algorithm that does not experience delay and sees onoe step into the future. Define

$$\bar{\pi}_h^k(a\mid s)=\frac{\exp\left(-\eta\sum_{j:j\leq k-1}Q_h^j(s,a)\right)}{\sum_{a'\in\mathcal{A}}\exp\left(-\eta\sum_{j:j\leq k-1}Q_h^j(s,a')\right)}.$$

We break the sum term in the following way:

$$\sum_k\left\langle Q_h^k(s,\cdot),\pi_h^k(\cdot\mid s)-\pi_h(\cdot\mid s)\right\rangle=\underbrace{\sum_k\left\langle Q_h^k(s,\cdot),\bar{\pi}_h^{k+1}(\cdot\mid s)-\pi_h(\cdot\mid s)\right\rangle}_{(B.1)}$$
$$+\underbrace{\sum_k\left\langle Q_h^k(s,\cdot),\pi_h^k(\cdot\mid s)-\bar{\pi}_h^{k+1}(\cdot\mid s)\right\rangle}_{(B.2)}.$$

Term (B.1) is the regret of the "cheating" algorithm. Using (Joulani et al., 2020, Theorem 3),[4]

$$(B.1)\leq\frac{\ln(A)}{\eta}. \tag{7}$$

[4]We choose the regularizers in (Joulani et al., 2020) to be $q_0(x)=r_1(x)=\frac{1}{\eta}\sum_i x_i\log x_i$ and the rest are zero, which makes the update of their ADA-MD algorithm as in our policy improvement step. The statement now follows from (Joulani et al., 2020, Theorem 3), the fact that the Bregman divergence is positive and that entropy is bounded by $\log A$.

Now, using the definition of $\bar{\pi}_h^k$,

$$\begin{aligned}
\frac{\bar{\pi}_h^{k+1}(a \mid s)}{\pi_h^k(a \mid s)} &= \frac{\exp\left(-\eta \sum_{j:j\le k} Q_h^j(s,a)\right)}{\sum_{a'} \exp\left(-\eta \sum_{j:j\le k} Q_h^j(s,a')\right)} \cdot \frac{\sum_{a'} \exp\left(-\eta \sum_{j:j+d^j\le k-1} Q_h^j(s,a')\right)}{\exp\left(-\eta \sum_{j:j+d^j\le k-1} Q_h^j(s,a)\right)} \\
&= \exp\left(-\eta \sum_{j:j\le k, j+d^j\ge k} Q_h^j(s,a)\right) \cdot \frac{\sum_{a'} \exp\left(-\eta \sum_{j:j+d^j\le k-1} Q_h^j(s,a')\right)}{\sum_{a'} \exp\left(-\eta \sum_{j:j\le k} Q_h^j(s,a')\right)} \\
&\ge \exp\left(-\eta \sum_{j:j\le k, j+d^j\ge k} Q_h^j(s,a)\right) \\
&\ge 1 - \eta \sum_{j:j\le k, j+d^j\ge k} Q_h^j(s,a)
\end{aligned}$$

where in the first inequality we have used $\sum_{j:j+d^j\le k-1} Q_h^j(s,a) \le \sum_{j:j\le k} Q_h^j(s,a')$, and for the second inequality we have used the fact that $e^x \ge 1+x$ for any x. Using the above,

$$\begin{aligned}
(B.2) &= \sum_k \left\langle Q_h^k(s,\cdot), \pi_h^k(\cdot \mid s) - \bar{\pi}_h^{k+1}(\cdot \mid s)\right\rangle \\
&= \sum_k \sum_{a\in\mathcal{A}} Q_h^k(s,a)\left(\pi_h^k(a \mid s) - \bar{\pi}_h^{k+1}(a \mid s)\right) \\
&= \sum_k \sum_{a\in\mathcal{A}} Q_h^k(s,a)\left(\pi_h^k(a \mid s)\left(1 - \frac{\bar{\pi}_h^{k+1}(a \mid s)}{\pi_h^k(a \mid s)}\right)\right) \\
&\le \eta \sum_k \sum_{a\in\mathcal{A}} \pi_h^k(a \mid s) Q_h^k(s,a) \sum_{j:j\le k, j+d^j\ge k} Q_h^j(s,a) \\
&\le \eta \sum_k \sum_{a\in\mathcal{A}} \pi_h^k(a \mid s) Q_h^k(s,a) \sum_{j:j\le k, j+d^j\ge k} \left(\hat{c}_h^j(s,a) + \langle p(\cdot \mid s,a), V_{h+1}^j\rangle\right) \\
&\le \eta \sum_k \sum_{a\in\mathcal{A}} \pi_h^k(a \mid s) Q_h^k(s,a) \sum_{j:j\le k, j+d^j\ge k} \frac{H}{\gamma} \qquad (\hat{c}_h^j(s,a), V_{h+1}^j(s) \le \tfrac{H}{\gamma}) \\
&\le \eta \frac{H}{\gamma} \sum_k |\{j : j\le k, j+d^j\ge k\}| \sum_{a\in\mathcal{A}} \pi_h^k(a \mid s) Q_h^k(s,a) \\
&= \eta \frac{H}{\gamma} \sum_k |\{j : j\le k, j+d^j\ge k\}| V_h^k(s).
\end{aligned}$$

Under the good event (in particular, $\neg F_4^{cond}$),

$$\sum_k |\{j : j\le k, j+d^j\ge k\}| \left(V_h^k(s) - V_h^{\pi^k}(s)\right) \le d_{max} \frac{H}{\gamma} \ln\frac{H}{\delta}.$$

Therefore,

$$\begin{aligned}
(B.2) &\le \eta \frac{H}{\gamma} \sum_k |\{j : j\le k, j+d^j\ge k\}| V_h^{\pi^k}(s) + \frac{\eta}{\gamma^2} d_{max} H^2 \ln\frac{H}{\delta} \\
&\le \eta \frac{H^2}{\gamma} \sum_k |\{j : j\le k, j+d^j\ge k\}| + \frac{\eta}{\gamma^2} d_{max} H^2 \ln\frac{H}{\delta} \\
&\le \eta \frac{H^2}{\gamma}(K+D) + \frac{\eta}{\gamma^2} d_{max} H^2 \ln\frac{H}{\delta}. \qquad (8)
\end{aligned}$$

where the last inequality holds since,

$$\sum_k |\{j : j\le k, j+d^j\ge k\}| = \sum_j \sum_k \mathbb{I}\{j \le k \le j+d^j\} \le \sum_j (1+d^j) \le K+D.$$

Combining (7) and (8) and summing over h, completes the proof. □

B.2.3. Bounding Term (C)

Lemma 7. *Conditioned on the good event G,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}^{\pi}\left[Q_h^k(s_h^k,a_h^k)-c_h^k(s_h^k,a_h^k)-\left\langle p_h(\cdot\mid s_h^k,a_h^k),V_{h+1}^k\right\rangle\right]\leq\frac{H}{2\gamma}\log\frac{SAHK}{\delta'}.$$

Proof. Using Bellman equations,

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}^{\pi}\left[Q_h^k(s_h^k,a_h^k)-c_h^k(s_h^k,a_h^k)-\left\langle p_h(\cdot\mid s_h^k,a_h^k),V_{h+1}^k\right\rangle\right]&=\\ =\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}^{\pi}\left[\hat{c}_h^k(s_h^k,a_h^k)-c_h^k(s_h^k,a_h^k)\right]}_{(C.1)}&\\ +\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}^{\pi}\left[\left\langle\hat{p}_h^k(\cdot\mid s_h^k,a_h^k),V_{h+1}^k\right\rangle-\left\langle p_h(\cdot\mid s_h^k,a_h^k),V_{h+1}^k\right\rangle\right]}_{(C.2)}.&\end{aligned}$$

For any h, s and a, under the good event,

$$\sum_{k=1}^{K}\hat{c}_h^k(s,a)-c_h^k(s,a)\leq\sum_{k=1}^{K}\hat{c}_h^k(s,a)-\frac{q_h^k(s)}{u_h^k(s)}c_h^k(s,a)\leq\frac{\log(\frac{SAHK}{\delta'})}{2\gamma},$$

where the first inequality is due to the fact that under the good event $p\in\mathcal{P}^{k-1}$ and so $u_h^k(s)=\max_{\hat{p}\in\mathcal{P}^{k-1}}q_h^{\hat{p},\pi^k}(s)\geq q_h^{p,\pi^k}(s)=q_h^k(s)$. The second inequality follows directly from $\neg F_4^{basic}$. Therefore,

$$(C.1)\leq\frac{H}{2\gamma}\log\frac{SAHK}{\delta'}.$$

Once again, since under the good event $p\in\mathcal{P}^k$, then for all h, s and a,

$$\left\langle\hat{p}_h^k(\cdot\mid s,a),V_{h+1}^k\right\rangle=\min_{\hat{p}_h(\cdot|s,a)\in\mathcal{P}^k}\left\langle\hat{p}_h^k(\cdot\mid s,a),V_{h+1}^k\right\rangle\leq\left\langle p_h(\cdot\mid s,a),V_{h+1}^k\right\rangle.$$

Therefore $(C.2)\leq 0$, which completes the proof of the lemma. □

B.3. Proof of Theorem 3 under full-information feedback

The proof follows almost immediately from the proof in Appendix B.2, by noting that some of the terms become zero since we use the actual cost function and not an estimated one. In addition, in this setting we use the explicit exploration which yield a better bound on term (A).

Let $\mathcal{K}_{exp}(s,h)$ be the episodes in which we used the uniform policy in state s at time h, because we did not receive enough feedback from that state. That is,

$$\mathcal{K}_{exp}(s,h)=\left\{k\in[K]:s_h^k=s,n_h^k(s)\leq d_{max}\log\frac{HSA}{\delta}\right\}.$$

Also, define $\mathcal{K}_{exp} = \bigcup_{s,h} \mathcal{K}_{exp}(s,h)$. Recall that the algorithm keeps track of $\mathcal{K}_{exp}$, and preforms the policy improvement step only with respect to rounds that are not in $\mathcal{K}_{exp}$. For any (k,s,h) we have that $m_h^k(s) - n_h^k(s) \le d_{max}$ and thus,

$$\begin{aligned}|\mathcal{K}_{exp}(s,h)| &= \left|\left\{k : s_h^k = s, n_h^k(s) \le d_{max} \log\frac{HSA}{\delta}\right\}\right| \\ &\le \left|\left\{k : s_h^k = s, m_h^k(s) \le 2d_{max} \log\frac{HSA}{\delta}\right\}\right| \lesssim d_{max},\end{aligned}$$

By taking the union over s and h we have, $|\mathcal{K}_{exp}| \lesssim HSd_{max}$.

Similarly to the proof in Appendix B.2, we use the value difference lemma of Shani et al. (2020), on episodes that are not in $\mathcal{K}_{exp}$,

$$\begin{aligned}\mathcal{R}_K &\lesssim H^2 S d_{max} + \sum_{k\notin\mathcal{K}_{exp}} V_1^{\pi^k}(s_1^k) - V_1^{\pi}(s_1^k) \\ &= H^2 S d_{max} + \underbrace{\sum_{k\notin\mathcal{K}_{exp}} V_1^{\pi^k}(s_1^k) - V_1^{k}(s_1^k)}_{(A)} \\ &\quad + \underbrace{\sum_{k\notin\mathcal{K}_{exp}}\sum_{h=1}^{H} \mathbb{E}^{\pi}\left[\langle Q_h^k(s_h^k,\cdot), \pi_h^k(\cdot \mid s_h^k) - \pi_h^k(\cdot \mid s_h^k)\rangle\right]}_{(B)} \\ &\quad + \underbrace{\sum_{k\notin\mathcal{K}_{exp}}\sum_{h=1}^{H} \mathbb{E}^{\pi}\left[\langle \hat{p}_h^k(\cdot \mid s_h^k, a_h^k) - p_h(\cdot \mid s_h^k, a_h^k), V_{h+1}^k\rangle\right]}_{(C)}. \end{aligned} \tag{9}$$

We continue bounding each of these terms separately. Term (A) is bounded in Appendix B.3.1 by $\widetilde{O}\big(H^2S\sqrt{AK} + H^2S^{3/2}A^{3/2}\sqrt{d_{max}} + H^2S^2A^2 + H^2Sd_{max}\big)$. Terms (B) and (C) are bounded in Appendix B.3.2 by $\widetilde{O}\big(\frac{H}{\eta} + \eta H^3(D+K)\big)$. This gives a total regret bound of

$$\begin{aligned}\mathcal{R}_K &= \widetilde{O}\Big(H^2S\sqrt{AK} + \frac{H}{\eta} + \eta H^3(D+K) + H^2S^{3/2}A^{3/2}\sqrt{d_{max}} + H^2Sd_{max} + H^2S^2A^2\Big) \\ &\le \widetilde{O}\Big(H^2S\sqrt{AK} + H^2Sd_{max} + H^2S^2A^3 + \frac{H}{\eta} + \eta H^3(D+K)\Big),\end{aligned}$$

where the last is because $H^2S^{3/2}A^{3/2}\sqrt{d_{max}} \le O(H^2Sd_{max} + H^2S^2A^3)$. Choosing $\eta = \frac{1}{H\sqrt{K+D}}$ gives the theorem's statement.

Remark 3. *Recall that we compute $\hat{p}^k$ at the end of episode $k + d^k$. However, in our analysis we utilize only feedbacks that returned before round k (formally the shift is done in Lemma 2). This was done to ensure the validity of our concentration bounds. For example, the concentration in F_3^{basic} (see Appendix B.1) requires r_h^k to depend only on the history up to round k. But, by round $k + d^k$ we might observe feedbacks from episodes that occur after episode k.*

With that being said, if $k + d^k$ is strictly monotone in k (e.g., under fixed delay), then all feedback of episodes $j < k$ are available at time $k + d^k$. In that case we can essentially achieve the bounds of Theorem 1, even when trajectory feedback is delayed, and even without explicit exploration.

B.3.1. BOUNDING TERM (A) UNDER FULL-INFORMATION FEEDBACK

Term (A) under full-information can be written as,

$$\begin{aligned}(A) &= \sum_{k\notin\mathcal{K}_{exp}} V_1^{\pi^k}(s_1^k) - V_1^k(s_1^k) \\ &= \sum_{k\notin\mathcal{K}_{exp}} \sum_{h=1}^{H} \mathbb{E}^{\pi^k}\left[\langle p_h(\cdot \mid s_h^k, a_h^k) - \hat{p}_h^k(\cdot \mid s_h^k, a_h^k), V_{h+1}^k\rangle\right] \\ &\le H \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'} \mathbb{E}^{\pi^k}\left[|p_h(s' \mid s_h^k, a_h^k) - \hat{p}_h^k(s' \mid s_h^k, a_h^k)|\right].\end{aligned}$$

The last is bounded by Lemma 8, which is analogues to Lemma 5. This gives us the following bound on term (A):

$$(A) \lesssim H^2 S\sqrt{AK} + H^2 S^{3/2}A^{3/2}\sqrt{d_{max}} + H^2 S^2 A^2 + H^2 S d_{max}. \tag{10}$$

Lemma 8. *Under the good event, with explicit exploration ($UseExplicitExploration = \texttt{true}$),*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}} \mathbb{E}^{\pi^k}\left[\left|p_h(s' \mid s_h^k, a_h^k) - \hat{p}_h^k(s' \mid s_h^k, a_h^k)\right|\right] \lesssim HS\sqrt{AK} + HSd_{max} + HS^2A^2 + HS^{3/2}A^{3/2}\sqrt{d_{max}}.$$

Proof. Similarly to the proof of Lemma 5,

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}} \mathbb{E}^{\pi^k}\left[|p_h(s' \mid s_h^k, a_h^k) - \hat{p}_h^k(s' \mid s_h^k, a_h^k)|\right] = \\ &\quad= \sum_{k=1}^{K}\sum_{s,a,h} q_h^k(s,a)\sum_{s'\in\mathcal{S}} |p_h(s' \mid s,a) - \hat{p}_h^k(s' \mid s,a)| \\ &\quad\le \sum_{k=1}^{K}\sum_{s,a,h} q_h^k(s,a)\min\{2, r_h^k(s,k)\} \\ &\quad= \sum_{k=1}^{K}\sum_{s,a,h} \left(q_h^k(s,a) - \mathbb{I}\{s_h^k = s, a_h^k = a\}\right)\min\{2, r_h^k(s,k)\} \\ &\qquad+ \sum_{k=1}^{K}\sum_{s,a,h} \mathbb{I}\{s_h^k = s, a_h^k = a\}\min\{2, r_h^k(s,k)\} \\ &\quad\lesssim \sqrt{K} + \sum_{k=1}^{K}\sum_{s,a,h} \mathbb{I}\{s_h^k = s, a_h^k = a\}\min\{2, r_h^k(s,k)\} && \text{(by } \neg F_3^{basic}\text{)} \\ &\quad\le \sqrt{K} + 2\sum_{s,h}\sum_{k\in\mathcal{K}_{exp}(s,h)} \underbrace{\sum_{a} \mathbb{I}\{s_h^k = s, a_h^k = a\}}_{\le 1} \\ &\qquad+ \sum_{k\notin\mathcal{K}_{exp}}\sum_{s,a,h} \mathbb{I}\{s_h^k = s, a_h^k = a\} r_h^k(s,k) \\ &\quad\lesssim \sqrt{K} + HSd_{max} && (|\mathcal{K}_{exp}(s,h)| \lesssim d_{max}) \\ &\qquad+ \sqrt{S}\sum_{s,a,h}\sum_{k\notin\mathcal{K}_{exp}} \frac{\mathbb{I}\{s_h^k = s, a_h^k = a\}}{\sqrt{n_h^k(s,a)\vee 1}} + S\sum_{s,a,h}\sum_{k\notin\mathcal{K}_{exp}} \frac{\mathbb{I}\{s_h^k = s, a_h^k = a\}}{n_h^k(s,a)\vee 1}.\end{aligned}$$

The last two terms are bounded using Lemmas 9 and 10, which completes the proof. □

Lemma 9. *It holds that*

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h=1}^{H}\sum_{k\notin\mathcal{K}_{exp}}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{n_h^k(s,a)\vee 1}}\lesssim H\sqrt{SAK}+HSA^{3/2}\sqrt{d_{max}}.$$

Proof. For any s,a and h,

$$\begin{aligned}
&\sum_{k\notin\mathcal{K}_{exp}}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{n_h^k(s,a)\vee 1}}\leq\\
&\leq\sum_{k\notin\mathcal{K}_{exp}(s,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\cdot\sqrt{\frac{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}{1\vee\sum_{j:j+d^j\leq k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\\
&\leq\sum_{k\notin\mathcal{K}_{exp}(s,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\cdot\sqrt{1+\frac{1\vee\sum_{j:j<k,j+d^j\geq k}\mathbb{I}\{s_h^j=s,a_h^j=a\}}{1\vee\sum_{j:j+d^j\leq k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\\
&\leq\sum_{k\notin\mathcal{K}_{exp}(s,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\cdot\sqrt{1+\frac{1+\sum_{j:j<k,j+d^j\geq k}\mathbb{I}\{s_h^j=s,a_h^j=a\}}{1\vee\sum_{j:j+d^j\leq k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\\
&\leq\sum_{k\notin\mathcal{K}_{exp}(s,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\cdot\sqrt{2+\frac{\sum_{j:j<k,j+d^j\geq k}\mathbb{I}\{s_h^j=s,a_h^j=a\}}{1\vee\sum_{j:j+d^j\leq k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\\
&\lesssim\underbrace{\sum_{k=1}^{K}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}}_{(D.1)}\\
&\quad+\underbrace{\sum_{k\notin\mathcal{K}_{exp}(s,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\cdot\sqrt{\frac{\sum_{j:j<k,j+d^j\geq k}\mathbb{I}\{s_h^j=s,a_h^j=a\}}{1\vee\sum_{j:j+d^j\leq k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}}_{(D.2)}.
\end{aligned}$$

As mentioned before, $(D.1)$ appears in the non-delayed setting and can be bounded when summing over (s,a,h) by $\tilde{O}(H\sqrt{SAK})$. For bounding $(D.2)$ we need another notion:

$$\mathcal{K}_{exp}(s,a,h)=\left\{k\in[K]:s_h^k=s,a_h^k=a,n_h^k(s,a)\leq d_{max}\right\},$$

and notice that $|\mathcal{K}_{exp}(s,a,h)|\leq 2d_{max}$ since every visit is observable after d_{max} episodes.

Now summing $(D.2)$ over s and a,

$$\sum_{s,a}\sum_{k\notin\mathcal{K}_{exp}(s,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\cdot\sqrt{\frac{\sum_{j:j<k,j+d^j\geq k}\mathbb{I}\{s_h^j=s,a_h^j=a\}}{1\vee\sum_{j:j+d^j\leq k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}$$

$$=\underbrace{\sum_{s,a}\sum_{\substack{k\notin\mathcal{K}_{exp}(s,h)\\k\in\mathcal{K}_{exp}(s,a,h)}}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\cdot\sqrt{\frac{\sum_{j:j<k,j+d^j\geq k}\mathbb{I}\{s_h^j=s,a_h^j=a\}}{1\vee\sum_{j:j+d^j\leq k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}}_{(D.2.1)}$$

$$+\underbrace{\sum_{s,a}\sum_{\substack{k\notin\mathcal{K}_{exp}(s,h)\\k\notin\mathcal{K}_{exp}(s,a,h)}}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\cdot\sqrt{\frac{\sum_{j:j<k,j+d^j\geq k}\mathbb{I}\{s_h^j=s,a_h^j=a\}}{1\vee\sum_{j:j+d^j\leq k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}}_{(D.2.2)}.$$

Now, under the good event $(\neg F_5^{basic})$ we have that for any $k\notin\mathcal{K}_{exp}(s,h)$,

$$\sum_{j:j+d^j\leq k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}=n_h^k(s,a)\geq\Omega\left(\frac{d_{max}}{A}\right).$$

Also, deterministically we have $\sum_{j:j<k,j+d^j\geq k}\mathbb{I}\{s_h^j=s,a_h^j=a\}\leq d_{max}$. Hence,

$$(D.2.1)\leq\sqrt{A}\sum_{s,a}\sum_{k\notin\mathcal{K}_{exp}(s,h),k\in\mathcal{K}_{exp}(s,a,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\lesssim SA^{3/2}\sqrt{d_{max}},$$

where the last inequality follows from the fact that for any time the nominator is 1, the sum in the denominator has increased by 1 as well. Hence it is bounded by the sum $\sum_{i=1}^{2d_{max}}\frac{1}{\sqrt{i}}\leq O(\sqrt{d_{max}})$. For last, by definition, for all $k\notin\mathcal{K}_{exp}(s,a,h)$, $\sum_{j:j+d^j\leq k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}\geq d_{max}$. Hence,

$$\begin{aligned}(D.2.2)&\leq\sum_{s,a}\sum_{k\notin\mathcal{K}_{exp}(s,h),k\notin\mathcal{K}_{exp}(s,a,h)}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}\\&\leq\sum_{s,a}\underbrace{\sum_{k=1}^{K}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{\sqrt{1\vee\sum_{j=1}^{k-1}\mathbb{I}\{s_h^j=s,a_h^j=a\}}}}_{(D.1)}\lesssim\sqrt{SAK}.\end{aligned}$$

□

Lemma 10. *It holds that*

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h=1}^{H}\sum_{k\notin\mathcal{K}_{exp}}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{n_h^k(s,a)\vee1}\lesssim HSA^2.$$

Proof. For any s, a and h,

$$\begin{aligned}
&\sum_{k \notin \mathcal{K}_{exp}} \frac{\mathbb{I}\{s_h^k = s, a_h^k = a\}}{n_h^k(s,a) \vee 1} \leq \\
&\leq \sum_{k \notin \mathcal{K}_{exp}(s,h)} \frac{\mathbb{I}\{s_h^k = s, a_h^k = a\}}{1 \vee \sum_{j=1}^{k-1} \mathbb{I}\{s_h^j = s, a_h^j = a\}} \cdot \frac{1 \vee \sum_{j=1}^{k-1} \mathbb{I}\{s_h^j = s, a_h^j = a\}}{1 \vee \sum_{j:j+d^j \leq k-1} \mathbb{I}\{s_h^j = s, a_h^j = a\}} \\
&\leq \sum_{k \notin \mathcal{K}_{exp}(s,h)} \frac{\mathbb{I}\{s_h^k = s, a_h^k = a\}}{1 \vee \sum_{j=1}^{k-1} \mathbb{I}\{s_h^j = s, a_h^j = a\}} \cdot \Big(2 + \frac{\sum_{j:j<k, j+d^j \geq k} \mathbb{I}\{s_h^j = s, a_h^j = a\}}{1 \vee \sum_{j:j+d^j \leq k-1} \mathbb{I}\{s_h^j = s, a_h^j = a\}}\Big) \\
&\leq \sum_{k \notin \mathcal{K}_{exp}(s,h)} \frac{\mathbb{I}\{s_h^k = s, a_h^k = a\}}{1 \vee \sum_{j=1}^{k-1} \mathbb{I}\{s_h^j = s, a_h^j = a\}} \cdot \Big(2 + \frac{d_{max}}{d_{max}/A}\Big) \\
&\leq 3A \sum_{k \notin \mathcal{K}_{exp}(s,h)} \frac{\mathbb{I}\{s_h^k = s, a_h^k = a\}}{1 \vee \sum_{j=1}^{k-1} \mathbb{I}\{s_h^j = s, a_h^j = a\}} \lesssim A \log(KH),
\end{aligned}$$

where we used the fact that $\sum_{j:j<k, j+d^j \geq k} \mathbb{I}\{s_h^j = s, a_h^j = a\}$ is always bounded by d_{max}, and that $\sum_{j:j+d^j \leq k-1} \mathbb{I}\{s_h^j = s, a_h^j = a\}$ is at least $\Omega(d_{max}/A)$ for episodes that are not in $\mathcal{K}_{exp}$ under the good event. The last inequality follows from standard arguments, since there is no delay involved in this sum. □

B.3.2. Bounding Terms (B) and (C) under full-information feedback

Following the analysis of term (B.2) in the proof of Lemma 6, since we run the policy improvement step only over $[K] \backslash \mathcal{K}_{exp}$, and since under full-information $V_h^k \leq H$ and the costs are bounded by 1,

$$(B.2) \leq \eta H^2 (K + D).$$

Hence, term (B) can be bounded by,

$$(B) \leq \frac{H \log A}{\eta} + \eta H^3 (D + K). \tag{11}$$

For last, term $(C.1)$ is now zero, and so

$$(C) \leq 0. \tag{12}$$

B.4. Proof of Theorem 1

When the trajectories are observed without delay, Term (A) is bounded similarly to Shani et al. (2020) since it is no longer affected by the delay. Moreover, since there is no explicit exploration we also do not have the extra $H^2 S d_{max}$ factor. Terms (B) and (C) remain unchanged.

Thus, with bandit feedback, we obtain the regret bound

$$\mathcal{R}_K = \widetilde{O}\Big(H^2 S \sqrt{AK} + H^2 S^2 A + \gamma K H S A + \frac{H}{\eta} + \frac{\eta}{\gamma} H^3 (K + D) + \frac{H}{\gamma} + \frac{\eta}{\gamma^2} H^3 d_{max}\Big),$$

and choosing $\eta = \frac{1}{H(A^{3/2}K + D)^{2/3}}$ and $\gamma = \frac{1}{(A^{3/2}K + D)^{1/3}}$ gives the theorem's statement.

Similarly, with full-information feedback, we obtain the regret bound

$$\mathcal{R}_K = \widetilde{O}\Big(H^{3/2} S \sqrt{AK} + H^2 S^2 A + \frac{H}{\eta} + \eta H^3 (K + D)\Big)$$

and choosing $\eta = \frac{1}{H\sqrt{K + D}}$ gives the theorem's statement.

We note that Shani et al. (2020) bound Term (A) by $\widetilde{O}(H^2S\sqrt{AK})$, but with full-information feedback it can actually be bounded by $\widetilde{O}(H^{3/2}S\sqrt{AK})$. This is obtained by known Bernstein-based confidence bounds analysis (Azar et al., 2017; Zanette & Brunskill, 2019). For example, one can follow Lemmas 4.6,4.7,4.8 of Rosenberg et al. (2020) which are more general.

The reason that this bound (that improves by a factor of $\sqrt{H}$) does not hold with delayed trajectory feedback is Lemma 10 where we bound $\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h=1}^{H}\sum_{k\notin\mathcal{K}_{exp}}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{n_h^k(s,a)\vee 1}$ by $\widetilde{O}(HSA^2)$ instead of $\widetilde{O}(HSA)$ when the trajectory feedback is not delayed. Thus, the analysis of Rosenberg et al. (2020) gets a bound of $\widetilde{O}(H^{3/2}SA\sqrt{K})$ instead of $\widetilde{O}(H^{3/2}S\sqrt{AK})$, since it bounds Term (A) roughly by

$$H\sqrt{SK}\sqrt{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h=1}^{H}\sum_{k\notin\mathcal{K}_{exp}}\frac{\mathbb{I}\{s_h^k=s,a_h^k=a\}}{n_h^k(s,a)\vee 1}}.$$

B.5. Proof of Theorem 2

When dynamics are known, we use the actual transition function instead of the estimated one. Under the bandit-feedback, the terms $(A.2)$, $(A.3)$ in Appendix B.2 become zero. Since we use the actual occupancy measure of the policy (and do not compute it using some transition function from the confidence set), Term $(A.1.1)$ is now bounded by $\gamma KHSA$. Term $(A.2.1)$ remains unchanged. Therefore,

$$(A)\lesssim\gamma KHSA+H\sqrt{K}.$$

Term (B) remains unchanged and Term (C) is now bounded by $\tilde{O}(H/\gamma)$, as (C.2) zeroes.

Thus, with known transition function and bandit feedback, we obtain the regret bound

$$\mathcal{R}_K=\widetilde{O}\Big(H\sqrt{K}+\gamma KHSA+\frac{H}{\eta}+\frac{\eta}{\gamma}H^3(K+D)+\frac{H}{\gamma}+\frac{\eta}{\gamma^2}H^3d_{max}\Big),$$

and choosing $\eta=\frac{1}{H(A^{3/2}K+D)^{2/3}}$ and $\gamma=\frac{1}{(A^{3/2}K+D)^{1/3}}$ gives the theorem's statement.

Similarly, with known transition function and full-information feedback (the only non-zero term now is Term (B)), we obtain the regret bound

$$\mathcal{R}_K=\widetilde{O}\Big(\frac{H}{\eta}+\eta H^3(K+D)\Big)$$

and choosing $\eta=\frac{1}{H\sqrt{K+D}}$ gives the theorem's statement.

C. Skipping scheme for handling large delays

In this section we show that by skipping episodes with large delays, we can substitute the d_{max} term in Theorems 1 to 3 by $\sqrt{D}$. This was presented by Thune et al. (2019) for MAB with delays and can be easily applied to our setting as well. The idea is to skip episodes with delay larger than $\sqrt{D}$ and bound the regret on skipped episodes trivially by H. That way, effectively the maximum delay is $\sqrt{D}$ and the number of skipped episodes is at most $\sqrt{D}$ as well. The skipping scheme can be generalized for arbitrary threshold as presented in Algorithm 4.

Algorithm 4 Skipping Wrapper

Input: Algorithm ALG, Skipping threshold $\beta > 0$.
for $k = 1, 2, 3, ...$ **do**
  Get policy π^k from ALG and play the k-th episode with π^k.
  Observe feedback from all episodes in $\mathcal{F}^k = \{j : j + d^j = k\}$.
  Feed ALG all episodes $j \in \mathcal{F}^k$ such that $d^j \leq \beta$.
end for

Lemma 11. *Assume that we have a regret bound for ALG that depends on the number of episodes, the sum of delays and the maximum delay: $R^{ALG}(K, D, d_{max})$. Assume also that the ALG choices depend only on the feedback. Then the regret of Algorithm 4 when simulating ALG with a threshold $\beta > 0$ is at most,*

$$R^{ALG}(|\mathcal{K}_\beta|, D_\beta, \beta) + H(K - |\mathcal{K}_\beta|)$$

where $\mathcal{K}_\beta = \{k : d^k \leq \beta\}$ and $D_\beta = \sum_{k \in \mathcal{K}_\beta} d^k$.

Proof. Fix some threshold β and a policy π.

$$\sum_{k=1}^{K} V_1^{\pi^k}(s_1^k) - V_1^{\pi}(s_1^k) = \sum_{k \in \mathcal{K}_\beta} V_1^{\pi^k}(s_1^k) - V_1^{\pi}(s_1^k) + \sum_{k \notin \mathcal{K}_\beta} V_1^{\pi^k}(s_1^k) - V_1^{\pi}(s_1^k).$$

Since the the algorithm policies π^k are affected only by feedback from $\mathcal{K}_\beta$, and the total delay on those rounds is D_β, the first sum is bounded by $R^{ALG}(|\mathcal{K}_\beta|, D_\beta, \beta)$.

Since the value function is bounded by H, the second sum is bounded by $H(K - |\mathcal{K}_\beta|)$. □

Remark 4. *The proof of Lemma 11 relies on the fact that the algorithm does not observe feedback outside of $\mathcal{K}_\beta$. However, if the trajectory feedback is available immediately at the end of the episode, we can also feed the algorithm with trajectory feedback outside of $\mathcal{K}_\beta$. This can only shrink the confidence intervals and reduce the regret.*

Lemma 12. *For any threshold $\beta > 0$, the number of skipped rounds under Algorithm 4 is bounded by $K - \mathcal{K}_\beta < \frac{D}{\beta}$.*

Proof. First note that,

$$K - \mathcal{K}_\beta = \sum_{k=1}^{K} \left(1 - \mathbb{I}\{d^k \leq \beta\}\right) = \sum_{k=1}^{K} \mathbb{I}\{d^k > \beta\}.$$

Now, we can bound the sum of delays by,

$$D \geq \sum_{k=1}^{K} d^k \mathbb{I}\{d^k > \beta\} > \sum_{k=1}^{K} \beta \mathbb{I}\{d^k > \beta\} = \beta \left(K - \mathcal{K}_\beta\right).$$

Dividing both sides by β completes the proof. □

Using the last two lemmas and the regret guarantees that we show in previous sections, we can now deduce regret bounds for delayed OPPO, when simulated by Algorithm 4. In some settings there was no dependence on d_{max}, and thus no skipping is needed.

Corollary 1. *Running Delayed OPPO, results in the following regret bounds (with probability at least* $1-\delta$*):*

- *Under bandit feedback, known dynamics, and with threshold* $\beta = \sqrt{D/HS}$*,*

$$\mathcal{R}_K = \widetilde{O}\left(HS\sqrt{A}K^{2/3} + H^2D^{2/3}\right).$$

- *Under bandit feedback, unknown dynamics, non-delayed trajectory feedback, and with threshold* $\beta = \sqrt{D/HS}$*,*

$$\mathcal{R}_K = \widetilde{O}\left(HS\sqrt{A}K^{2/3} + H^2D^{2/3} + H^2S^2A\right).$$

- *Under full-information feedback, unknown dynamics delayed trajectory feedback, and with threshold* $\beta = \sqrt{D/HS}$*,*

$$\mathcal{R}_K = \widetilde{O}\left(H^2S\sqrt{AK} + H^{3/2}\sqrt{SD} + H^2S^2A^3\right).$$

- *Under bandit feedback, unknown dynamics, delayed trajectory feedback, and with threshold* $\beta = \sqrt{D/HSA}$*,*

$$\mathcal{R}_K = \widetilde{O}\left(HS\sqrt{A}K^{2/3} + H^2D^{2/3} + H^2S^2A^3 + S^3A^3\right).$$

Proof. The first three regret bounds follow immediately from the regret bounds we show for Delayed OPPO, Lemma 11 and Lemma 12. For the last bound, we directly get a bound of,

$$\tilde{O}\left(HS\sqrt{A}K^{2/3} + H^2D^{2/3} + H^{3/2}\sqrt{SAD} + H^2S^2A^3 + S^3A^3\right).$$

Note that if the third term dominates over the second, then $D \leq (SA/H)^3$, which implies that

$$H^{2/3}\sqrt{SD} \leq S^3A^3.$$

□

D. Doubling trick for handling unknown number of episodes and total delay

Denote by M^k the number of missing samples at the end of episode k. That is, $M^k = k - \sum_{j=1}^{k} |\mathcal{F}^j|$.

Algorithm 5 for unknown D and K, uses the well-known doubling trick. It estimates the value $(D+K)$ and initializes a new phase whenever the estimation doubles itself. At time k we estimate the value of $(D+K)$ by $k + \sum_{j=1}^{k} M^j$.

Note that this is an optimistic estimation of $(D+K)$. If the feedback from episode j arrived, then we estimate its delay exactly by d^j. However, if the feedback did not arrive, we estimate it as if the feedback will arrive in the next episode.

We also include in the algorithm the skipping scheme from previous section, in order to avoid the dependence on d_{max}.

Algorithm 5 Delayed OPPO with known transition function, bandit feedback and unknown D and K

Input: State space $\mathcal{S}$, Action space $\mathcal{A}$, Transition function p.
Initialization: Set $\pi_h^1(a \mid s) = 1/A$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $e = 1$, $\eta_e = H^{-1}2^{-2e/3}$, $\gamma_e = 2^{-e/3}$, $\beta_e = 2^{e/2}$.
for $k = 1, 2, \ldots, K$ **do**
  Play episode k with policy π^k.
  Observe feedback from all episodes $j \in \mathcal{F}^k$.
  # Policy Evaluation
  for $j \in \mathcal{F}^k$ such that $d^j \leq \beta_e$ **do**
    $\forall s \in \mathcal{S} : V_{H+1}^j(s) = 0$.
    for $h = H, \ldots, 1$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
      $\hat{c}_h^j(s,a) = \frac{c_h^j(s,a) \cdot \mathbb{I}\{s_h^j = s, a_h^j = a\}}{q_h^{p,\pi^j}(s)\pi_h^j(a|s) + \gamma_e}$.
      $Q_h^j(s,a) = \hat{c}_h^j(s,a) + \langle p_h(\cdot \mid s,a), V_{h+1}^j \rangle$.
      $V_h^j(s) = \langle Q_h^j(s,\cdot), \pi_h^j(\cdot \mid s) \rangle$.
    end for
  end for
  # Policy Improvement
  for $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ **do**
    $\pi_h^{k+1}(a|s) = \frac{\pi_h^k(a|s) \exp(-\eta_e \sum_{j \in \mathcal{F}^k : d^j \leq \beta_e} Q_h^j(s,a))}{\sum_{a' \in \mathcal{A}} \pi_h^k(a'|s) \exp(-\eta_e \sum_{j \in \mathcal{F}^k : d^j \leq \beta_e} Q_h^j(s,a'))}$.
  end for
  # Doubling
  if $k + \sum_{j=1}^{k} M^j = k + \sum_{j=1}^{k} (j - \sum_{i=1}^{j} |\mathcal{F}^i|) > 2^e$ **then**
    Set $\pi_h^1(a \mid s) = 1/A$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $e = e+1$, $\eta_e = H^{-1}2^{-2e/3}$, $\gamma_e = 2^{-e/3}$, $\beta_e = 2^{e/2}$.
  end if
end for

Theorem 5. *Under bandit feedback, if the transition function is known, then (with probability at least $1-\delta$), the regret of Algorithm 5 is bounded by,*

$$\mathcal{R}_K \leq \tilde{O}\Big(HSAK^{2/3} + H^2 D^{2/3}\Big).$$

Proof. The proof follows the proof technique of Theorem 2 in Bistritz et al. (2019). Let $\mathcal{K}_e$ be the episodes in phase e, $\mathcal{N}_e' = \{k \in \mathcal{K}_e : k + d^k \notin \mathcal{K}_e\}$, $\mathcal{N}_{\beta_e} = \{k \in \mathcal{K}_e : d^k \geq \beta_e\}$, and $\mathcal{N}_e = \mathcal{N}_e' \cup \mathcal{N}_{\beta_e}$ which is the set of episodes with missing feedback in phase e. Also, denote the last episode of phase e by K_e. Using Theorem 2 and Lemma 11.

$$\sum_{k \in \mathcal{K}_e} V_1^{\pi^k}(s_1^k) - V_1^{\pi}(s_1^k) = \tilde{O}\Big(\gamma_e |\mathcal{K}_e| HSA + \frac{H}{\eta_e} + \frac{\eta_e}{\gamma_e} H^3 \sum_{k \in \mathcal{K}_e \setminus \mathcal{N}_e} (1 + d^k) + \frac{H}{\gamma_e} + H^3 \frac{\eta_e}{\gamma_e^2} \beta_e + H\sqrt{|\mathcal{K}_e|} + H|\mathcal{N}_e|\Big). \tag{13}$$

By definition of our doubling scheme,

$$\begin{aligned}
2^e &\geq \sum_{k=1}^{K_e}(1+M^k) \\
&= K_e + \sum_{k=1}^{K_e}\sum_{j=1}^{K}\mathbb{I}\{j \leq k, j+d^j > k\} \\
&= K_e + \sum_{j=1}^{K}\sum_{k=1}^{K_e}\mathbb{I}\{j \leq k, j+d^j > k\} \\
&\geq K_e + \sum_{j\in\mathcal{K}_e\setminus\mathcal{N}_e}\sum_{k=1}^{K_e}\mathbb{I}\{j \leq k, j+d^j > k\} \\
&\underset{(*)}{\geq} |\mathcal{K}_e| + \sum_{j\in\mathcal{K}_e\setminus\mathcal{N}_e} d^j \\
&\geq \sum_{j\in\mathcal{K}_e\setminus\mathcal{N}_e}(d^j+1),
\end{aligned} \tag{14}$$

where $(*)$ holds since $j + d^j \leq K_e$ for any $j \in \mathcal{K}_e\setminus\mathcal{N}_e$. We now bound $|\mathcal{N}'_e|$. Similarly to the above,

$$\begin{aligned}
2^e &\geq \sum_{j=1}^{K}\sum_{k=1}^{K_e}\mathbb{I}\{j \leq k, j+d^j > k\} \\
&\geq \sum_{j\in\mathcal{N}'_e}\sum_{k=K_{e-1}+1}^{K_e}\mathbb{I}\{j \leq k, j+d^j > k\} \\
&\underset{(*)}{=} \sum_{j\in\mathcal{N}'_e} K_e - j + 1 \\
&\underset{(**)}{\geq} \sum_{j=K_e-|\mathcal{N}'_e|}^{K_e} K_e - j + 1 \\
&\geq \sum_{j=1}^{|\mathcal{N}'_e|+1} j \\
&= \frac{1}{2}(|\mathcal{N}'_e|+1)(|\mathcal{N}'_e|+2) \\
&\geq \frac{1}{2}|\mathcal{N}'_e|^2,
\end{aligned}$$

where $(*)$ follows from the fact that for any $k \in \mathcal{K}_e$ we have that $j + d^j > k$, and $(**)$ follows by choosing the smallest possible $|\mathcal{N}'_e|$ indices. The above implies that,

$$|\mathcal{N}'_e| \leq 2^{\frac{e+1}{2}}. \tag{15}$$

For last, we bound $|\mathcal{N}_{\beta_e} \setminus \mathcal{N}'_e|$,

$$\begin{aligned}
2^e &\geq \sum_{j=1}^{K}\sum_{k=1}^{K_e} \mathbb{I}\{j \leq k, j + d^j > k\} \\
&\geq \sum_{j \in \mathcal{N}_{\beta_e} \setminus \mathcal{N}'_e} \sum_{k=1}^{K_e} \mathbb{I}\{j \leq k, j + d^j > k\} \\
&\underset{(*)}{=} \sum_{j \in \mathcal{N}_{\beta_e} \setminus \mathcal{N}'_e} d^j \\
&\underset{(**)}{\geq} \sum_{j \in \mathcal{N}_{\beta_e} \setminus \mathcal{N}'_e} \beta_e \\
&= |\mathcal{N}_{\beta_e} \setminus \mathcal{N}'_e| \beta_e,
\end{aligned}$$

where $(*)$ follows because $j \notin \mathcal{N}'_e$ and so $j + d^j \leq K_e$. And $(**)$ follows the definition of $\mathcal{N}_{\beta_e}$ This implies that,

$$|\mathcal{N}_{\beta_e} \setminus \mathcal{N}'_e| \leq 2^{\frac{e}{2}}.$$

Combining with (15) gives us,

$$|\mathcal{N}_e| = |\mathcal{N}'_e \cup \mathcal{N}_{\beta_e}| \leq 2^{\frac{e+3}{2}}. \tag{16}$$

Plugging the bounds of (14) and (16) into (13), noting that $|\mathcal{K}_e| \leq 2^e$, plugging the values of η_e, γ_e and β_e, and summing over all phases gives us,

$$\mathcal{R}_K \leq \tilde{O}\left(HSA \sum_{e=1}^{E} |\mathcal{K}_e| 2^{-\frac{e}{3}} + H^2 \sum_{e=1}^{E} 2^{\frac{2e}{3}}\right),$$

where $E \leq \log(K + D)$ is the number of phases. The second sum above is a sum of geometric series and can be bounded by $O\big((D + K)^{2/3}\big)$. The first term is bounded by the value of the following maximization problem,

$$\begin{aligned}
&\max \sum_{e=1}^{E} |\mathcal{K}_e| 2^{-\frac{e}{3}} \\
&\text{subject to } |\mathcal{K}_e| \leq 2^e, \sum_{e=1}^{E} |\mathcal{K}_e| = K,
\end{aligned}$$

which is necessarily bounded by

$$\begin{aligned}
&\max \sum_{e=1}^{\infty} x_e 2^{-\frac{e}{3}} \\
&\text{subject to } 0 \leq x_e \leq 2^e, \sum_{e=1}^{\infty} x_e = K.
\end{aligned}$$

Since $2^{-\frac{e}{3}}$ is decreasing, this is maximized whenever the first x_es are at maximum value. There are at most $\lceil \log K \rceil$ non-zero x_es and so,

$$\sum_{e=1}^{E} |\mathcal{K}_e| 2^{-\frac{e}{3}} \leq \sum_{e=1}^{\lceil \log K \rceil} 2^{\frac{2e}{3}} \leq O\left(K^{\frac{2}{3}}\right),$$

which gives us the desired bound. □

Remark 5. *The exact same proof holds for the rest of settings, in which we get the exact same regret as in Corollary 1, Theorem 1 and Theorem 2 under full-information. Under bandit feedback the $HS\sqrt{A}K^{2/3}$ becomes $HSAK^{2/3}$ as in Theorem 5.*

E. Delayed O-REPS

Given an MDP $\mathcal{M}$, a policy π induces an occupancy measure $q = q^\pi$, which satisfies the following:

$$\sum_{a,s'} q_1(s_{\text{init}}, a, s') = 1 \tag{17}$$

$$\sum_{s,a,s'} q_h(s, a, s') = 1 \qquad \forall h = 1, \ldots, H \tag{18}$$

$$\sum_{s',a} q_h(s, a, s') = \sum_{s',a} q_{h-1}(s', a, s) \qquad \forall s \in \mathcal{S} \text{ and } h = 2, \ldots, H \tag{19}$$

If q satisfies (17), (18) and (19), then it induces a policy and a transition function in the following way:

$$\pi_h^q(a \mid s) = \frac{\sum_{s'} q_h(s, a, s')}{\sum_{a'} \sum_{s'} q_h(s, a', s')}$$

$$p_h^q(s' \mid s, a) = \frac{q(s', a, s)}{\sum_{s,a} q(s', a, s)}$$

The next lemma characterize the occupancy measures induced by some policy π.

Lemma 13 (Lemma 3.1, Rosenberg & Mansour (2019b))**.** *An occupancy measure q that satisfies (17), (18) and (19) is induced by some policy π **if and only if** $p^q = p$.*

Definition 1. *Given an MDP $\mathcal{M}$, we define $\Delta(\mathcal{M})$ to be the set of all $q \in [0,1]^{H \times S \times A \times S}$ that satisfies (17), (18) and (19) such that $p^q = p$ (where p is the transition function of $\mathcal{M}$).*

For connivance, in this section we let the cost functions to be a function of the current state, the action taken and the next state: $c_h(s, a, s')$. So the value of a policy, π is given by

$$V^\pi(s_{\text{init}}) = \langle q^\pi, c \rangle = \sum_h \langle q_h^\pi, c_h \rangle$$

and the regret with respect to a policy π is given by

$$\mathcal{R}_K = \sum_{k=1}^{K} \langle q^{\pi^k} - q^\pi, c^k \rangle.$$

The above can be treated as an online linear optimization problem. Indeed, O-REPS (Zimin & Neu, 2013) treats it as such by running Online Mirror Decent (OMD) on the set of occupancy measures. That is, at each episode the algorithm plays the policy induced by the occupancy measure q^k and updates the occupancy measure for the next episode by,

$$q^{k+1} = \arg \min_{q \in \Delta(\mathcal{M})} \{\eta \langle q, c^k \rangle + D_R(q \| q^k)\},$$

where is R the unnormalized negative entropy. That is,

$$R(q) = \sum_h \sum_{s,a,s'} q_h(s, a, s') \log q_h(s, a, s') - q_h(s, a, s')$$

and D_R is the Bregman divergence associated with R (D_R is also known as the Kullback-Leibler divergence). If the feedback is delayed we would update the occupancy measure using all the feedback that arrives at the end of the current episode, i.e.,

$$q^{k+1} = \arg \min_{q \in \Delta(\mathcal{M})} \{\eta \langle q, \sum_{j \in \mathcal{F}^k} c^j \rangle + D_R(q \| q^k)\}.$$

Whenever the transition function is unknown, $\Delta(\mathcal{M})$ can not be computed. In this case we adopt the method of Rosenberg & Mansour (2019b) and extend $\Delta(\mathcal{M})$ by the next definition.

Definition 2. *For any $k \in [K]$, we define $\Delta(\mathcal{M}, k)$ to be the set of all $q \in [0,1]^{H \times S \times A \times S}$ that satisfies (17), (18), (19) and*

$$\forall h, s', a, s : |p_h^q(s' \mid s, a) - \bar{p}_h^k(s' \mid s, a)| \leq \epsilon_h^k(s' \mid s, a).$$

The update step will now be with respect to $\Delta(\mathcal{M}, k)$. We have that with high probability $\Delta(\mathcal{M}, k)$ contain $\Delta(\mathcal{M})$ for all k, and so the estimation of the value function is again optimistic. Delayed O-REPS for unknown dynamics is presented in Algorithm 6.

Algorithm 6 Delayed O-REPS

Input: State space $\mathcal{S}$, Action space $\mathcal{A}$, Learning rate $\eta > 0$, Confidence parameter $\delta > 0$.
Initialization: Set $\pi_h^1(a \mid s) = 1/A$, $q_h^1(s, a, s') = 1/S^2 A$ 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.
  Observe feedback from all episodes $j \in \mathcal{F}^k$, and last trajectory $U^k = \{(s_h^k, a_h^k)\}_{h=1}^H$.
  Update transition function estimation.
  # Update Occupancy Measure
  if transition function is known **then**
    $q^{k+1} = \arg\min_{q \in \Delta(\mathcal{M})} \{\eta \langle q, \sum_{j \in \mathcal{F}^k} c^j \rangle + D_R(q \| q^k)\}$
  else
    $q^{k+1} = \arg\min_{q \in \Delta(\mathcal{M}, k)} \{\eta \langle q, \sum_{j \in \mathcal{F}^k} c^j \rangle + D_R(q \| q^k)\}$
  end if
  # Update Policy
  Set $\pi_h^{k+1}(a \mid s) = \frac{\sum_{s'} q_h^{k+1}(s,a,s')}{\sum_{a'} \sum_{s'} q_h^{k+1}(s,a',s')}$ for every $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \times [H]$.
end for

The update step can be implemented by first solving the unconstrained convex optimization problem,

$$\tilde{q}^{k+1} = \arg\min \left\{ \eta \left\langle q, \sum_{j \in \mathcal{F}^k} c^j \right\rangle + D_R(q \| q^k) \right\}, \tag{20}$$

and then projecting onto the set $\Delta(\mathcal{M}, k)$ with respect to $D_R(\cdot \| \tilde{q}^{k+1})$. That is,

$$q^{k+1} = \arg \min_{q \in \Delta(\mathcal{M}, k)} D_R(q \| \tilde{q}^{k+1}).$$

The solution for (20) is simply given by,

$$\tilde{q}_h^{k+1}(s, a, s') = q_h^k(s, a, s') e^{-\eta \sum_{j \in \mathcal{F}^k} c^j(s,a,s')}.$$

Theorem 6. *Running Delayed O-REPS under full-information feedback and delayed cost feedback guarantees the following regret, with probability at least $1 - \delta$,*

$$\mathcal{R}_K = \widetilde{O}(H^{3/2} S \sqrt{AK} + H\sqrt{D} + H^2 S^2 A).$$

Moreover, if the transition function is known, we obtain regret of

$$\mathcal{R}_K = \widetilde{O}(H\sqrt{K + D}).$$

E.1. Proof of Theorem 6

Given a policy π and a transition p, we denote the occupancy measure of π with respect to p, by $q^{p,\pi}$. That is, $q_h^{p,\pi}(s, a, s') = \Pr[s_h^k = s, a_h^k = a, s_{h+1}^k = s' \mid s_1^k = s_{\text{init}}, \pi, p]$. Also, denote $p^k = p^{q^k}$, and note that by definition $q^{p^k, \pi^k} = q^k$. We define the following good event $G = \neg F_1^{basic}$ where F_1^{basic} defined in Appendix B.1. As shown in Lemma 1, G occurs with

probability of at least $1-\delta$. As consequence we have that for all episodes, $\Delta(\mathcal{M}) \subseteq \Delta(\mathcal{M}, k)$. From this point we analyse the regret given that G occurred.

We break the regret in the following way:

$$\begin{aligned}\sum_{k=1}^{K}\langle q^{p,\pi^k} - q^{p,\pi}, c^k\rangle &= \sum_{k=1}^{K}\langle q^{p,\pi^k} - q^{p^k,\pi^k}, c^k\rangle + \sum_{k=1}^{K}\langle q^{p^k,\pi^k} - q^{p,\pi}, c^k\rangle \\ &= \sum_{k=1}^{K}\langle q^{p,\pi^k} - q^{p^k,\pi^k}, c^k\rangle + \sum_{k=1}^{K}\langle q^{k} - q^{\pi}, c^k\rangle. \end{aligned} \tag{21}$$

The first term, under the good event, is bounded similarly as in the proof of Theorem 1 by,

$$\sum_{k=1}^{K}\langle q^{p,\pi^k} - q^{p^k,\pi^k}, c^k\rangle \lesssim H^{3/2}S\sqrt{AK} + H^2S^2A. \tag{22}$$

For the second term, we adopt the approach of (Thune et al., 2019; Bistritz et al., 2019), and modify the standard analysis of OMD. We start with Lemma 14 which bounds the regret of playing π^{k+d^k} at episode k.

Lemma 14. *If $\Delta(\mathcal{M}) \subseteq \Delta(\mathcal{M}, k)$ for all k, then for any $q \in \Delta(\mathcal{M})$, delayed O-REPS satisfies*

$$\sum_{k=1}^{K}\langle c^k, q^{k+d^k} - q\rangle \leq \frac{2H\log(HSA)}{\eta} + \eta HK.$$

Proof. Note that $\tilde{q}_h^{k+1}(s,a,s') = q_h^k(s,a,s')e^{-\eta\sum_{j\in\mathcal{F}^k} c_h^j(s,a,s')}$. Taking the log,

$$\eta\sum_{j\in\mathcal{F}^k} c_h^j(s,a,s') = \log q_h^k(s,a,s') - \log\tilde{q}_h^{k+1}(s,a,s').$$

Hence for any q

$$\begin{aligned}\eta\left\langle\sum_{j\in\mathcal{F}^k} c_h^j, q^k - q\right\rangle &= \left\langle\log q^k - \log\tilde{q}^{k+1}, q^k - q\right\rangle \\ &= D_R(q||q^k) - D_R(q||\tilde{q}^{k+1}) + D_R(q^k||\tilde{q}^{k+1}) \\ &\leq D_R(q||q^k) - D_R(q||q^{k+1}) - D_R(q^{k+1}||\tilde{q}^{k+1}) + D_R(q^k||\tilde{q}^{k+1}) \\ &\leq D_R(q||q^k) - D_R(q||q^{k+1}) + D_R(q^k||\tilde{q}^{k+1}),\end{aligned}$$

where the first equality follows directly the definition of Bregman divergence. The first inequality is by (Zimin, 2013, Lemma 1.2) and the assumption that $\Delta(\mathcal{M}) \subseteq \Delta(\mathcal{M}, k)$. The second inequality is since Bregman divergence is non-negative. Now, the last term is bounded by,

$$\begin{aligned}D_R(q^k||\tilde{q}^{k+1}) &\leq D_R(q^k||\tilde{q}^{k+1}) + D_R(\tilde{q}^{k+1}||q^k) \\ &= \sum_h\sum_{s,a,s'}\tilde{q}_h^{k+1}(s,a,s')\log\frac{\tilde{q}_h^{k+1}(s,a,s')}{q_h^k(s,a,s')} \\ &\quad + \sum_h\sum_{s,a,s'} q_h^k(s,a,s')\log\frac{q_h^k(s,a,s')}{\tilde{q}_h^{k+1}(s,a,s')} \\ &= \langle q^k - \tilde{q}^{k+1}, \log q^k - \log\tilde{q}^{k+1}\rangle \\ &= \eta\left\langle q^k - \tilde{q}^{k+1}, \sum_{j\in\mathcal{F}^k} c^j\right\rangle.\end{aligned}$$

We get that

$$\eta \left\langle \sum_{j \in \mathcal{F}^k} c^j, q^k - q \right\rangle \leq D_R(q||q^k) - D_R(q||q^{k+1}) + \eta \left\langle q^k - \tilde{q}^{k+1}, \sum_{j \in \mathcal{F}^k} c^j \right\rangle.$$

Summing over k and dividing by η, we get

$$\begin{aligned}
\underbrace{\sum_{k=1}^{K} \sum_{j \in \mathcal{F}^k} \langle c^j, q^k - q \rangle}_{(*)} &\leq \frac{D_R(q||q^1) - D_R(q||q^{K+1})}{\eta} + \sum_{k=1}^{K} \left\langle q^k - \tilde{q}^{k+1}, \sum_{j \in \mathcal{F}^k} c^j \right\rangle \\
&\leq \frac{D_R(q||q^1)}{\eta} + \sum_{k=1}^{K} \left\langle q^k - \tilde{q}^{k+1}, \sum_{j \in \mathcal{F}^k} c^j \right\rangle \\
&\leq \frac{2H \log(SA)}{\eta} + \underbrace{\sum_{k=1}^{K} \left\langle q^k - \tilde{q}^{k+1}, \sum_{j \in \mathcal{F}^k} c^j \right\rangle}_{(**)},
\end{aligned}$$

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

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

In a similar way,

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

This gives us,

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

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

$$\begin{aligned}
\sum_k \langle q^{k+d^k} - \tilde{q}^{k+d^k+1}, c^k \rangle &= \sum_k \sum_h \sum_{s,a,s'} c_h^k(s,a,s')(q_h^{k+d^k}(s,a,s') - \tilde{q}_h^{k+d^k+1}(s,a,s')) \\
&= \sum_k \sum_h \sum_{s,a,s'} c_h^k(s,a,s') \left(q_h^{k+d^k}(s,a,s') - q_h^{k+d^k}(s,a,s') e^{-\eta \sum_{j \in \mathcal{F}^{k+d^k+1}} c_h^k(s,a,s')} \right) \\
&\leq \sum_k \sum_h \sum_{s,a,s'} q_h^{k+d^k}(s,a,s') \left(1 - e^{-\eta \sum_{j \in \mathcal{F}^{k+d^k+1}} c_h^k(s,a,s')} \right) \\
&\leq \eta \sum_k \sum_h \sum_{s,a,s'} q_h^{k+d^k}(s,a,s') \left(\sum_{j \in \mathcal{F}^k + d^k + 1} c_h^k(s,a,s') \right) && (1 - e^{-x} \leq x) \\
&\leq \eta \sum_k \sum_h \sum_{s,a,s'} q_h^{k+d^k}(s,a,s') \left| \mathcal{F}^{k+d^k+1} \right| && (c_h^k(s,a,s') \leq 1) \\
&= \eta H \sum_k \left| \mathcal{F}^{k+d^k+1} \right| \\
&\leq \eta H K.
\end{aligned}$$

This completes the proof of the lemma. □

Using Lemma 14, we can bounds the regret as,

$$\sum_{k=1}^K \langle c^k, q^k - q \rangle \leq \frac{2H \log(HSA)}{\eta} + \eta H K + \sum_{k=1}^K \langle c^k, q^k - q^{k+d^k} \rangle. \tag{23}$$

The next lemma is a generalization of Lemma 1 in (Zimin & Neu, 2013), which allows us to bound the distance between two consecutive occupancy measures.

Lemma 15. *For q_h^k that are generated by delayed O-REPS, we have that,*

$$\sum_h D_R(q_h^k \| q_h^{k+1}) \leq \sum_h \sum_{s,a,s'} q_h^k(s,a,s') \frac{(\eta \sum_{j \in \mathcal{F}^k} c_h^j(s,a,s'))^2}{2}.$$

Proof. First we present some notations. Given $v_h, e_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, define

$$B_h^k(s,a,s' \mid v, e) = e_h(s,a,s') + v_h(s,a,s') - \eta \sum_{j \in \mathcal{F}^k} c_h^j(s,a,s') - \sum_{s''} \bar{p}_h^k(s'' \mid s,a) v_{h+1}(s,a,s'').$$

Given $\beta_h : \mathcal{S} \to \mathbb{R}$ and $\mu_h^+, \mu_h^- : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$, define

$$\begin{aligned}
v^{\mu_h}(s,a,s') &= \mu_h^-(s,a,s') - \mu_h^+(s,a,s') \\
e_h^{\mu_h, \beta_h}(s,a,s') &= (\mu_h^-(s,a,s') + \mu_h^+(s,a,s')) \epsilon_h^k(s' \mid s,a) + \beta_h(s) - \beta_{h+1}(s'),
\end{aligned}$$

where we always set $\beta_1 = \beta_H = 0$. For last, define

$$Z_h^k(v,e) = \sum_{s,a,s'} q_h^k(s,a,s') e^{B_h^k(s,a,s' \mid v,e)}.$$

By (Rosenberg & Mansour, 2019b, Theorem 4.2), we have that

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

where

$$\mu^k, \beta^k = \arg\min_{\beta,\mu\geq 0} \sum_{h=1}^{H} \log Z_h^k(v^{\mu_h}, e^{\mu_h,\beta_h}).$$

Now, we have that

$$\begin{aligned}
\sum_h D_R(q_h^k \| q_h^{k+1}) &= \sum_h \sum_{s,a,s'} q_h^k(s,a,s') \log \frac{q_h^k(s,a,s')}{q_h^{k+1}(s,a,s')} \\
&= \sum_h \sum_{s,a,s'} q_h^k(s,a,s') \log \frac{Z_h^k(v^{\mu_h^k}, e^{\mu_h^k,\beta_h^k})}{e^{B_h^k(s,a,s'|v^{\mu_h^k}, e^{\mu_h^k,\beta_h^k})}} \\
&= \underbrace{\sum_h \log Z_h^k(v^{\mu_h^k}, e^{\mu_h^k,\beta_h^k})}_{(A)} - \underbrace{\sum_h \sum_{s,a,s'} q_h^k(s,a,s') B_h^k(s,a,s' \mid v^{\mu_h^k}, e^{\mu_h^k,\beta_h^k})}_{(B)}.
\end{aligned}$$

By definition of μ_h^k, β_h^k, term (A) can be bounded by

$$\begin{aligned}
(A) &\leq \sum_h \log Z_h^k(0,0) \\
&= \sum_h \log(\sum_{s,a,s'} q_h^k(s,a,s') e^{B_h^k(s,a,s'|0,0)}) \\
&= \sum_h \log(\sum_{s,a,s'} q_h^k(s,a,s') e^{-\eta \sum_{j\in\mathcal{F}^k} c_h^j(s,a,s')}) \\
&\leq \sum_h \log(\sum_{s,a,s'} q_h^k(s,a,s')(1 - \eta \sum_{j\in\mathcal{F}^k} c_h^j(s,a,s') + \frac{(\eta \sum_{j\in\mathcal{F}^k} c_h^j(s,a,s'))^2}{2})) \qquad (\forall x \geq 0 : e^{-x} \leq 1 - x + \tfrac{x^2}{2}) \\
&= \sum_h \log(1 - \eta \sum_{s,a,s'} \sum_{j\in\mathcal{F}^k} q_h^k(s,a,s') c_h^j(s,a,s') + \sum_{s,a,s'} q_h^k(s,a,s') \frac{(\eta \sum_{j\in\mathcal{F}^k} c_h^j(s,a,s'))^2}{2}) \\
&\leq -\eta \sum_h \sum_{s,a,s'} \sum_{j\in\mathcal{F}^k} q_h^k(s,a,s') c_h^j(s,a,s') + \sum_h \sum_{s,a,s'} q_h^k(s,a,s') \frac{(\eta \sum_{j\in\mathcal{F}^k} c_h^j(s,a,s'))^2}{2}. \qquad (\log(1+x) \leq x)
\end{aligned}$$

Term (B) can be rewritten as

$$
\begin{aligned}
(B) &= \sum_h \sum_{s,a,s'} q_h^k(s,a,s')(e^{\mu_h^k,\beta_h^k}(s,a,s') + v^{\mu_h^k}(s,a,s') \\
&\quad - \eta \sum_{j\in\mathcal{F}^k} c_h^j(s,a,s') - \sum_{s''} p^k(s'' \mid s,a) v^{\mu_h^k}(s,a,s'')) \\
&= \sum_h \sum_{s,a,s'} q_h^k(s,a,s') e^{\mu_h^k,\beta_h^k}(s,a,s') + \sum_h \sum_{s,a,s'} q_h^k(s,a,s') v^{\mu_h^k}(s,a,s') \\
&\quad - \eta \sum_h \sum_{s,a,s'} \sum_{j\in\mathcal{F}^k} q_h^k(s,a,s') c_h^j(s,a,s') \\
&\quad - \sum_h \sum_{s,a,s'} \sum_{s''} q_h^k(s,a,s') p_h^k(s'' \mid s,a) v^{\mu_h^k}(s,a,s'') \\
&= \sum_h \sum_{s,a,s'} q_h^k(s,a,s') e^{\mu_h^k,\beta_h^k}(s,a,s') + \sum_h \sum_{s,a,s'} q_h^k(s,a,s') v^{\mu_h^k}(s,a,s') \\
&\quad - \eta \sum_h \sum_{s,a,s'} \sum_{j\in\mathcal{F}^k} q_h^k(s,a,s') c_h^j(s,a,s') \\
&\quad - \sum_h \sum_{s,a} \sum_{s''} q_h^k(s,a) p^k(s'' \mid s,a) v^{\mu_h^k}(s,a,s'') \\
&= \sum_h \sum_{s,a,s'} q_h^k(s,a,s') e^{\mu_h^k,\beta_h^k}(s,a,s') + \sum_h \sum_{s,a,s'} q_h^k(s,a,s') v^{\mu_h^k}(s,a,s') \\
&\quad - \eta \sum_h \sum_{s,a,s'} \sum_{j\in\mathcal{F}^k} q_h^k(s,a,s') c_h^j(s,a,s') - \sum_h \sum_{s,a,s''} q_h^k(s,a,s'') v^{\mu_h^k}(s,a,s'') \\
&= \sum_h \sum_{s,a,s'} q_h^k(s,a,s') e^{\mu_h^k,\beta_h^k}(s,a,s') - \eta \sum_h \sum_{s,a,s'} \sum_{j\in\mathcal{F}^k} q_h^k(s,a,s') c_h^j(s,a,s').
\end{aligned}
$$

Overall we get

$$
\begin{aligned}
\sum_h D_R(q_h^k \| q_h^{k+1}) &\le \sum_h \sum_{s,a,s'} q_h^k(s,a,s') \frac{(\eta \sum_{j\in\mathcal{F}^k} c_h^j(s,a,s'))^2}{2} \\
&\quad - \sum_h \sum_{s,a,s'} q_h^k(s,a,s') e^{\mu_h^k,\beta_h^k}(s,a,s') \\
&\le \sum_h \sum_{s,a,s'} q_h^k(s,a,s') \frac{(\eta \sum_{j\in\mathcal{F}^k} c_h^j(s,a,s'))^2}{2} \qquad (\mu_h^k \ge 0) \\
&\quad - \sum_h \sum_{s,a,s'} q_h^k(s,a,s')(\beta_h^k(s) - \beta_{h+1}^k(s')).
\end{aligned}
$$

For last, we show that the second term is 0, which completes the proof

$$\begin{aligned}
&\sum_h \sum_{s,a,s'} q_h^k(s,a,s')(\beta_h^k(s) - \beta_{h+1}^k(s')) = \\
&\qquad = \sum_h \left[\sum_{s,a,s'} q_h^k(s,a,s')\beta_h^k(s) - \sum_{s,a,s'} q_h^k(s,a,s')\beta_{h+1}^k(s') \right] \\
&\qquad = \sum_h \left[\sum_s q_h^k(s)\beta_h^k(s) - \sum_{s'} q_{h+1}^k(s')\beta_{h+1}^k(s') \right] \\
&\qquad = \sum_s \sum_h \left[q_h^k(s)\beta_h^k(s) - q_{h+1}^k(s)\beta_{h+1}^k(s) \right] \\
&\qquad = \sum_s \left[q_1^k(s)\beta_1^k(s) - q_H^k(s)\beta_H^k(s) \right] = 0,
\end{aligned}$$

where the last equality follows since $\beta_1^k = \beta_H^k = 0$. □

We now use the lemma above to bound the last term in (23):

$$\begin{aligned}
\sum_{k=1}^K \langle c^k, q^k - q^{k+d^k} \rangle &\le \sum_{k=1}^K \sum_h \sum_{s,a,s'} |q_h^k(s,a,s') - q_h^{k+d^k}(s,a,s')| \\
&\le \sum_{k=1}^K \sum_{j=k}^{k+d^k-1} \sum_h \sum_{s,a,s'} |q_h^j(s,a,s') - q_h^{j+1}(s,a,s')| \\
&\le 2 \sum_{k=1}^K \sum_{j=k}^{k+d^k-1} \sum_h \sqrt{2 D_R(q_h^j \| q_h^{j+1})} && \text{(by Pinsker's inequality)} \\
&\le 2 \sum_{k=1}^K \sum_{j=k}^{k+d^k-1} \sqrt{2H \sum_h D_R(q_h^j \| q_h^{j+1})} && \text{(by Jensen's inequality)} \\
&\le \sum_{k=1}^K \sum_{j=k}^{k+d^k-1} \sqrt{H \sum_h \sum_{s,a,s'} q_h^j(s,a,s')(\eta \sum_{i \in \mathcal{F}^j} c_h^i(s,a,s'))^2} && \text{(by Lemma 15)} \\
&\le \sum_{k=1}^K \sum_{j=k}^{k+d^k-1} \sqrt{H \sum_h \sum_{s,a,s'} q_h^j(s,a,s')(\eta |\mathcal{F}^j|)^2} \\
&= \eta H \sum_{k=1}^K \sum_{j=k}^{k+d^k-1} |\mathcal{F}^j| \le \eta H D,
\end{aligned}$$

where the last inequality is shown in the proof of Theorem 1 in (Thune et al., 2019). Combining the above with (21), (22) and (23) completes the proof.

F. Full the details of the empirical evaluation and more experiments

We conducted our experiments on a grid-world of size 10×10 i.e., $S = 100$ and horizon $H = 50$. There are four types of states: *Initial state*, s_{init}, which is always the top-left corner, *goal state*, s_{goal}, which is always the buttom-right corner, *wall states* which are not reachable and *regular states* which are the rest of the states on the grid. There are four actions $A = \{\texttt{up}, \texttt{down}, \texttt{right}, \texttt{left}\}$. After taking an action, the agent moves with probability 0.9 towards the adjacent state in the corresponding direction, provided that this is not a wall state or falls outside the grid. With probability 0.1 the direction is perturbed uniformly. The cost function is defined as $c(s, a) = \mathbb{I}\{s \neq s_{\text{goal}}\}$.

Implementation of the algorithms. As presented by (Shani et al., 2020), under stochastic MDP, the estimated transition $\hat{p}_h^j$ can be replaced with the observed empirical transitions, and instead reduce the cost by order of $1/\sqrt{n_h^j(s,a)}$ during the policy evaluation step. This forces the algorithm to explore states that were not visited enough and keeps the estimated Q-function optimistic as our algorithm does. For better computational efficiency we implemented the policy evaluation step in our algorithm with these kind of estimates. All algorithms where run under full-information cost feedback. All algorithms where tested with a fixed learning rate $\eta = 0.1$. Reduction maintains $d_{max} + 1$ copies of OPPO where d_{max} is realized maximal delay. Effectively each copy suffers no delay so this is reduced to standard OPPO (Shani et al., 2020). Each of our experiment take 2-5 hours of computation time on a CPU.

Adding wall states. The results of Fig. 1 where tested on a simple grid without wall states. In Fig. 2 we added wall states so that a more complex dynamics needs to be learned.

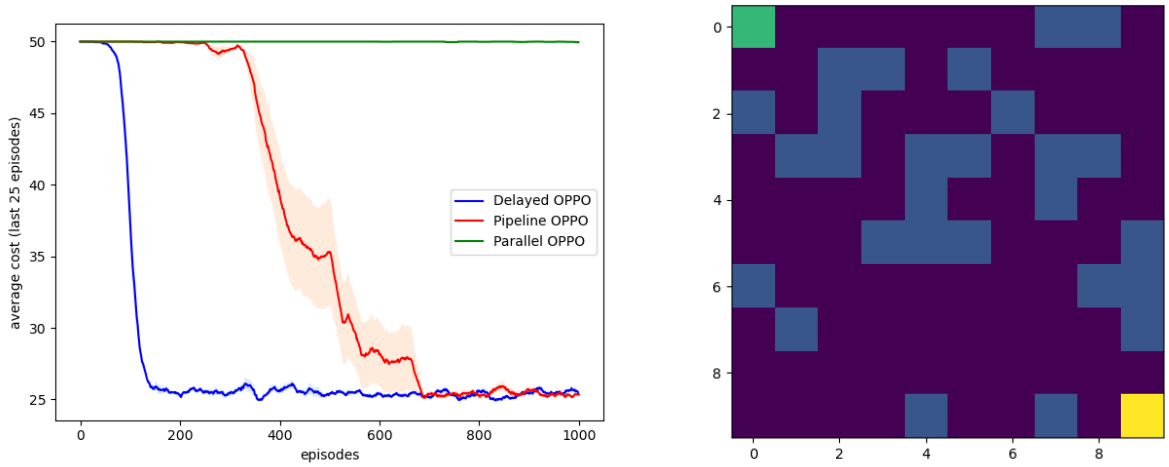

Figure 2. On the left: Average cost of delayed algorithms in grid world with walls with geometrically distributed delays with mean 10. On the right: the grid environment with the wall states, where green is s_{init}, yellow is s_{goal}, dark blue are regular states and blue are the wall states.

Note that the convergence time of all algorithms increases compared to Fig. 1, as a more complex policy need to be learned. However, Delayed OPPO still keeps its great advantage over the other two alternatives.

Convergence time of Parallel-OPPO. In all the experiments we presented so far, Parallel-OPPO have not shown an improvement over time. In order to show the difference on convergence time, we changed the delay distribution to be geometric with mean 2 and increased the number of episodes to $K = 2000$. The results are averaged over 10 runs and appear in Fig. 3.

The maximal delay scale approximately as $2\log(K) \approx 15$. While each copy of OPPO that Parallel-OPPO maintains suffers effectively no delay, this is insignificant compared to the fact that Delayed OPPO observe approximately 15 times more observation then each copy. Pipeline-OPPO performs quite well in this case, as the maximal delay is quite small and approximately after 15 episodes it has a pipeline of observations. With that being said, note that even when the maximal delay quite small, Delayed OPPO definitely outperforms Pipline-OPPO. In addition, it is sufficient to have a single large

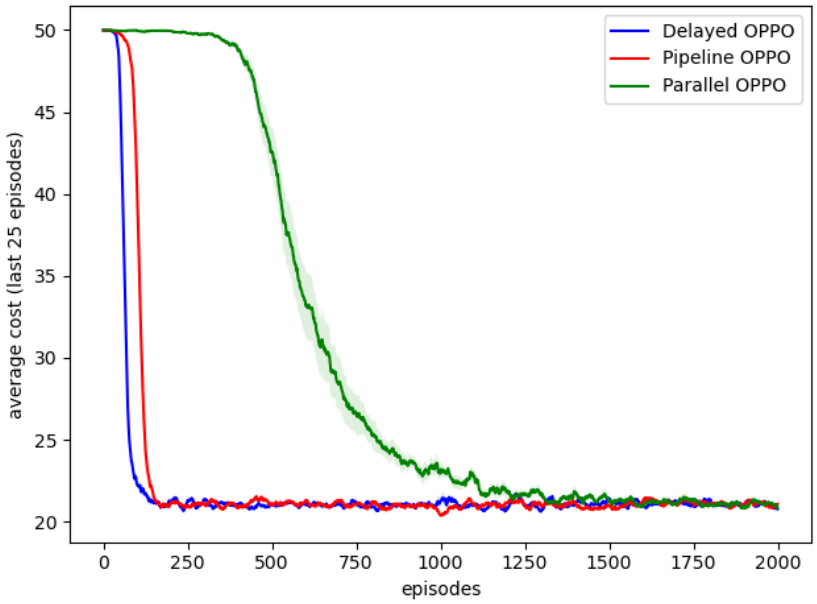

Figure 3. Average cost of delayed algorithms in grid world with walls with geometrically distributed delays with mean 2.

delay in order to have a major reduction in the performance of Pipline-OPPO. In real-world application it is quite common to have few large delays. In fact, few delays might be infinite, for example due to packet loss over a network. This would make Pipline-OPPO and Parallel-OPPO completely degenerate, while the effect of few missing observations on Delayed-OPPO is only minor.