Delay-Adapted Policy Optimization and Improved Regret for Adversarial MDP with Delayed Bandit Feedback

Tal Lancewicki [1] **Aviv Rosenberg** [2] **Dmitry Sotnikov** [2]

Abstract

Policy Optimization (PO) is one of the most popular methods in Reinforcement Learning (RL). Thus, theoretical guarantees for PO algorithms have become especially important to the RL community. In this paper, we study PO in adversarial MDPs with a challenge that arises in almost every real-world application – *delayed bandit feedback*. We give the first near-optimal regret bounds for PO in tabular MDPs, and may even surpass state-of-the-art (which uses less efficient methods). Our novel Delay-Adapted PO (DAPO) is easy to implement and to generalize, allowing us to extend our algorithm to: (i) infinite state space under the assumption of linear Q-function, proving the first regret bounds for delayed feedback with function approximation. (ii) deep RL, demonstrating its effectiveness in experiments on MuJoCo domains.

1. Introduction

Policy Optimization (PO) is one of the most widely-used methods in Reinforcement Learning (RL). It has demonstrated impressive empirical success (Levine & Koltun, 2013; Schulman et al., 2017; Haarnoja et al., 2018), leading to increasing interest in understanding its theoretical guarantees. While in recent years we have seen great advancement in theory of PO (Shani et al., 2020b; Luo et al., 2021; Chen et al., 2022b), our understanding is still very limited when considering *delayed feedback* – an important challenge that occurs in most practical applications. For example, recommendation systems often learn the utility of a recommendation based on the number of user conversions, which may happen with a variable delay after the recommendation was issued. Other notable examples include communication between agents (Chen et al., 2020a), video streaming (Changuel et al., 2012) and robotics (Mahmood et al., 2018). To mitigate the gap in the PO literature, we study PO in the challenging adversarial MDP model (i.e., costs change arbitrarily) under bandit feedback with arbitrary unrestricted delays.

PO with delays was previously studied by Lancewicki et al. (2022b), but their regret bounds are far from optimal and scale with $(K+D)^{2/3}$, where K is the number of episodes and D is the total delay. Recently, Jin et al. (2022) achieved near-optimal $\widetilde{\mathcal{O}}(H^2S\sqrt{AK} + (HSA)^{1/4}H\sqrt{D})$ regret (ignoring logarithmic factors), where S, A and H are the number of states, actions, and the episode length, respectively. However, their algorithm is not based on PO, but on the O-REPS method (Zimin & Neu, 2013) which requires solving a computationally expensive global optimization problem and cannot be extended to function approximation (FA). On the other hand, PO algorithms build on highly efficient local-search and extend naturally to FA (Tomar et al., 2022).

Our Contributions. In this paper, we vastly expand our understanding of PO and delayed feedback. We propose a novel Delay-Adapted PO method, called DAPO, which measures changes in the agent's policy over the time of the delays and adapts its updates accordingly. First, we establish the power of DAPO in tabular MDPs, i.e., finite number of states and actions. We prove DAPO attains the first near-optimal regret bound for PO with delayed feedback $\widetilde{\mathcal{O}}(H^3S\sqrt{AK} + H^3\sqrt{D})$. This bound is tighter than (Jin et al., 2022) when the delay term is dominant and the number of states is significantly larger than the horizon (which is the common case). Moreover, it matches the lower bound of Lancewicki et al. (2022b) in the delay term up to factors of H, showing for the first time that the delay term in the regret does not need to scale with S or A. Importantly, if there is no delay, it matches the best known regret for PO (Luo et al., 2021).

Next, we show DAPO is easy to implement and naturally extends to function approximation in two important settings:

1. *Linear-Q.* We extend DAPO to MDPs with linear FA under standard assumptions (Luo et al., 2021) that Q-functions are linear in some known low-dimensional

[1]Tel Aviv University (Research conducted while the author was an intern at Amazon Science) [2]Amazon Science. Correspondence to: Tal Lancewicki <lancewicki@mail.tau.ac.il>, Aviv Rosenberg <avivros007@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

features and also a simulator is available. We prove that DAPO achieves the first sub-linear regret for delayed feedback with FA, i.e., non-tabular MDP.

2. Deep RL. We show that the famous PPO algorithm (Schulman et al., 2017) can be easily combined with DAPO, and demonstrate superior empirical performance even in the presence of simple delays in experiments on MuJoCo domains (Todorov et al., 2012).

Throughout the paper we handle several technical challenges which are unique to PO algorithms with delayed feedback. The main challenge is to control the stability of the algorithm. This problem is more challenging in MDPs compared to multi-armed bandit (where there is no transition function to estimate), and is enhanced even further in Policy Optimization algorithms due to their local-search nature, as opposed to the global update of O-REPS methods – see more details in the proof sketch of Theorem 3.1. Further, the linear Q setting with delayed feedback was not studied before and requires a careful new algorithmic design and analysis. In particular, a-priori, it is highly unclear how to design a delay-adapted estimator and delay-adapted bonus term which sufficiently stabilize the algorithm – see more details in Section 4.

While the main contribution of this paper is the novel Delay-Adapted PO method, we also make substantial technical contributions that might be of independent interest. Our algorithms are based on PO with dilated bonuses (Luo et al., 2021), which dilate towards further horizons and do not satisfy standard Bellman equations. However, we are able to achieve the same regret guarantees without using dilated bonuses. Instead, we compute local bonuses and use them to construct a Q-function that operates as exploration bonuses. This has an important practical benefit – now bonuses can be approximated similarly to the Q-function. It also has a theoretical benefit – it greatly simplifies the analysis, making it more natural and easy to extend to new scenarios. Moreover, utilizing our new simplified analysis, we are able to give regret guarantees with high probability in the Linear Q setting and not just in expectation (as in Luo et al. (2021)). Finally, we also develop new analyses for handling delayed feedback when losses can be negative. This was not addressed in the delayed multi-armed bandit literature (or in previous papers on delays in MDPs), but cannot be avoided in our case since exploration bonuses are crucial to guaranteeing near-optimal regret but they might turn losses to negative.

1.1. Additional Related Work

Due to lack of space, this section only gives a brief overview of related work - for a full literature review see Appendix A.

There is a rich literature on regret minimization in tabular MDPs, initiated with the seminal UCRL algorithm (Jaksch et al., 2010) for stochastic losses that is based on the fundamental concept of Optimism Under Uncertainty. Their model was later extended to the more general adversarial MDP, where most algorithms are based on either the framework of occupancy measures (a.k.a, O-REPS) (Zimin & Neu, 2013; Jin et al., 2020a) or on the more practical PO (Even-Dar et al., 2009; Shani et al., 2020b). In recent years this line of research was extended beyond the tabular model to linear function approximation. For stochastic losses, existing algorithms are mostly based on optimism (Jin et al., 2020b), whereas most algorithms for adversarial losses are based on PO (Cai et al., 2020; Luo et al., 2021; Neu & Olkhovskaya, 2021) which extends much more naturally than O-REPS to function approximation. On the practical side, some of the most successful deep RL algorithms are built upon PO principles. These include the famous Trust Region PO (TRPO; Schulman et al. (2015)) as well as Proximal PO (PPO; Schulman et al. (2017)) which we will further discuss and adapt to delayed feedback in Section 5.

Regret minimization with delayed feedback was initially studied in Online Optimization and Multi-armed bandit (MAB) in both the stochastic setting (Agarwal & Duchi, 2012; Pike-Burke et al., 2018) and the adversarial setting (Cesa-Bianchi et al., 2016; Thune et al., 2019). As a natural extension, this line of work was generalized to delayed feedback in MDPs, where (Howson et al., 2021) consider the more restrictive stochastic model. Most related to our work are the works of Lancewicki et al. (2022b); Jin et al. (2022) that were mentioned earlier, and the work of Dai et al. (2022). They recently showed that Follow-The-Perturbed-Leader (FTPL) algorithms can also handle delayed feedback in adversarial MDPs. The efficiency of FTPL is similar to PO, but their regret bound is slightly weaker than Jin et al. (2022). Finally, a different line of work (Katsikopoulos & Engelbrecht, 2003; Walsh et al., 2009) consider delays in observing the current state. That setting is inherently different than ours (see Appendix A for more details).

2. Preliminaries

A finite-horizon episodic adversarial MDP is defined by a tuple $M = (S, A, H, p, \{c^k\}_{k=1}^K)$, where S and A are state and action spaces of sizes $|S| = S$ and $|A| = A$, respectively, H is the horizon and K is the number of episodes. $p : S \times A \times [H] \to \Delta_S$ is the transition function such that $p_h(s'|s,a)$ is the probability to move to s' when taking action a in state s at time h. $\{c^k : S \times A \times [H] \to [0,1]\}_{k=1}^K$ are cost functions chosen by an oblivious adversary, where $c_h^k(s,a)$ is the cost for taking action a at (s,h) in episode k.

A policy $\pi : S \times [H] \to \Delta_A$ is a function that gives the probability $\pi_h(a|s)$ to take action a when visiting state s at time h. The value $V_h^\pi(s;c)$ is the expected cost of π with respect to cost function c starting from s in time h, i.e.,

$V_h(s;c) = \mathbb{E}\left[\sum_{h'=h}^{H} c_{h'}(s_{h'},a_{h'}) \,|\, s_h = s\right]$, where the expectation is with respect to policy π and transition function p, that is, $a_{h'} \sim \pi_{h'}(\cdot|s_{h'})$ and $s_{h'+1} \sim p_{h'}(\cdot|s_{h'},a_{h'})$. The Q-function is defined by $Q_h(s,a;c) = c_h(s,a) + \langle p_h(\cdot|s,a), V_{h+1}(\cdot;c)\rangle$, where $\langle\cdot,\cdot\rangle$ is the dot product.

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

The goal of the learner is to minimize the regret, defined as the difference between the learner's cumulative expected cost and the best fixed policy in hindsight:

$$R_K = \sum_{k=1}^{K} V_1^{\pi^k}(s_{\text{init}};c^k) - \min_{\pi}\sum_{k=1}^{K} V_1^{\pi}(s_{\text{init}};c^k).$$

Algorithm 1 DAPO with Known Transitions (Tabular)

Initialization: Set $\pi_h^1(a|s) = 1/A$ for every (s,a,h).
for $k = 1,2,\ldots,K$ do
  Play episode k with policy π^k.
  # Policy Evaluation
  for j such that $j + d^j = k$ do
    Observe bandit feedback $\{c_h^j(s_h^j,a_h^j)\}_{h=1}^H$.
    Compute delay-adapted estimator $\hat{Q}_h^j(s,a)$ defined in Eq. (3).
    Compute delay-adapted bonus $B_h^j(s,a)$ as the Q-function of π^j with respect to the costs $b_h^j(s)$ defined in Eq. (4). I.e., compute recursively for $h = H,\ldots,1$, $B_h^j(s,a) = b_h^j(s) + \mathbb{E}_{s'\sim p_h(\cdot|s,a),\, a'\sim\pi_{h+1}^j(\cdot|s')} B_{h+1}^j(s',a')$.
  end for
  # Policy Improvement
  Define the policy π^{k+1} for every (s,a,h) by:

$$\pi_h^{k+1}(a|s) \propto e^{-\sum_{j:j+d^j<k}(\hat{Q}_h^j(s,a) - B_h^j(s,a))} \quad (1)$$

end for

In Section 3 we consider tabular MDPs, i.e., MDPs with a finite number of states and actions. In Section 4 we consider the more general case that allows infinite number of states but under the assumption that the Q-function is linear for all policies. We follow the standard definition (Abbasi-Yadkori et al., 2019; Neu & Olkhovskaya, 2021; Luo et al., 2021), which also assumes the number of actions is finite.

Assumption 2.1 (Linear-Q). Let $\phi : S \times A \times [H] \to \mathbb{R}^n$ be a known feature mapping. Assume that for every episode k, policy π and step h there exist an unknown vector $\theta_h^{k,\pi} \in \mathbb{R}^n$ such that $Q_h(s,a;c^k) = \phi_h(s,a)^\top \theta_h^{k,\pi}$ for all (s,a). Moreover, $\|\phi_h(s,a)\|_2 \le 1$ and $\|\theta_h^{k,\pi}\|_2 \le H\sqrt{n}$.

Additional notations. Episode indices appear as superscripts and in-episode steps as subscripts. The total delay is $D = \sum_k d^k$, the maximal delay is $d_{\max}$ and the number of episodes that their feedback arrives in the end of episode k is $m^k = |\{j : j + d^j = k\}|$. The occupancy measure $q_h(s,a) = \Pr[s_h = s, a_h = a \,|\, s_1 = s_{\text{init}}]$ is the distribution that policy π induces over state-action pairs in step h, and $q_h(s) = \sum_{a\in A} q_h(s,a)$. The notations $\tilde{O}(\cdot)$ and $\lesssim$ hide poly-logarithmic factors including $\log(K/\delta)$ for confidence parameter δ. $[n] = \{1,2,\ldots,n\}$ and the indicator of event E is $\mathbb{I}\{E\}$. Finally, denote by π^* the best fixed policy in hindsight and use the notations $V_h^k(s), Q_h^k(s,a), q_h^k(s,a)$ when the policy and cost are π^k and c^k, respectively.

Simplifying assumptions. Similarly to Jin et al. (2022), we assume that K, D and $d_{\max}$ are known. This assumption simplifies presentation and can be easily removed with doubling for delayed feedback (Bistritz et al., 2021; Lancewicki et al., 2022b). Bounds in the main text hide low-order terms and additive dependence on $d_{\max}$ (see Remark C.2 on removing $d_{\max}$ dependence). For full bounds see Appendix.

3. DAPO for Tabular MDP

In this section we present our novel Delayed-Adapted Policy Optimization algorithm (DAPO; presented in Algorithm 1) for the tabular case, where the number of states is finite. We use this fundamental model to develop a generic method for handling delayed feedback with Policy Optimization. Our approach consists of two important algorithmic features: a new delay-adapted importance-sampling estimator for the Q-function and a novel delay-adapted bonus term to drive exploration. Remarkably, this method extends naturally to both linear function approximation as we show in Section 4, and to deep RL with the extremely practical PPO algorithm (Schulman et al., 2017) as we show in Section 5.

To simplify presentation and focus on the contributions of our delay adaptation method, in this section we assume that the agent knows the transition function in advance. Generalizing DAPO to unknown transitions in the tabular case is fairly straightforward, and follows the common approach of optimism and confidence sets (Jaksch et al., 2010). Due to lack of space, in the main text we only provide sketches for the algorithms and proofs. The full versions (for both known and unknown transitions), together with the detailed analyses, can be found in Appendices B and C.

[1] If $d^k \equiv 0$, we get standard online learning in adversarial MDP.

PO algorithms follow the algorithmic paradigm of Policy Iteration (see, e.g., Sutton & Barto (2018)). That is, in every iteration they perform an evaluation of the current policy and then a step of policy improvement. The improvement step is regularized to be "soft", and is practically implemented by running an online multi-armed bandit algorithm, such as Hedge (Freund & Schapire, 1997), locally in each state. The losses that are fed to the algorithm are the estimated Q-functions, but in order to achieve the optimal regret, they are also combined with a bonus term which aims to stabilize the algorithm and drive exploration (keeping the estimated Q-function optimistic). The actual policy update step is based on exponential weights and presented in Eq. (1).

DAPO adapts to delays through the policy evaluation step. Remarkably, it adapts to delays near-optimally by computing the following simple ratio, which measures the local change in the agent's policy through the time of the delay,

$$r_h^k(s,a) = \frac{\pi_h^k(a \mid s)}{\max\{\pi_h^k(a \mid s), \pi_h^{k+d^k}(a \mid s)\}}. \qquad (2)$$

In order to get our delay-adapted Q-function estimation, we simply multiply $r_h^k(s,a)$ by the standard importance-sampling estimator from the non-delayed setting (Luo et al., 2021). The result is:

$$\widehat{Q}_h^k(s,a) = r_h^k(s,a)\,\frac{\mathbb{I}\{s_h^k = s, a_h^k = a\} L_h^k}{q_h^k(s)\,\pi_h^k(a \mid s) + \gamma}; \qquad (3)$$

where $L_h^k = \sum_{h'=h}^{H} c_{h'}^k(s_{h'}^k, a_{h'}^k)$ is the realized cost-to-go from step h and γ is an exploration parameter (Neu, 2015) needed to guarantee regret with high probability.

Intuitively, incorporating $r_h^k(s,a)$ helps us control the variance of the estimator $\widehat{Q}_h^k(s,a)$ in the presence of delays, since it will be used only in episode $k + d^k$ where actions are chosen according to π^{k+d^k} and not π^k. The maximum in the denominator is needed in order to keep the bias small. We note that this estimator is inspired by Jin et al. (2022), but there are two major differences. First, as Jin et al. (2022) perform their update globally in the space of state-action occupancy measures, their adaptation occurs in the state space as well. On the other hand we perform the update locally in each state, so our adaptation takes place only in the action space. Second, they directly change importance-sampling weighting, while we simply multiply standard estimators by the delay-adapted ratio. This seemingly minor nuance is critical in more complex (non-tabular) regimes, where its generality allows to utilize existing procedures from the non-delayed case (see more details in Sections 4 and 5).

Finally, to complement our new estimator, we devise an appropriate delay-adapted bonus $B_h^k(s,a)$ based on the following delay-adapted local bonus (again obtained by combining $r_h^k(s,a)$ with the original local bonus of Luo et al. (2021)),

$$b_h^k(s) = \sum_{a \in A} r_h^k(s,a)\,\frac{3H\,\pi_h^{k+d^k}(a \mid s)}{q_h^k(s)\,\pi_h^k(a \mid s) + \gamma}. \qquad (4)$$

At this point, Luo et al. (2021) compute $B_h^k(s,a)$ using a dilated Bellman equation that is not very intuitive. Instead, we compute $B_h^k(s,a)$ with the regular Bellman equations, making it a proper Q-function. This is an important contribution that might be of independent interest for two reasons – theoretically the analysis becomes much simpler, and practically the bonuses can be approximated like a Q-function.

Next, we present the regret guarantees of DAPO in tabular MDPs, and the key steps in the analysis, which highlight the intuition behind our algorithm design.

Theorem 3.1. Running DAPO in a tabular adversarial MDP guarantees with probability $1-\delta$, for known transition,

$$R_K = \widetilde{O}(H^2\sqrt{SAK} + H^3\sqrt{S(K+D)})$$

when setting $\eta = \sqrt{H^2SAK + H^4(K+D)}^{-1/2}$ and $\gamma = 2\eta H$, and for unknown transition,

$$R_K = \widetilde{O}(H^3S\sqrt{AK} + H^3\sqrt{SD})$$

when setting $\eta = \min\{\frac{1}{H}, \sqrt{H^2SAK + H^4(K+D)}^{-1/2}\}$ and $\gamma = 2\eta H$.

This is a big improvement compared to the best known regret for PO (Lancewicki et al., 2022b) that scales as $(K+D)^{2/3}$ (ignoring dependencies in H, S, A). It is also better than the current state-of-the-art regret bound of Jin et al. (2022) in the case that there is significant delay and $D \ll KH$ (which occurs in almost every practical application). While their bound has better dependency in H (this is a known weakness of PO (Chen et al., 2022b)), we improve the dependency in S and A. Lancewicki et al. (2022b) also show a lower bound of $\Omega(H^{3/2}\sqrt{SAK} + H\sqrt{SD})$. Thus, our bound shows for the first time that under the optimal regret, the delay term does not scale with S or A. The first term in our regret bound matches the state-of-the-art regret of non-delayed PO (Luo et al., 2021), and matches the best known regret for (non-delayed) adversarial MDPs in general up to factors of H (Jin et al., 2020a). Moreover, DAPO is the first efficient algorithm to be consistent with the optimal regret in delayed MAB, i.e., for $H = S = 1$ we get the optimal regret of Thune et al. (2019); Bistritz et al. (2019). Finally, it is important to emphasize that PO algorithms are much more practical than O-REPS algorithms, and extend naturally to function approximation, as we show in Sections 4 and 5.

Proof sketch of Theorem 3.1. Much of the intuition for PO algorithms stems from a classic regret decomposition known as the value difference lemma (Even-Dar et al., 2009):

$R_K = \sum_{k,h} \mathbb{E}_{s \sim q_h}\langle \pi_h^k(\cdot|s) - \pi_h(\cdot|s), Q_h^k(s,\cdot)\rangle$. Fixing a state s and step h, the sum over k can be viewed as the regret of an online experts algorithm (e.g., Hedge) with respect to the losses $Q_h^k(s,\cdot)$. We propose to further decompose the regret as follows,

$$\begin{aligned}
R_K = &\sum_{k,h} \mathbb{E}_{s \sim q_h}\langle \pi_h^{k+d^k}(\cdot|s) - \pi_h(\cdot|s), Q_h^k(s,\cdot) - \hat{Q}_h^k(s,\cdot)\rangle \\
&\quad \underbrace{\qquad\qquad}_{\text{BIAS}_1} \\
&+ \sum_{k,h} \mathbb{E}_{s \sim q_h}\langle \pi_h(\cdot|s) - \pi_h(\cdot|s), \hat{Q}_h^k(s,\cdot) - Q_h^k(s,\cdot)\rangle \\
&\quad \underbrace{\qquad\qquad}_{\text{BIAS}_2} \\
&+ \sum_{k,h} \mathbb{E}_{s \sim q_h}\langle \pi_h^k(\cdot|s) - \pi_h(\cdot|s), B_h^k(s,\cdot)\rangle \qquad (5) \\
&\quad \underbrace{\qquad\qquad}_{\text{BONUS}} \\
&+ \sum_{k,h} \mathbb{E}_{s \sim q_h}\langle \pi_h^{k+d^k}(\cdot|s) - \pi_h(\cdot|s), \hat{Q}_h^k(s,\cdot) - B_h^k(s,\cdot)\rangle \\
&\quad \underbrace{\qquad\qquad}_{\text{REG}} \\
&+ \sum_{k,h} \mathbb{E}_{s \sim q_h}\langle \pi_h^k(\cdot|s) - \pi_h^{k+d^k}(\cdot|s), Q_h^k(s,\cdot) - B_h^k(s,\cdot)\rangle : \\
&\quad \underbrace{\qquad\qquad}_{\text{DRIFT}}
\end{aligned}$$

Indeed, the policy update step in Eq. (1) is an Hedge-style exponential weights update. This allows us to bound the REG term as it represents the regret of Hedge with respect to the losses $\hat{Q}_h^k(s,a) - B_h^k(s,a)$. Note that the delayed feedback causes a shift of the agent's policies from π^k to π^{k+d^k}. As a result, we can bound (using Corollary E.7 in Appendix E):

$$\begin{aligned}
\text{REG} &\le \frac{H}{\eta} + \sum_{k,h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)\big(\hat{Q}_h^k(s,a) - B_h^k(s,a)\big)^2 \\
&\le \frac{H}{\eta} + H^5 K + \sum_{k,h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)\hat{Q}_h^k(s,a)^2 \\
&\qquad\qquad \underbrace{\qquad\qquad}_{(*)}
\end{aligned}$$

where the second inequality is since $|B_h^k(s,a)| \le H^2$. To bound the last term, we start with a concentration bound. This allows us to substitute the indicator in Eq. (3) by its expectation (which is $q_h^k(s,a)$) and cancel out the denominator once. The resulting bound is:

$$(*) \le H^2 \sum_{k,h,s,a} \frac{q_h(s)\,\pi_h^{k+d^k}(a|s)}{q_h^k(s)\,\pi_h^k(a|s) + \gamma} r_h^k(s,a). \qquad (6)$$

Now, the first issue we need to address is the mismatch between $q_h(s)$ in the nominator and $q_h^k(s)$ in the denominator. A similar challenge also arises in the non-delayed analysis, but in the case of delayed feedback, this requires a carefully constructed delay-adapted local bonus (defined Eq. (4)). Note that our definition of $b_h^k(s)$ with $\beta = 3\eta/H$ implies that Eq. (6) is equal to $\sum_{k,h,s} q_h(s) b_h^k(s)$. Next, we apply the value difference lemma a second time, to show that $\text{BONUS} = \sum_{k,h,s} q_h^k(s) b_h^k(s) - \sum_{k,h,s} q_h(s) b_h^k(s)$. Essentially, this means that by summing REG and BONUS, we can substitute $q_h(s)$ in Eq. (6) by $q_h^k(s)$.

The second issue, which is unique to delayed feedback, is the mismatch between $\pi_h^{k+d^k}(a|s)$ and $\pi_h^k(a|s)$. It is important to note that while in MAB $\pi_h^{k+d^k}(a|s)/\pi_h^k(a|s)$ is always bounded by a constant, in MDPs this ratio can be as large as $\gamma^{-d_{\max}}$ (see Remark B.5 in Appendix B). Thus, the standard importance-sampling estimator (without $r_h^k(s,a)$) will not work in this type of analysis. The main idea behind our delay-adapted estimator is that $r_h^k(s,a)\,\pi_h^{k+d^k}(a|s) \le \pi_h^k(a|s)$ which guarantees that the ratio is simply bounded by 1. Overall, we get $(*) \le \text{BONUS} + H^2 \sum_{k,h,s,a} 1 = \frac{H}{\eta} + H^5 K + H^3 SAK$, and then, $\text{REG} + \text{BONUS} \le \frac{H}{\eta} + H^5 K + H^3 SAK$.

While $r_h^k(s,a)$ in our delay-adapted estimator reduces variance, it increases bias. Remarkably, this additional bias scales similarly to the DRIFT term. More specifically, for BIAS_1 we first use a variant of Freedman's inequality which is highly sensitive to the estimator's variance. This brings similar issues to the ones we faced in Eq. (6), which are treated in a similar manner. Then, we show that the additional bias introduced by the ratio $r_h^k(s,a)$ scales as

$$\begin{aligned}
&\sum_{k,h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)\big(1 - r_h^k(s,a)\big) Q_h^k(s,a) \\
&\qquad \le H \sum_{k,h,s} q_h(s) \big\| \pi_h^{k+d^k}(\cdot|s) - \pi_h^k(\cdot|s) \big\|_1, \qquad (7)
\end{aligned}$$

where the inequality follows by plugging in the definition of $r_h^k(s,a)$ and some simple algebra ($|Q_h^k(s,a)| \le H$). Utilizing the exponential weights update form, we bound the ℓ_1-distance above by $\sum_{j \in M^k} \sum_a \pi_h^{j+d^j}(a|s)\hat{Q}_h^j(s,a)$, where M^k is the set of episodes that their feedback arrives between episodes k and $k + d^k$. Then, we sum over k and apply a concentration bound over $\hat{Q}_h^k(s,a)$, which is smaller than $Q_h^k(s,a)$ in expectation (since $r_h^k(s,a) \le 1$). Thus, we get that the right-hand-side (RHS) of Eq. (7) is bounded by $H^3 \sum_k M^k$, which in turn we bound by $H^3(K + D)$ using standard delayed feedback analysis (Lemma E.10).

We can also bound the DRIFT term by the RHS of Eq. (7), up to a factor of H^2 since $|Q_h^k(s,a)| + |B_h^k(s,a)| \le H^2$. Thus, we get that $\text{DRIFT} \le H^5(K + D)$ in total. The previous state-of-the-art (Jin et al., 2022) was only able to bound the DRIFT term with an additional $\sqrt{SA}$ factor, in part due to their complex update rule that requires solving a global optimization problem. This is a great demonstration of how a simple update rule is not only beneficial on the practical side, but also for enhanced provable guarantees.

Finally, $\text{BIAS}_2 \le H/\eta$ by standard arguments for optimistic

Algorithm 2 DAPO for Linear Q-function

Initialization: Define $\pi_h^1(a|s) = 1/A$ for every (s,a,h).
for $k = 1,2,\ldots,K$ **do**
  Play episode k with policy π^k.
  # Policy Evaluation
  for j such that $j + d^j = k$ **do**
    Observe bandit feedback $\{c_h^j(s_h^j, a_h^j)\}_{h=1}^H$.
    Compute the estimated inverse covariance matrix $\{\widehat{\Sigma}_h^{j,+}\}_{h=1}^H$ via Matrix Geometric Resampling, and the estimated Q-function weights $\{\widehat{w}_h^j\}_{h=1}^H$ defined in Eq. (8).
    Define the delay-adapted estimator $\widehat{Q}_h^j(s,a)$ using Eq. (9), and estimate the bonuses $\widehat{B}_h^j(s,a)$ using Algorithm 6 with respect to the local bonuses defined in Eq. (10).
  end for
  # Policy Update
  Define the policy π^{k+1} for every (s,a,h) by:

$$\pi_h^{k+1}(a|s) \propto e^{-\sum_{j:j+d^j<k} (\widehat{Q}_h^j(s,a) - \widehat{B}_h^j(s,a))}$$

end for

estimators. To finish the proof, sum the regret from all terms and set $\eta = 1/\sqrt{SAK + H^2(K+D)}$ and $\gamma = \eta/H$. □

4. DAPO for Linear-Q

In this section we extend DAPO to linear function approximation under the Linear-Q assumption (see Assumption 2.1), which generalizes Linear MDPs (Jin et al., 2020b) and in particular is much more general than the tabular setting. This enables our algorithm to scale to MDPs with a huge (possibly infinite) number of states, and gives the first regret bound for delayed feedback in non-tabular MDPs.

DAPO for Linear-Q (presented in Algorithm 2) follows the same framework as in Section 3. That is, in each episode there is a policy evaluation step and then a policy improvement step. Since the improvement takes the same exponential weights form, we focus on the evaluation step. Specifically, we describe the new estimator $\widehat{Q}_h^k(s,a)$ and bonus $\widehat{B}_h^k(s,a)$, as these are the only changes compared to the tabular setting. Here we only provide sketches for the algorithm and analysis, but the full details are found in Appendix D.

Just like in the tabular case, our Q-function estimator will simply take the original estimator from the non-delayed setting (Luo et al., 2021) and multiply it by the delay-adapted ratio. Here, the difference between our delay adaptation approach and that of Jin et al. (2022) becomes evident. While their approach of directly changing the importance-sampling weights simply does not apply anymore, our delay-adapted ratio is easily computed locally in the current state. For completeness, we now briefly describe the estimator.

Recall that, by Assumption 2.1, the Q-function of policy π^k is parameterized by H vectors $\{w_h^k\}_{h=1}^H$, so instead of constructing an estimate $\widehat{Q}_h^k(s,a)$ for each state (which is not feasible anymore), we directly estimate w_h^k. To that end, we first construct an estimate $\widehat{\Sigma}_h^{k,+}$ of the inverse covariance matrix $(\Sigma_h^k + \gamma I)^{-1}$, where $\Sigma_h^k = \mathbb{E}_{s,a\sim q_h^k}[\phi_h(s,a)\phi_h(s,a)^\top]$ and $\gamma > 0$ is an exploration parameter. $\widehat{\Sigma}_h^{k,+}$ is computed via the Matrix Geometric Resampling procedure (Neu & Olkhovskaya, 2021) which samples trajectories of the policy using the simulator. Now, the estimator of w_h^k is defined by

$$\widehat{w}_h^k = \widehat{\Sigma}_h^{k,+}\phi_h(s_h^k, a_h^k)L_h^k, \tag{8}$$

where L_h^k is the cost-to-go from (s_h^k, a_h^k). Now, to get $\widehat{Q}_h^k(s,a)$ we compute the delay adapted-ratio $r_h^k(s,a)$ (as in Eq. (2)) and multiply it by $\phi_h(s,a)^\top \widehat{w}_h^k$, i.e.,

$$\widehat{Q}_h^k(s,a) = r_h^k(s,a)\,\phi_h(s,a)^\top \widehat{w}_h^k. \tag{9}$$

Intuitively, this estimator is a direct generalization of the tabular importance-sampling estimator (Eq. (3)) since $\widehat{\Sigma}_h^{k,+}$ corresponds to $\frac{1}{q_h^k(s,a)+\gamma}$, making $\widehat{Q}_h^k(s,a)$ an unbiased estimate of $Q_h^k(s,a)$ up to γ and approximation errors.

Next, we design the local bonuses $b_h^k(s,a)$ to go with our delay-adapted estimator. It is defined as the sum of the following local bonuses (where $\{\epsilon_i\}_i$ are parameters):

$$\begin{aligned}
b_h^{k,v}(s) &= \epsilon_v m^{k+d^k} \sum_a r_h^k(s,a)\,\pi_h^{k+d^k}(a|s)\,\|\phi_h(s,a)\|^2_{\widehat{\Sigma}_h^{k,+}} \\
b_h^{k,1}(s) &= \epsilon_1 \sum_a r_h^k(s,a)\,\pi_h^{k+d^k}(a|s)\,\|\phi_h(s,a)\|_{\widehat{\Sigma}_h^{k,+}} \\
b_h^{k,2}(s,a) &= \epsilon_2 r_h^k(s,a)\,\|\phi_h(s,a)\|_{\widehat{\Sigma}_h^{k,+}} \\
b_h^{k,r}(s,a) &= \epsilon_r(1 - r_h^k(s,a)) \\
b_h^{k,f}(s) &= \epsilon_f \sum_a r_h^k(s,a)\,\pi_h^{k+d^k}(a|s)\,\|\phi_h(s,a)\|^2_{\widehat{\Sigma}_h^{k,+}} \\
b_h^{k,g}(s,a) &= \epsilon_g r_h^k(s,a)\,\|\phi_h(s,a)\|^2_{\widehat{\Sigma}_h^{k,+}},
\end{aligned} \tag{10}$$

where $\|x\|_A = \sqrt{x^\top Ax}$ for $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$. Each of these terms plays a different important role in the analysis. $b_h^{k,v}(s)$ helps us control the variance of the estimator. It is inspired by Luo et al. (2021) and adapted to delay via the ratio $r_h^k(s,a)$. $b_h^{k,1}(s)$ and $b_h^{k,2}(s,a)$ help us to control the bias of the estimator. These, on the other hand, are constructed in a different manner than Luo et al. (2021). Importantly, the corresponding bonus terms in (Luo et al., 2021) might be of order $K^{1/3}$, while with our construction the local bonus is bounded by $O(H\sqrt{n})$. This novel construction is what allows us to avoid dilated Bellman equations in the definition of the global bonus $B_h^k(s,a)$, and by that to greatly

simplify both the algorithm and the analysis. $b_h^{k,r}(s,a)$ is specially designed to enhance exploration under delayed feedback, and in particular to control the additional bias due to the delay-adapted ratio $\frac{1}{w_h^k(s,a)}$. Finally, we add the novel terms $b_h^{k,f}(s)$ and $b_h^{k,g}(s,a)$ to ensure regret with high probability (see events $\mathcal{E}^f$ and $\mathcal{E}^g$ in Lemma D.3 in Appendix D). The exact role and interpretation of each of the bonus terms is further described through the main steps in the proof sketch of Theorem 4.1.

Given the local bonuses, we define $B_h^k(s,a)$ to be the Q-function with respect to $b_h^k(s,a)$. However, due to the possibly infinite number of states, B^k is infeasible to compute without additional structure. Instead we compute an unbiased estimate $\widehat{B}^k$ using the simulator (see further details in Appendix D). Importantly, this estimate satisfies the Bellman equations in expectation and thus inherit some of the desired properties of Q-functions such as the validity of the value difference lemma. We note that while b^k, Q^k, B^k are defined for all states, it is sufficient to calculate them on-the-fly only over the visited states. Finally, Algorithm 6 (which we borrow from Luo et al. (2021)) that computes $\widehat{B}^k$ is not sample efficient. However, with some additional structure we can replace it with an efficient procedure recently presented by Sherman et al. (2023) and obtain the same regret (see Remark D.2).

Theorem 4.1. Running DAPO in a Linear-Q adversarial MDP with $\eta = \sqrt{\ln(|A|)/(n|S|K)}$, $\gamma = \min\{\frac{1}{10Hd_{\max}}, \frac{1}{H(K+D)^{3/4}}\}$ and access to a simulator guarantees, with probability $1-\delta$, that

$$R_K = \widetilde{O}(H^3 n^{5/4} K^{3/4} + H^2 D^{3/4}).$$

This is the first sub-linear regret for non-tabular MDPs with delayed feedback. Moreover, like the optimal bound for tabular MDP, the delay term does not depend on the dimension n. Importantly, our analysis is relatively simple even compared to the non-delayed case. By that we lay solid foundations for improved regret bounds in future work, and manage to bound the regret with high probability and not just in expectation. One significant difference between the tabular case and Linear-Q is that now the estimator $\widehat{Q}_h^k(s,a)$ might be negative (specifically, $-H/\gamma$). This is not a problem in the non-delayed setting which does not have a DRAFT term, and in fact, with proper hyper-parameter tuning we get the same $\widetilde{O}(H^2 n^{2/3} K^{2/3})$ bound of Luo et al. (2021) without delays. However, in the presence of delays this issue induces new challenges and requires a more involved analysis and algorithmic design (and leads to worse regret).

Proof sketch of Theorem 4.1. We start by decomposing the regret as in Eq. (5). To bound REG we can no longer apply Corollary E.7 like we did in the tabular case, because it heavily relies on the losses being not too negative (now they might be $-O(H/\gamma)$). Instead, we prove a novel bound (Lemma E.8) that bounds REG, for sufficiently small η, by

$$\frac{H}{\eta} + \sum_{k,h} \mathbb{E}_{s\sim q_h, a\sim \pi^{k+d^k}_h} m^{k+d^k} (\widehat{Q}_h^k(s,a) - \widehat{B}_h^k(s,a))^2;$$

where $m^k = |\{j : j + d^j = k\}|$. Next, follow similar steps to Theorem 3.1. Specifically, we use $|\widehat{B}_h^k(s,a)| \le H^2\sqrt{n}$, apply a concentration bound on $\widehat{Q}_h^k(s,a)^2$ around its expectation and further bound the expectation. This allows us to show that: $\text{REG} \le \frac{H}{\eta} + H^5 nK + H^2 \sum_{k,h} \mathbb{E}_{s\sim q_h, a\sim \pi^{k+d^k}_h} m^{k+d^k} r_h^k(s,a) \|\ell_h(s,a)\|^2_{\Lambda_h^{k,+}}$.

Now we once again face the issue that the expectation is taken over states generated by q_h and actions generated by π^{k+d^k}, while $\Lambda_h^{k,+}$ is constructed from trajectories generated by π^k. Remarkably, our technique from the tabular case extends naturally to Linear-Q. Since $\widehat{B}_h^k(s,a)$ satisfies the Bellman equations in expectation, we can use the value difference lemma to show that $\text{BONUS} \le \sum_{k,h} \mathbb{E}_{s,a\sim q_h} [b_h^k(s,a)] - \sum_{k,h} \mathbb{E}_{s,a\sim q_h^k} [b_h^k(s,a)]$.

Recall that $b_h^k(s,a)$ is the sum of the 6 local bonuses defined in Eq. (10), so we can write $\text{BONUS} = \text{BONUS}_v + \text{BONUS}_1 + \text{BONUS}_2 + \text{BONUS}_r + \text{BONUS}_f + \text{BONUS}_g$, where each term corresponds to its local bonus. We set $b_h^{k,v} = H^2$ to get that $\sum_{k,h} \mathbb{E}_{s\sim q_h} [b_h^{k,v}(s)] = \Gamma(\cdot) + \text{BONUS}_v$ is bounded by,

$$H^2 \sum_{k,h} \mathbb{E}_{s\sim q_h^k, a\sim \pi^{k+d^k}_h} m^{k+d^k} r_h^k(s,a) \|\ell_h(s,a)\|^2_{\Lambda_h^{k,+}} \le H^2 \sum_{k,h} \mathbb{E}_{s\sim q_h^k, a\sim \pi^k_h} m^{k+d^k} \|\ell_h(s,a)\|^2_{\Lambda_h^{k,+}}; \quad (11)$$

where the last step is due to our delay-adapted ratio, demonstrating its power compared to directly changing the importance-sampling weights (Jin et al., 2022). It lets us adjust the expectation to be over trajectories sampled with π^k, so it is aligned with the construction of $\Lambda_h^{k,+}$ via Matrix Geometric Resampling. To further bound Eq. (11), we may now utilize standard techniques from non-delayed analysis (e.g., Jin et al. (2020b); Luo et al. (2021)). We plug in the definition of the matrix norm, then we can consider its trace and use its linearity and invariance under cyclic permutations. This enables us to bound the expectation in Eq. (11) by $\text{tr}\left(\Lambda_h^{k,+} \mathbb{E}_{s,a\sim q_h^k} \ell_h(s,a)\ell_h(s,a)^\top\right) = \text{tr}\left(\Lambda_h^{k,+} \Sigma_h^k\right)$. Finally, since $\Lambda_h^{k,+}$ approximates $(\Sigma_h^k + \gamma I)^{-1}$, the last term is approximately bounded by nK. Thus, we get that $\text{BONUS}_v + \text{REG} \le \frac{H}{\eta} + H^5 nK$ because $\sum_k m^{k+d^k} \le K$.

For the analysis of BIAS_1, we first show it is mainly bounded by two terms: the first comes from the standard estimator while the second is the additional bias due to the delay-

adapted ratio $r^k_h(s,a)$. That is, we bound BIAS_1 by:

$$H\sqrt{n}\sum_{k,h}\mathbb{E}_{s\sim q_h,a\sim \pi^{k+d^k}_h}\, r^k_h(s,a)\,\|\hat{\ell}_h(s,a)\|_{\Sigma^{k,\gamma}_h} \quad (12)$$

$$+\, H\sqrt{n}\sum_{k,h}\mathbb{E}_{s\sim q_h,a\sim \pi^{k+d^k}_h}\,(1-r^k_h(s,a))\,. \quad (13)$$

Once again, with the proper tuning $\gamma = H\sqrt{n}$, we can get $(12) = \sum_{k,h}\mathbb{E}_{s\sim q_h}[b^{k,1}_h(s)]$, and then combine with the corresponding BONUS term BONUS_1, while utilizing the delay-adapted ratio. This gives us $(12) + \text{BONUS}_1 \le H\sqrt{n}\sum_{k,h}\mathbb{E}_{s,a\sim q^k_h}\|\ell_h(s,a)\|_{(\Sigma^{k,\gamma}_h)^{-1}} \le \sqrt{H^2nK}$.

For Eq. (13), we first bound $\mathbb{E}_{a\sim\pi^{k+d^k}_h}(1-r^k_h(s,a)) \le \|\pi^{k+d^k}_h(\cdot|s) - \pi^k_h(\cdot|s)\|_1$ and then follow similar arguments to the tabular case regarding the multiplicative weights update form. The main difference is that now $\hat{Q}^k_h(s,a)$ can be negative, resulting in weaker guarantees (see more details in Appendix D.6). Overall, we get $(13) \le H^3\sqrt{n}(K+D)$.

The analysis of BIAS_2 is very different from both the tabular case and the non-delayed Linear-Q (Luo et al., 2021). This is mainly for two reasons: First, in the tabular case, the added bias makes the estimator $\hat{Q}^k_h(s,a)$ optimistic, but this is no longer the case in Linear-Q since now the estimator might also be negative. Second, BIAS_2 contains the inner product with $\ell_h(\cdot|s)$ which cannot be aligned with the denominator of the delay-adapted ratio $r^k_h(s,a)$. Thus, we need a novel more involved analysis for BIAS_2.

We start in a similar way to BIAS_1, and bound BIAS_2 by:

$$H\sqrt{n}\sum_{k,h}\mathbb{E}_{s,a\sim q_h}\, r^k_h(s,a)\,\|\hat{\ell}_h(s,a)\|_{\Sigma^{k,\gamma}_h} \quad (14)$$

$$+\, H\sqrt{n}\sum_{k,h}\mathbb{E}_{s,a\sim q_h}\,(1-r^k_h(s,a))\,. \quad (15)$$

We handle Eq. (14) like Eq. (12), so for $\gamma = H\sqrt{n}$ we get that $(14) + \text{BONUS}_2 \le \sqrt{H^2nK}$. Term (15) on the other hand might be of order K due to mismatch between $\ell_h(a|s)$ in the expectation and $\max\{\pi^{k+d^k}_h(a|s), \pi^k_h(a|s)\}$ in the denominator of $r^k_h(s,a)$. To address that, we design the novel bonus term $b^{k,r}_h(s,a)$. Summing Eq. (15) with BONUS_r essentially allows us to substitute $\ell_h(a|s)$ by $\pi^k_h(a|s)$, which gives: $(14) + \text{BONUS}_r \le \sum_{k,h}\mathbb{E}_{s\sim q_h}\|\pi^{k+d^k}_h(\cdot|s) - \pi^k_h(\cdot|s)\|_1 \le H^3\sqrt{n}(K+D)$.

Finally, DRIFT is also bounded by the ℓ_1-distance above and further by $H^4(K+D)$. Summing the regret from all terms and optimizing over η and γ completes the proof. □

5. Delay-Adapted PPO and Experiments

In this section we show how our generic delay adaptation method extends to state-of-the-art deep RL methods, and demonstrate its great potential through simple experiments on popular MuJoCo environments (Todorov et al., 2012). We note that here we follow the standard convention that use rewards in deep RL rather than costs as in the rest of the paper. Due to lack of space, here we only provide an overview of the method and main experiment. For additional experiments and full implementation details see Appendix F.

The highly successful TRPO algorithm (Schulman et al., 2015) is a deep RL policy-gradient method that builds on the same PO principles discussed in this paper. Specifically, it follows the Policy Iteration paradigm with a "soft" policy improvement step, where the policy is now approximated by a Deep Neural Network with parameters θ. While our update step (Eq. (1)) is equivalent to maximizing an objective with a KL-regularization term (Shani et al., 2020a), TRPO replaces it by a constraint which results in maximizing the objective $L^k_{\text{TRPO}}(\theta) = \sum_{h=1}^H \frac{\pi_\theta(a_h|s_h)}{\pi^k(a_h|s_h)}\hat{A}^k_h$ subject to a constraint that keeps the new and the old policies close in terms of KL-divergence. Here $\hat{A}^k_h$ is an estimate of the advantage function which replaces the Q-function in our formulation to further reduce variance (Sutton et al., 1999).

While successful, TRPO's constrained optimization is computationally expensive. Thus, PPO (Schulman et al., 2017) removes the explicit constraint and replaces it with a sophisticated clipping technique that allows to keep strong empirical performance while only optimizing the following (non-constrained) objective:

$$L^k(\theta) = \sum_{h=1}^H \min\left\{g^k_h(\theta)\hat{A}^k_h;\ \text{clip}_1\left(g^k_h(\theta)\right)\hat{A}^k_h\right\}, \quad (16)$$

where $g^k_h(\theta) = \frac{\pi_\theta(a_h|s_h)}{\pi^k(a_h|s_h)}$ and $\text{clip}_1(x)$ clips x between $1-\varepsilon$ and $1+\varepsilon$. This objective essentially zeros the gradient whenever the policy changes too much, and thus replaces the need for an explicit constraint (see more details in (Schulman et al., 2017)). Next, we adapt PPO to delayed feedback.

With delayed feedback, the trajectory that arrives at time k was generated using policy π^{k-d^k}. Thus, a naive adaptation would be to optimize Eq. (16) but replacing π^k with π^{k-d^k}. We will refer to this algorithm as Delayed PPO (DPPO). As this paper shows, DPPO is likely to suffer from large variance which can be reduced when multiplying the objective by the delay-adapted ratio $r^k_h(s,a)$. The result is our novel Delay-Adapted PPO (DAPPO) which optimizes:

$$L^k_{\text{DA}}(\theta) = \sum_{h=1}^H \min\left\{R^k_h(\theta)\hat{A}^k_h;\ \text{clip}_1\left(R^k_h(\theta)\right)\hat{A}^k_h\right\},$$

for $R^k_h(\theta) = \frac{\pi_\theta(a_h|s_h)}{\max\{\pi^{k-d^k}(a_h|s_h);\ \pi^k(a_h|s_h)\}}$. Another alternative, which we call Non-Delayed PPO (NDPPO), is to run the original PPO and ignore the fact that feedback is

Figure 1.Training curves: DAPPO vs DPPO.Plots show average reward and std over 5 seeds. x-axis is number of timesteps up to 5M.

Figure 2.Training curves with different xed delay length: DAPPO vs DPPO with different delay, alongside PPO without delays. Plots show average reward and std over 5 seeds. x-axis is number of timesteps up to 5M.

delayed. This results in a highly unstable algorithm which is likely to suffer from large bias due to the mismatch between the policy $^{k\ d}$ that generated the trajectory and the policy ^{k} which is used to re-weight the estimator.

Fig. 1 compares the performance of DPPO and DAPPO over 8 MuJoCo environments. We use a xed delay of 10^5 timesteps while the total number of timestamps is $5 \cdot 10^6$. Results are averaged over 5 runs and the shaded areas around the curves indicate standard deviation. Note that the only difference between the algorithms is the objective, we did not tune hyper-parameters or modify network architecture. DAPPO outperforms DPPO in at least 4 environments and is on par with DPPO in the rest. The only exception is InvertedDoublePendulum , but it is important to note that this environment is extremely noisy.

Fig. 2 compares the training curves of DPPO vs DAPPO in the SWIMMER environment (for more environments see Appendix F.3) with different delays in $\{10000, 25000, 50000, 75000, 100000\}$, alongside the training curve of PPO without delay. As expected, when the delay is relatively small (e.g. 10000), there is no signi cant difference between learning with or without delayed feedback. As the delay becomes larger, the performance of all algorithms drops (but at different rates).

These empirical results support our claim that handling delays via the delay-adapted ratio extends naturally beyond the tabular and Linear Q settings to practical deep function approximation. Surprisingly, even in this simple case, delays cause signi cant drop in performance which demonstrates the great importance of delay-adapted algorithms. Our novel method makes a signi cant step towards practical deep RL algorithms that are robust to delayed feedback. Finally, we note that NDPPO is omitted from the graphs because it does not converge (as expected). Instead, it oscillates between high and low reward (see Appendix F for more details).

6. Future Work

We leave a few open questions for future work. In the tabular case, it still remains unclear what is the optimal dependency under delayed feedback in terms of the horizon H. Another future direction is to further improve our results in the Linear Function Approximation setting and extend them to the case where a simulator is unavailable. This is in particular important in light of very recent advancement in the non-delayed setting (Sherman et al., 2023; Dai et al., 2023) which signi cantly improve Luo et al. (2021). Finally, our experiment demonstrate the potential that delay-adaptation methods have for deep RL applications. However, a much more thorough empirical study needs to be done in order to fully understand the implications of these methods on deep RL.

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. Politex: Regret bounds for policy iteration using expert prediction. In International Conference on Machine Learning, pp. 3692–3702. PMLR, 2019.

Agarwal, A. and Duchi, J. C. Distributed delayed stochastic optimization. In 2012 IEEE 51st IEEE Conference on Decision and Control (CDC), pp. 5451–5452. IEEE, 2012.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On theory, pp. 64–66. PMLR, 2020.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning, pp. 463–474. PMLR, 2020.

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.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 19–26. JMLR Workshop and Conference Proceedings, 2011.

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.

Bistritz, I., Zhou, Z., Chen, X., Bambos, N., and Blanchet, J. No discounted-regret learning in adversarial bandits with delays. arXiv preprint arXiv:2103.04552, 2021.

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.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. Delay and cooperation in nonstochastic bandits. The Journal of Machine Learning Research, 20(1):613–650, 2019.

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.

Chen, L., Jafarnia-Jahromi, M., Jain, R., and Luo, H. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. Advances in Neural Information Processing Systems, 2021.

Chen, L., Jain, R., and Luo, H. Improved no-regret algorithms for stochastic shortest path with linear mdp. In International Conference on Machine Learning, pp. 3204–3245. PMLR, 2022a.

Chen, L., Luo, H., and Rosenberg, A. Policy optimization for stochastic shortest path. In Loh, P. and Raginsky, M. (eds.), Conference on Learning Theory, 2-5 July 2022, London, UK, volume 178 of Proceedings of Machine Learning Research, pp. 982–1046. PMLR, 2022b.

Cohen, A., Daniely, A., Drori, Y., Koren, T., and Schain, M. Asynchronous stochastic optimization robust to arbitrary delays. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pp. 9024–9035, 2021a.

Cohen, A., Efroni, Y., Mansour, Y., and Rosenberg, A. Minimax regret for stochastic shortest path. Advances in Neural Information Processing Systems, 34, 2021b.

Dai, Y., Luo, H., and Chen, L. Follow-the-perturbed-leader for adversarial markov decision processes with bandit feedback. arXiv preprint arXiv:2205.13457, 2022.

Dai, Y., Luo, H., Wei, C.-Y., and Zimmert, J. Refined regret for adversarial mdps with linear function approximation. arXiv preprint arXiv:2301.12942, 2023.

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.

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

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Ef cient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Arti cial Intelligence*, pp. 169–178, 2011.

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.

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

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.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3389–3396. IEEE, 2017.

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

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.

He, J., Zhou, D., and Gu, Q. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Arti cial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4259–4280. PMLR, 28–30 Mar 2022.

Howson, B., Pike-Burke, C., and Filippi, S. Delayed feedback in episodic reinforcement learning. *arXiv preprint arXiv:2111.07615*, 2021.

Howson, B., Pike-Burke, C., and Filippi, S. Delayed feedback in generalised linear bandits revisited. *arXiv preprint arXiv:2207.10786*, 2022.

Ito, S., Hatano, D., Sumita, H., Takemura, K., Fukunaga, T., Kakimura, N., and Kawarabayashi, K.-I. Delay and cooperation in nonstochastic linear bandits. *Advances in Neural Information Processing Systems*, 33:4872–4883, 2020.

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 ef cient? 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 ef cient 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.

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

Jin, T., Lancewicki, T., Luo, H., Mansour, Y., and Rosenberg, A. Near-optimal regret for adversarial mdp with delayed bandit feedback. *arXiv preprint arXiv:2201.13178*, 2022.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Kakade, S. M. A natural policy gradient. Advances in neural information processing systems, 14:1531–1538, 2001.

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.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Lancewicki, T., Segal, S., Koren, T., and Mansour, Y. Stochastic multi-armed bandits with unrestricted delay distributions. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, pp. 5969–5978. PMLR, 2021.

Lancewicki, T., Rosenberg, A., and Mansour, Y. Cooperative online learning in stochastic and adversarial mdps. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of Proceedings of Machine Learning Research, pp. 11918–11968. PMLR, 2022a.

Lancewicki, T., Rosenberg, A., and Mansour, Y. Learning adversarial markov decision processes with delayed feedback. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pp. 7281–7289, 2022b.

Levine, S. and Koltun, V. Guided policy search. In International conference on machine learning, pp. 1–9. PMLR, 2013.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research, 17(1):1334–1373, 2016.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971, 2015.

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.

Luo, H., Wei, C.-Y., and Lee, C.-W. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. Advances in Neural Information Processing Systems, 34, 2021.

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) 4635–4640. IEEE, 2018.

Masoudian, S., Zimmert, J., and Seldin, Y. A best-of-both-worlds algorithm for bandits with delayed feedback. arXiv preprint arXiv:2206.14906, 2022.

Min, Y., He, J., Wang, T., and Gu, Q. Learning stochastic shortest path with linear function approximation. In International Conference on Machine Learning, pp. 15584–15629. PMLR, 2022.

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

Neu, G. and Olkhovskaya, J. Online learning in mdps with linear function approximation and bandit feedback. Advances in Neural Information Processing Systems, 34, 10407–10417, 2021.

Neu, G., György, A., and Szepesvári, C. The online loop-free stochastic shortest-path problem. In COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010, pp. 231–243, 2010a.

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

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.

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.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. Stable-baselines3: Reliable reinforcement learning implementations. Journal of Machine Learning Research, 2021.

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. PMLR, 2019b.

Rosenberg, A. and Mansour, Y. Oracle-ef cient regret minimization in factored mdps with unknown structure. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 11148–11159, 2021a.

Rosenberg, A. and Mansour, Y. Stochastic shortest path with adversarially changing costs. In Zhou, Z. (ed.), *Proceedings of the Thirtieth International Joint Conference on Arti cial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 2936–2942. ijcai.org, 2021b.

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. *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., and Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *The Thirty-Fourth AAAI Conference on Arti cial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Arti cial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Arti cial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 5668–5675. AAAI Press, 2020a.

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, 2020b.

Sherman, U., Koren, T., and Mansour, Y. Improved regret for ef cient online reinforcement learning with linear function approximation. *arXiv preprint arXiv:2301.13087*, 2023.

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

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

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.

Tarbouriech, J., Zhou, R., Du, S. S., Pirotta, M., Valko, M., and Lazaric, A. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Tomar, M., Shani, L., Efroni, Y., and Ghavamzadeh, M. Mirror descent policy optimization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. 2022.

Van Der Hoeven, D. and Cesa-Bianchi, N. Nonstochastic bandits and experts with arm-dependent delays. In *International Conference on Arti cial Intelligence and Statistics*. PMLR, 2022.

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

Vernade, C., Carpentier, A., Lattimore, T., Zappella, G., Ermis, B., and Brueckner, M. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pp. 9712–9721. PMLR, 2020.

Vial, D., Parulekar, A., Shakkottai, S., and Srikant, R. Regret bounds for stochastic shortest path problems with linear function approximation. In *International Conference on Machine Learning*, pp. 22203–22233. PMLR, 2022.

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.

Wei, C.-Y., Jahromi, M. J., Luo, H., and Jain, R. Learning in nite-horizon average-reward mdps with linear function approximation. In *International Conference on Arti cial Intelligence and Statistics*, pp. 3007–3015. PMLR, 2021.

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., 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, D. and Gu, Q. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. arXiv preprint arXiv:2205.11507, 2022.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In Conference on Learning Theory, pp. 4532–4576. PMLR, 2021.

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. 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.

Appendix

A. Related Work

In this section we provide a full review of the literature related to regret minimization in adversarial MDP with delayed feedback. For completeness, we include topics that are not directly related to this paper.

Delays in RL without Regret Analysis. Delays were studied in the practical RL literature (Schuitema et al., 2010; Liu et al., 2014; Changuel et al., 2012; Mahmood et al., 2018; Derman et al., 2021), but this is not related the topic of this paper. In the theory literature, most previous work (Katsikopoulos & Engelbrecht, 2003; Walsh et al., 2009) considered delays in the observation of the state. That is, when the agent takes an action she is not certain what is the current state, and will only observe it in delay. This setting is much more related to partially observable MDPs (POMDPs) and motivated by scenarios like robotics system delays. Unfortunately, even planning is computationally hard (exponential in the delay) for delayed state observability (Walsh et al., 2009). The topic studied in this paper (i.e., delayed feedback) is inherently different, and is motivated by settings like recommendation systems. Importantly, unlike delayed state observability, it is not computationally hard to handle delayed feedback. The challenges of delayed feedback are very different than the ones of delayed state observability, and include policy updates that occur in delay and exploration without observing feedback (Lancewicki et al., 2022b).

Delays in RL with Regret Analysis. This line of work is related to this paper the most. Howson et al. (2021) studied delayed feedback in stochastic MDPs, and assume that the delays are also stochastic, i.e., sampled i.i.d from a fixed (unknown) distribution. This is a restrictive assumption since it does not allow dependencies between costs and delays that are very common in practice. In adversarial MDPs, delayed feedback was first studied by Lancewicki et al. (2022b). They proposed Policy Optimization algorithms that handle delays, but focused on the case of full-information feedback where the agent observes the entire cost function in the end of the episode instead of bandit feedback (where the agent observes only costs along its trajectory). Full-information feedback is not a realistic assumption in most applications, and for bandit feedback they only prove sub-optimal regret of $\widetilde{O}(H^2S^2A(K+D)^{2/3})$ (ignoring dependencies in S, A, H). Later, Jin et al. (2022) managed to achive a near-optimal regret bound of $\widetilde{O}(H^2S^2\sqrt{AK} + (HSA)^{1/4}H\sqrt{D})$ for the case of known transition function and $\widetilde{O}(H^2S^3\sqrt{AK} + (HSA)^{1/4}H\sqrt{D})$ for the case of unknown transitions. However, their algorithm is based on the O-REPS method (Zimin & Neu, 2013) which requires solving a computationally expensive global optimization problem and cannot be extended to function approximation. Recently, Dai et al. (2022) showed that delayed feedback in adversarial MDPs can also be dealt with using Follow-The-Perturbed-Leader (FTPL) algorithms. The efficiency of FTPL algorithms is similar to Policy Optimization, but their regret bound is only $\widetilde{O}(H^2S\sqrt{AK} + \sqrt{HSA}H\sqrt{D})$.

Delays in multi-arm bandit (MAB). Delays were extensively studied in MAB and online optimization both in the stochastic setting (Dudik et al., 2011; Agarwal & Duchi, 2012; Vernade et al., 2017; 2020; Pike-Burke et al., 2018; Cesa-Bianchi et al., 2018; Zhou et al., 2019; Gael et al., 2020; Lancewicki et al., 2021; Cohen et al., 2021a; Howson et al., 2022), and the adversarial setting (Quanrud & Khashabi, 2015; Cesa-Bianchi et al., 2016; Thune et al., 2019; Bistritz et al., 2019; Zimmert & Seldin, 2020; Ito et al., 2020; Gyorgy & Joulani, 2021; Van Der Hoeven & Cesa-Bianchi, 2022; Masoudian et al., 2022). However, as discussed in (Lancewicki et al., 2022b), delays introduce new challenges in MDPs that do not appear in MAB.

Regret minimization in Tabular RL. There exists a rich literature on regret minimization in tabular MDPs. In the stochastic case, the algorithms are mainly built on the optimism in face of uncertainty approach (Jaksch et al., 2010; Azar et al., 2017; Dann et al., 2017; Jin et al., 2018; Efroni et al., 2019; Tarbouriech et al., 2020; 2021; Rosenberg & Mansour, 2021a; Rosenberg et al., 2020; Cohen et al., 2021b; Chen et al., 2021). In the adversarial case, while a few algorithms use FTPL (Neu et al., 2012; Dai et al., 2022), most algorithms are based on either the O-REPS method (Zimin & Neu, 2013; Rosenberg & Mansour, 2019b;a; 2021b; Jin et al., 2020a; 2021; Jin & Luo, 2020; Lancewicki et al., 2022a; Chen et al., 2020b; Chen & Luo, 2021) or on Policy Optimization (Even-Dar et al., 2009; Neu et al., 2010a;b; 2014; Shani et al., 2020b; Luo et al., 2021; Chen et al., 2022b). Note that regret minimization in standard episodic MDPs is a special case of the model considered in this paper where $d^k = 0$ for every episode k.

Regret minimization in RL with Linear Function Approximation. In recent years the literature on regret minimization in RL has expanded to linear function approximation. While in the stochastic case algorithms are still based on optimism (Jin et al., 2020b; Yang & Wang, 2019; Zanette et al., 2020a;b; Ayoub et al., 2020; Zhou et al., 2021; Zhou & Gu, 2022; Vial et al., 2022; Chen et al., 2022a; Min et al., 2022), in the adversarial case O-REPS cannot be extended to linear function

approximation without additional assumptions so algorithms are mostly based on Policy Optimization (Cai et al., 2020; Abbasi-Yadkori et al., 2019; Agarwal et al., 2020; He et al., 2022; Neu & Olkhovskaya, 2021; Luo et al., 2021; Wei et al., 2021).

Policy Optimization in Deep RL. Policy Optimization is among the most widely used methods in deep Reinforcement Learning (Lillicrap et al., 2015; Levine et al., 2016; Gu et al., 2017). The origins of these algorithms are Policy Gradient (Sutton & Barto, 2018), Conservative Policy Iteration (Kakade & Langford, 2002) and Natural Policy Gradient (Kakade, 2001). These have evolved into some of the state-of-the-art algorithms in RL, e.g., Trust Region Policy Optimization (TRPO) (Schulman et al., 2015), Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Soft Actor-Critic (SAC) (Haarnoja et al., 2018). Recently, the connections between deep RL policy optimization algorithms and online learning regularization methods (like Follow-The-Regularized-Leader and Online-Mirror-Descent) were studied and explained (Shani et al., 2020a; Tomar et al., 2022).

Remark A.1 (The Loop-Free Assumption.) We warn the readers that some of the works mentioned in this section (mainly in the adversarial MDP literature, e.g., Luo et al. (2021)) present a slightly different dependence in the Horizon H. The reason is that they make the loop-free assumption, i.e., they assume that the state space consists of H disjoint sets $S = S_1 \cup S_2 \cup \cdots \cup S_H$ such that in step h the agent can only be found in states from the set S_h. Effectively, this means that their state space is larger than ours by a factor of H. So when they present a regret bound of $\widetilde{O}(H^2 S \sqrt{AK})$, this implies a bound of $\widetilde{O}(H^3 S \sqrt{AK})$ in the model presented in this paper. We emphasize that these differences are only due to different models, and not due to actual different regret bounds.

B. Delay-Adapted Policy Optimization for (Tabular) Adversarial MDP with Known Transition

Algorithm 3 Delay-Adapted Policy Optimization with Known Transition Function (Tabular)

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 trajectory $\{(s_h^k, a_h^k)\}_{h=1}^H$ and compute $c_h^k(s)$ for every $(s,h) \in \mathcal{S} \times [H]$.
  # Policy Evaluation
  for j such that $j + d^j = k$ do
    Observe bandit feedback $\{c_h^j(s_h^j, a_h^j)\}_{h=1}^H$ and set $B_{H+1}^j(s,a) = 0$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$.
    for $h = H, H-1, \ldots, 1$ do
      for $(s,a) \in \mathcal{S} \times \mathcal{A}$ do
        Compute $r_h^j(s,a) = \frac{\pi_h^j(a|s)}{\max\{q_h^j(s,a), \pi_h^k(a|s)\gamma\}}$ and $L_h^j = \sum_{h'=h}^H c_{h'}^j(s_{h'}^j, a_{h'}^j)$.
        Compute $\widehat{Q}_h^j(s,a) = r_h^j(s,a)\frac{\mathbb{I}\{s_h^j = s, a_h^j = a\} L_h^j}{q_h^j(s)\pi_h^j(a|s) + \gamma}$ and $b_h^j(s) = \sum_{a \in \mathcal{A}} \frac{3H \pi_h^k(a|s) r_h^j(s,a)}{q_h^j(s)\pi_h^j(a|s) + \gamma}$.
        Compute $B_h^j(s,a) = b_h^j(s) + \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} p_h(s' \mid s,a)\pi_{h+1}^j(a' \mid s') B_{h+1}^j(s',a')$.
      end for
    end for
  end for
  # Policy Improvement
  Define the policy π^{k+1} for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ by:

$$\pi_h^{k+1}(a \mid s) = \frac{\pi_h^k(a \mid s) \exp\left(\sum_{j: j + d^j = k} (\widehat{Q}_h^j(s,a) - B_h^j(s,a))\right)}{\sum_{a' \in \mathcal{A}} \pi_h^k(a' \mid s) \exp\left(\sum_{j: j + d^j = k} (\widehat{Q}_h^j(s,a') - B_h^j(s,a'))\right)}.$$

end for

Theorem B.1. Set $\eta = \sqrt{H^2 SAK + H^4(K+D)}^{-1/2}$ and $\gamma = 2\eta H$. Running Algorithm 3 in an adversarial MDP $M = (\mathcal{S}, \mathcal{A}, H, p, \{c^k\}_{k=1}^K)$ with known transition function and delays $\{d^k\}_{k=1}^K$ guarantees, with probability at least $1-\delta$,

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

B.1. The good event

Let $\iota = 10\log\frac{10KHSA}{\delta}$, $\mathcal{H}^k$ be the history of episodes $\{j : j + d^j < k\}$, and define $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{H}^{k+d^k}]$. Define the following events:

$$\mathcal{E}^d = \left\{\forall (h,s): \sum_{k=1}^K \sum_{j=1}^K \sum_a \mathbb{I}\{j + d^k < j + d^j\} \pi_h^{k+d^k}(a \mid s)(\widehat{Q}_h^k(s,a) - Q_h^k(s,a)) \le \frac{10H^2 d_{\max}\log\frac{10HS}{\delta}}{\gamma}\right\},$$

$$\mathcal{E} = \left\{\sum_{k=1}^K \sum_{h,s,a} q_h(s,a)(\widehat{Q}_h^k(s,a) - Q_h^k(s,a)) \le \frac{H^2\log\frac{10HSA}{\delta}}{\gamma}\right\},$$

$$\mathcal{E}^b = \left\{\sum_{k=1}^K \sum_{h,s,a} \frac{q_h(s)\pi_h^{k+d^k}(a \mid s) r_h^k(s,a)\mathbb{I}\{s_h^k = s, a_h^k = a\}}{(q_h^k(s,a) + \gamma)^2} \le \sum_{k=1}^K \sum_{h,s,a} \frac{q_h(s)\pi_h^{k+d^k}(a \mid s) r_h^k(s,a)}{q_h^k(s,a) + \gamma} + \frac{H}{2}\ln\frac{10H}{\delta}\right\},$$

$$\mathcal{E}^f = \left\{\sum_{k=1}^K \mathbb{E}_k\left[\sum_{h,s} q_h(s)\left\langle \pi_h^{k+d^k}(\cdot \mid s), \widehat{Q}_h^k(s,\cdot)\right\rangle\right] \le \sum_{k=1}^K \sum_{h,s} q_h(s)\left\langle \pi_h^{k+d^k}(\cdot \mid s), \widehat{Q}_h^k(s,\cdot)\right\rangle + \frac{1}{3}\sum_{k=1}^K \sum_{h,s} q_h(s) b_h^k(s) + \frac{H^2}{\gamma}\ln\frac{10}{\delta}\right\},$$

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

Lemma B.2 (The Good Event) Let $G = E^d \cap E \cap E^b \cap E^f$ be the good event. It holds that $\Pr[G] \geq 1 - \delta$.

Proof. We'll show that each of the events E^d, E, E^b, E^f holds with probability of at most $\delta/4$ and so by the union bound $\Pr[G] \geq 1 - \delta$.

Event E^d: Fix s and h. For every (h', s', a) set:

$$z^k_{h'}(s', a) = \mathbb{I}\{s'_h = s, h' = h\} \sum_{j=1}^{K} \mathbb{I}\{j - d^j = k - d^k\} \pi^{k+d^k}_h(a \mid s) r^k_h(s, a) Q^j_h(s, a)$$

$$Z^k_{h'}(s', a) = \mathbb{I}\{s'_h = s, h' = h\} \sum_{j=1}^{K} \mathbb{I}\{j - d^j = k - d^k\} \pi^{k+d^k}_h(a \mid s) r^k_h(s, a) M^k_h(s, a)$$

$$M^k_h(s, a) = \mathbb{I}\{s^k_h = s, a^k_h = a\} \sum_{h'=1}^{H} c^k_{h'}(s^k_{h'}, a^k_{h'}) + (1 - \mathbb{I}\{s^k_h = s, a^k_h = a\}) Q^k_h(s, a).$$

Note that $\mathbb{E}_k[Z^k_{h'}(s', a)] = z^k_{h'}(s', a)$ and,

$$\sum_{h', s', a} \frac{\mathbb{I}\{s^k_h = s, a^k_h = a\} Z^k_h(s, a)}{q^k_h(s, a) + \gamma} = \sum_{j=1}^{K} \sum_{a} \mathbb{I}\{j - d^j = k - d^k\} \pi^{k+d^k}_h(a \mid s) \hat{Q}^k_h(s, a)$$

$$\leq \sum_{h', s', a} z^k_h(s, a) \sum_{j=1}^{K} \sum_{a} \mathbb{I}\{j - d^j = k - d^k\} \pi^{k+d^k}_h(a \mid s) Q^k_h(s, a),$$

where in the inequality we used the fact that $\hat{Q}^k_h(s, a) \leq 1$. Finally, we use Lemma E.5 with $\eta^k_h(s, a) = q^k_h(s, a)$ and $R = 2H d_{max}$ since the number of js such that $j - d^j = k - d^k$ is at most $2d_{max}$. Thus, the event holds for (s, h) with probability $1 - \frac{\delta}{10HS}$. By taking the union bound over all s and h, E^d holds with probability $1 - \frac{\delta}{10}$.

Event E (Lemma C.2 of Luo et al. (2021)): E holds with probability of at least $1 - \delta/10$ by applying Lemma E.5 with $\eta^k_{h'}(s', a) = q^k_{h'}(s', a)$, $z^k_h(s, a) = q_h(s, a) r^k_h(s, a) Q^k_h(s, a)$ and,

$$Z^k_h(s, a) = q_h(s, a) r^k_h(s, a) \left(\mathbb{I}\{s^k_h = s, a^k_h = a\} \sum_{h'=1}^{H} c^k_{h'}(s^k_{h'}, a^k_{h'}) + (1 - \mathbb{I}\{s^k_h = s, a^k_h = a\}) Q^k_h(s, a) \right).$$

Note that $R \leq H$, $\frac{\mathbb{I}\{s^k_h = s, a^k_h = a\} Z^k_h(s, a)}{q^k_h(s, a) + \gamma} = q_h(s, a) \hat{Q}^k_h(s, a)$ and $\frac{q^k_h(s, a) z^k_h(s, a)}{q^k_h(s, a)} \leq q_h(s, a) Q^k_h(s, a)$.

Event E^b: Similar to the last two events, E^b holds with probability of at least $1 - \delta/10$ by applying Lemma E.5 with $\eta^k_{h'}(s', a) = q^k_{h'}(s', a)$ and $z^k_h(s, a) = Z^k_h(s, a) = \frac{q_h(s) \pi^{k+d^k}_h(a \mid s) r^k_h(s, a)}{q^k_h(s, a) + \gamma}$.

Event E^f: Let $Y_k = \sum_{h, s} q_h(s) \left\langle \pi^{k+d^k}_h(\cdot \mid s), \hat{Q}^k_h(s, \cdot) \right\rangle$. We'll use a variant of Freedman's inequality (Lemma E.3) to

bound $\sum_{k=1}^{K} \mathbb{E}_k[Y_k] \le \sum_{k=1}^{K} Y_k$. Note that:

$$\begin{aligned}
\mathbb{E}_k\left[Y_k^2\right] &= \mathbb{E}_k\left[\left(\sum_{h,s,a} q_h(s)\,\pi_h^{k+d^k}(a\,|\,s)\widehat{Q}_h^k(s,a)\right)^2\right] \\
&\le \mathbb{E}_k\left[\left(\sum_{h,s,a} q_h(s)\,\pi_h^{k+d^k}(a\,|\,s)\right)\left(\sum_{h,s,a} q_h(s)\,\pi_h^{k+d^k}(a\,|\,s)\left(\widehat{Q}_h^k(s,a)\right)^2\right)\right] \\
&\le H\,\mathbb{E}_k\left[\sum_{h,s,a} q_h(s)\,\pi_h^{k+d^k}(a\,|\,s)\frac{r_h^k(s,a)^2(L_h^k)^2\,\mathbb{I}\{s_h^k=s,a_h^k=a\}}{(q_h^k(s,a)+\gamma)^2}\right] \\
&\le H^3\sum_{h,s,a} q_h(s)\,\pi_h^{k+d^k}(a\,|\,s)\frac{q_h^k(s,a)r_h^k(s,a)}{(q_h^k(s,a)+\gamma)^2} \\
&\le H^3\sum_{h,s,a}\frac{q_h(s)\,\pi_h^{k+d^k}(a\,|\,s)r_h^k(s,a)}{q_h^k(s,a)+\gamma},
\end{aligned}$$

where the first inequality is Cauchy-Schwartz inequality. Also $|Y_k| \le \frac{H^2}{\gamma}$. Therefore by Lemma E.3 with probability $1-\delta/10$,

$$\begin{aligned}
\sum_{k=1}^{K}\mathbb{E}_k[Y_k] &\le \sum_{k=1}^{K} Y_k + H^3\sum_{k=1}^{K}\sum_{h,s,a}\frac{q_h(s)\,\pi_h^{k+d^k}(a\,|\,s)r_h^k(s,a)}{q_h^k(s,a)+\gamma} + \frac{H^2}{\gamma}\ln\frac{10}{\delta} \\
&= \frac{1}{3}\sum_{k=1}^{K}\sum_{h,s} q_h(s)b_h^k(s) + \frac{H^2}{\gamma}\ln\frac{10}{\delta}. \qquad \square
\end{aligned}$$

B.2. Proof of the main theorem

Proof of Theorem B.1 By Lemma B.2, the good event holds with probability $1-\delta$. We now analyze the regret under the assumption that the good event holds. We start with the following regret decomposition,

$$\begin{aligned}
R_K &= \sum_{k=1}^{K}\sum_{h,s} q_h(s)\left\langle \pi_h^k(\cdot\,|\,s) - \pi_h(\cdot\,|\,s), Q_h^k(s,\cdot)\right\rangle = \underbrace{\sum_{k=1}^{K}\sum_{h,s} q_h(s)\left\langle \pi_h^{k+d^k}(\cdot\,|\,s), Q_h^k(s,\cdot) - \widehat{Q}_h^k(s,\cdot)\right\rangle}_{\textsc{Bias}_1} \\
&+ \underbrace{\sum_{k=1}^{K}\sum_{h,s} q_h(s)\left\langle \pi_h^{k+d^k}(\cdot\,|\,s) - \pi_h(\cdot\,|\,s), \widehat{Q}_h^k(s,\cdot) - Q_h^k(s,\cdot)\right\rangle}_{\textsc{Bias}_2} + \underbrace{\sum_{k=1}^{K}\sum_{h,s} q_h(s)\left\langle \pi_h^k(\cdot\,|\,s) - \pi_h(\cdot\,|\,s), B_h^k(s,\cdot)\right\rangle}_{\textsc{Bonus}} \\
&+ \underbrace{\sum_{k=1}^{K}\sum_{h,s} q_h(s)\left\langle \pi_h^{k+d^k}(\cdot\,|\,s) - \pi_h(\cdot\,|\,s), \widehat{Q}_h^k(s,\cdot) - B_h^k(s,\cdot)\right\rangle}_{\textsc{Reg}} \\
&+ \underbrace{\sum_{k=1}^{K}\sum_{h,s} q_h(s)\left\langle \pi_h^k(\cdot\,|\,s) - \pi_h^{k+d^k}(\cdot\,|\,s), Q_h^k(s,\cdot) - B_h^k(s,\cdot)\right\rangle}_{\textsc{Drift}},
\end{aligned}$$

where the first equality is by Lemma E.1 (value difference lemma), BIAS_2 is bounded under event $\mathcal{E}$ by $O\left(\frac{H^2}{}\right)$. The other four terms are bounded in Lemmas B.3, B.4, B.6 and B.7. Overall,

$$\begin{aligned}
R_K \le\ & \underbrace{\frac{2}{3}\sum_{k=1}^{K}\sum_{h,s} q_h(s)b_h^k(s) + O\left(H^4(K+D) + \frac{H^4 d_{\max}}{} + \frac{H^2}{}\right)}_{\text{BIAS}_1} + \underbrace{O\left(\frac{H^2}{}\right)}_{\text{BIAS}_2} \\
& + 3H^2SAK\underbrace{\sum_{k=1}^{K}\sum_{h,s} q_h(s)b_h^k(s)}_{\text{BONUS}} + \underbrace{\frac{H\ln A}{} + \frac{1}{3}\sum_{k=1}^{K}\sum_{h,s} q_h(s)b_h^k(s)}_{\text{REG}} + O\left(H^5K + \frac{H^3}{2}\right) \\
& + \underbrace{O\left(H^5(K+D) + \frac{H^5 d_{\max}}{}\right)}_{\text{DRIFT}} \\
\le\ & O\left(\frac{H\ln A}{} + H^2SAK + H^5(K+D) + \frac{H^3}{2} + \frac{H^2}{} + \frac{H^5 d_{\max}}{}\right).
\end{aligned}$$

For $= \sqrt{\frac{1}{H^2SAK + H^4(K+D)}}$ and $= 2H$, we get: $R_K \le O\left(H^2\sqrt{SAK} + H^3\sqrt{D} + H^3\sqrt{K} + H^4 d_{\max}\right)$. □

B.3. Bound on BIAS_1

Lemma B.3. Under the good event $\mathcal{E}$, $\text{BIAS}_1 \le \frac{2}{3}\sum_{k=1}^{K}\sum_{h,s} q_h(s)b_h^k(s) + O\left(H^4(K+D) + \frac{H^4 d_{\max}}{} + \frac{H^2}{}\right)$.

Proof. Let $Y_k = \sum_{h,s} q_h(s)\left(\ell_h^{k+d^k}(\cdot|s) - \hat{Q}_h^k(s,\cdot)\right)$. It holds that

$$\text{BIAS}_1 = \sum_{k=1}^{K}\sum_{h,s} q_h(s)\left(\ell_h^{k+d^k}(\cdot|s) - Q_h^k(s,\cdot)\right) \le \sum_{k=1}^{K} E_k[Y_k] + \sum_{k=1}^{K}\left(E_k[Y_k] - Y_k\right).$$

Under event $\mathcal{E}^f$ it holds that

$$\sum_{k=1}^{K} E_k[Y_k] - \sum_{k=1}^{K} Y_k \le \frac{1}{3}\sum_{k=1}^{K}\sum_{h,s} q_h(s)b_h^k(s) + \frac{H^2}{}\ln\frac{10}{}.$$

In addition,

$$\sum_{k=1}^{K}\sum_{h,s} q_h(s)\left\langle \pi_h^{k+d^k}(\cdot|s), Q_h^k(s,\cdot)\right\rangle \mathbb{E}\left[\sum_{k=1}^{K} Y_k\right] =$$

$$= \sum_{k=1}^{K}\sum_{h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)Q_h^k(s,a)\left(1 - \frac{q_h^k(s,a)r_h^k(s,a)}{q_h^k(s,a)+\gamma}\right)$$

$$= \sum_{k=1}^{K}\sum_{h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)Q_h^k(s,a)\left(1 - \frac{(q_h^k(s,a)+\gamma)r_h^k(s,a)}{q_h^k(s,a)+\gamma}\right)$$

$$+ \sum_{k=1}^{K}\sum_{h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)Q_h^k(s,a)\frac{r_h^k(s,a)}{q_h^k(s,a)+\gamma}$$

$$\sum_{k=1}^{K}\sum_{h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)Q_h^k(s,a)(1 - r_h^k(s,a)) + \sum_{k=1}^{K}\sum_{h,s,a} q_h(s)\frac{H\,\pi_h^{k+d^k}(a|s)r_h^k(s,a)}{q_h^k(s,a)+\gamma}$$

$$H\sum_{k=1}^{K}\sum_{h,s,a} q_h(s)\,\max\{\pi_h^k(a|s), \pi_h^{k+d^k}(a|s)\}\,\left|\pi_h^k(a|s) + \sum_{k=1}^{K}\sum_{h,s,a} q_h(s)\frac{H\,\pi_h^{k+d^k}(a|s)r_h^k(s,a)}{q_h^k(s,a)+\gamma}\right.$$

$$H\sum_{k=1}^{K}\sum_{h,s} q_h(s)\left\|\pi_h^{k+d^k}(\cdot|s) - \pi_h^k(\cdot|s)\right\|_1 + \frac{1}{3}\sum_{k=1}^{K}\sum_{h,s} q_h(s)b_h^k(s).$$

Finally, using Lemma B.8,

$$\begin{aligned}
\text{BIAS}_1 &\le \frac{2}{3}\sum_{k=1}^{K}\sum_{h,s} q_h(s)b_h^k(s) + H\sum_{h,s} q_h(s)\sum_{k=1}^{K}\eta^k\left\|\pi_h^{k+d^k}(\cdot|s) - \pi_h^k(\cdot|s)\right\|_1 + \frac{H^2}{\gamma}\\
&\le \frac{2}{3}\sum_{k=1}^{K}\sum_{h,s} q_h(s)b_h^k(s) + H\sum_{h,s} q_h(s)\;O\left(H^2(K+D) + \frac{H^2 d_{\max}}{\gamma}\right) + \frac{H^2}{\gamma}\\
&= \frac{2}{3}\sum_{k=1}^{K}\sum_{h,s} q_h(s)b_h^k(s) + O\left(H^4(K+D) + \frac{H^4 d_{\max}}{\gamma} + \frac{H^2}{\gamma}\right). \qquad \square
\end{aligned}$$

B.4. Bound on BONUS

Lemma B.4. It holds that $\text{BONUS} \le 3H^2SAK + \sum_{k,h,s} q_h(s)b_h^k(s)$.

Proof. Note that B_h^k is the Q-function of policy π^k with respect to the cost function b^k. Hence, by the value difference difference lemma (Lemma E.1),

$$\sum_{h,s} q_h^k(s)b_h^k(s) - \sum_{h,s} q_h(s)b_h^k(s) = V_1^{\pi^k}(s_{\text{init}}; b^k) - V_1^{\pi^{k+d^k}}(s_{\text{init}}; b^k) = \sum_{h,s} q_h(s)\left\langle \pi_h^k(\cdot|s) - \pi_h^{k+d^k}(\cdot|s), B_h^k(s,\cdot)\right\rangle.$$

Summing over k we get: $\text{BONUS} = \sum_{k,h,s} q_h^k(s)b_h^k(s) - \sum_{k,h,s} q_h(s)b_h^k(s)$. For last,

$$\sum_{k=1}^{K}\sum_{h,s} q_h^k(s)b_h^k(s) = 3H\sum_{k=1}^{K}\sum_{h,s,a}\frac{q_h^k(s)\,\pi_h^{k+d^k}(a|s)r_h^k(s,a)}{q_h^k(s)\,\pi_h^k(a|s)+\gamma} \le 3H^2SAK,$$

where the last uses the fact that $q_h^{k+d^k}(a|s)r_h^k(s,a) \le q_h^k(s)\,\pi_h^k(a|s)$. $\square$

Remark B.5. The adaptation to delay via the ratio $\frac{q_h^k(s,a)}{q_h^k(s,a)+\gamma}$ is simple, yet crucial. The main reason is the following. While in MAB the ratio $\pi_h^{k+d^k}(a|s)/\pi_h^k(a|s)$ is always bounded by a constant (Thune et al., 2019, Lemma 11), in MDPs it can be

as large as $e^{d_{max}}$. In fact, even the ratio $\pi_h^{k+1}(a \mid s) / \pi_h^k(a \mid s)$ can be of order $e^{d_{max}}$, because even if an action a is chosen with probability close to 1, the estimator $\hat{Q}_h^k(s,a)$ can be as large as $1/\gamma$, as long as the visitation probability to state s is smaller than γ. This can cause radical changes in the probability to take action a (and as a consequence in the probability of the rest of the actions in that state). For example, assume we have two actions a_1, a_2 with $\pi_h^k(a_1 \mid s) = 1/(e^{d_{max}}+1)$, $\pi_h^k(a_2 \mid s) = e^{d_{max}}/(e^{d_{max}}+1)$ and $q_h^k(s) \geq \gamma$. Now, assume that the feedback from d_{max} episodes arrive at at the end of episode k for which the agent visited in the state-action pair (s,a_1). Further assume the cost-to-go from (s,a_1) was of order H. This would imply that $\sum_{j : j + d^j = k} (\hat{Q}_h^j(s,a) - B_h^j(s,a)) \geq d_{max}$. Hence, the probability to take each of the actions under π^{k+1} would be approximately $1/2$. In particular, $\pi_h^{k+1}(a_2 \mid s) / \pi_h^k(a_2 \mid s) = ((e^{d_{max}}+1)/e^{d_{max}}) = (1 + e^{-d_{max}})$.

B.5. Bound on REG

Lemma B.6. For $\eta \leq \frac{\gamma}{2H}$ it holds that $\text{REG} \leq \frac{H \ln A}{\eta} + \frac{1}{3}\sum_{k,h,s} q_h^k(s) b_h^k(s) + O\left(H^5 K \eta + \frac{H^3}{\gamma}\right)$.

Proof. By Corollary E.7, since $\max_{k,h,s,a} B_h^k(s,a) \leq 3H^2$,

$$\begin{aligned}
\text{REG} &\leq \frac{H \ln A}{\eta} + 2\sum_{k,h,s,a} q_h^k(s)\, \pi_h^{k+d^k}(a \mid s)\, \left(\hat{Q}_h^k(s,a) - B_h^k(s,a)\right)^2 + O(H^5 K \eta) \\
&\leq \frac{H \ln A}{\eta} + 2\sum_{k,h,s,a} q_h^k(s)\, \pi_h^{k+d^k}(a \mid s) \hat{Q}_h^k(s,a)^2 + 2\sum_{k,h,s,a} q_h^k(s)\, \pi_h^{k+d^k}(a \mid s) B_h^k(s,a)^2 + O(H^5 K \eta). \qquad (17)
\end{aligned}$$

For the middle term

$$\begin{aligned}
2\sum_{k,h,s,a} q_h^k(s)\, \pi_h^{k+d^k}(a \mid s) \hat{Q}_h^k(s,a)^2 &\leq 2\sum_{k,h,s,a} q_h^k(s)\, \pi_h^{k+d^k}(a \mid s) \frac{H^2 r_h^k(s,a)^2 \mathbb{I}\{s_h^k = s, a_h^k = a\}}{(q_h^k(s,a) + \gamma)^2} \\
&\leq 2H^2 \sum_{k,h,s,a} \frac{q_h^k(s)\, \pi_h^{k+d^k}(a \mid s) r_h^k(s,a)}{q_h^k(s,a) + \gamma} + O\left(\frac{H^3}{\gamma}\right) \\
&\leq \frac{2}{3} H \sum_{k,h,s} q_h^k(s) b_h^k(s) + O\left(\frac{H^3}{\gamma}\right) \\
&\leq \frac{1}{3}\sum_{k,h,s} q_h^k(s) b_h^k(s) + O\left(\frac{H^3}{\gamma}\right),
\end{aligned}$$

where the second inequality if by $b \geq \gamma$ and the last is since $\eta \leq \frac{\gamma}{2H}$. For the last term in Eq. (17) we use $b_h^k(s) \leq 3H$ and therefore $B_h^k(s,a) \leq 3H^2$. Thus: $\sum_{k,h,s,a} q_h^k(s)\, \pi_h^{k+d^k}(a \mid s) B_h^k(s,a)^2 \leq 9 H^5 K$. Overall,

$$\text{REG} \leq \frac{H \ln A}{\eta} + \frac{1}{3}\sum_{k,h,s} q_h^k(s) b_h^k(s) + O\left(H^5 K \eta + \frac{H^3}{\gamma}\right). \qquad \square$$

B.6. Bound on DRIFT

Lemma B.7. If event E^d holds then $\text{DRIFT} \leq O\left(H^5 (K + D) \eta + \frac{H^5 d_{max}}{\gamma}\right)$.

Proof. We use Lemma B.8 and the fact that $|Q_h^k(s,a) - B_h^k(s,a)| \le 3H^2$ to obtain

$$\begin{aligned}
\textsc{Drift} &= \sum_{k=1}^{K} \sum_{h,s,a} q_h(s)\left(\pi_h^k(a \mid s) - \pi_h^{k+d^k}(a \mid s)\right)\left(Q_h^k(s,a) - B_h^k(s,a)\right) \\
&\le \sum_{k=1}^{K} \sum_{h,s,a} q_h(s)\left|\pi_h^k(a \mid s) - \pi_h^{k+d^k}(a \mid s)\right|\left|Q_h^k(s,a) - B_h^k(s,a)\right| \\
&\le 3H^2 \sum_{k=1}^{K} \sum_{h,s} q_h(s)\left\|\pi_h^k(\cdot \mid s) - \pi_h^{k+d^k}(\cdot \mid s)\right\|_1 \\
&\le O\left(H^5(K+D) + \frac{H^5 d_{\max}}{}\right).
\end{aligned}$$

□

Lemma B.8. If event E^d holds then, $\sum_{k=1}^{K} \left\|\pi_h^{k+d^k}(\cdot \mid s) - \pi_h^k(\cdot \mid s)\right\|_1 \le O\left(H^2(K+D) + \frac{H^2 d_{\max}}{}\right)$.

Proof. We first bound,

$$\begin{aligned}
&\sum_{k=1}^{K} \left\|\pi_h^{k+d^k}(\cdot \mid s) - \pi_h^k(\cdot \mid s)\right\|_1 \le \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \left\|\pi_h^{j+1}(\cdot \mid s) - \pi_h^j(\cdot \mid s)\right\|_1 \\
&= \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \sum_{a} \left|\pi_h^{j+1}(a \mid s) - \pi_h^j(a \mid s)\right|.
\end{aligned}$$

Now, we apply Lemma E.9 for each j in the summation above with $\ell(\cdot) = \pi_h^{j+1}(\cdot \mid s)$, $\pi(\cdot) = \pi_h^j(\cdot \mid s)$ and $\grave{\ell}(\cdot) = \sum_{i:i+d^i=j} (\hat{Q}_h^i(s,\cdot) - B_h^i(s,\cdot))$ $\sum_{i:i+d^i=j} 6H^2$ and observe that,

$$\begin{aligned}
&\sum_{k=1}^{K} \left\|\pi_h^{k+d^k}(\cdot \mid s) - \pi_h^k(\cdot \mid s)\right\|_1 \le \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \sum_{a} \pi_h^j(a \mid s) \sum_{i:i+d^i=j} (\hat{Q}_h^i(s,a) + 6H^2) \\
&\quad + \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \Big|\underbrace{\sum_{a}}_{=1}\Big| \sum_{a'} \pi_h^j(a' \mid s) \sum_{i:i+d^i=j} (\hat{Q}_h^i(s,a') + 6H^2) \\
&= 2 \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \sum_{a} \pi_h^j(a \mid s) \sum_{i:i+d^i=j} (\hat{Q}_h^i(a \mid s) + 6H^2) \\
&= 2 \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \sum_{i:i+d^i=j} \sum_{a} \pi_h^j(a \mid s)\hat{Q}_h^i(a \mid s) + 12H^2 \sum_{k=1}^{K} \sum_{i=1}^{K} \mathbb{I}\{k \le i + d^i < k + d^k\} \\
&\le 2 \sum_{k=1}^{K} \sum_{j=k}^{k+d^k-1} \sum_{i:i+d^i=j} \sum_{a} \pi_h^j(a \mid s)\hat{Q}_h^i(a \mid s) + 12H^2(D+K);
\end{aligned}$$

where the last inequality is by Lemma E.10. For the first term we use event E^d

$$\sum_{k=1}^{K}\sum_{j=k}^{k+d^k-1}\sum_{i:i+d^i=j}\sum_{a} {}^{j}_{h}(a\,j\,s)\hat{Q}^{i}_{h}(a\,j\,s) = \sum_{k=1}^{K}\sum_{j=k}^{k+d^k-1}\sum_{i:i+d^i=j}\sum_{a} {}^{i+d^i}_{h}(a\,j\,s)\hat{Q}^{i}_{h}(a\,j\,s)$$

$$= \sum_{k=1}^{K}\sum_{i=1}^{K}\sum_{a} \mathbb{I}\{k \;\; i+d^i < k+d^k\} \; {}^{i+d^i}_{h}(a\,j\,s)\hat{Q}^{i}_{h}(a\,j\,s)$$

$$\sum_{k=1}^{K}\sum_{i=1}^{K}\sum_{a} \mathbb{I}\{k \;\; i+d^i < k+d^k\} \; {}^{i+d^i}_{h}(a\,j\,s)Q^{i}_{h}(a\,j\,s) + O\left(\frac{H^2 d_{max}}{}\right)$$

$$H\sum_{k=1}^{K}\sum_{i=1}^{K} \mathbb{I}\{k \;\; i+d^i < k+d^k\} + O\left(\frac{H^2 d_{max}}{}\right)$$

$$H(D+K) + O\left(\frac{H^2 d_{max}}{}\right) : \qquad \square$$

C. Delay-Adapted Policy Optimization for (Tabular) Adversarial MDP with Unknown Transition

Algorithm 4 Delay-Adapted Policy Optimization with Unknown Transition Function (Tabular)

Input: state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, learning rate $\eta > 0$, exploration parameter $\gamma > 0$, confidence parameter $\delta > 0$.

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

Compute visit counters for every $(h,s,a,s') \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:

$$n_h^k(s,a,s') = \sum_{j: j + d^j < k} \mathbb{I}\{s_h^j = s, a_h^j = a, s_{h+1}^j = s'\}, \qquad n_h^k(s,a) = \sum_{j: j + d^j < k} \mathbb{I}\{s_h^j = s, a_h^j = a\}.$$

Compute empirical transition function $p_h^k(s' \mid s,a) = \frac{n_h^k(s,a,s')}{\max\{n_h^k(s,a), 1\}}$ and confidence set $\mathcal{P}^k = \{P_h^k(s,a)\}_{s,a,h}$ such that $p_h^0(\cdot \mid s,a) \in P_h^k(s,a)$ if and only if $\sum_{s'} p_h^0(s' \mid s,a) = 1$ and for every $s' \in \mathcal{S}$:

$$\left| p_h^0(s' \mid s,a) - p_h^k(s' \mid s,a) \right| \le 4 \sqrt{\frac{p_h^k(s' \mid s,a) \log\frac{10HSAK}{\delta}}{n_h^k(s,a) - 1}} + \frac{10 \log\frac{10HSAK}{\delta}}{n_h^k(s,a) - 1}.$$

Compute occupancy measures $\overline{q}_h^k(s) = \max_{p' \in \mathcal{P}^k} q_h^{\pi^k, p'}(s)$ and $\underline{q}_h^k(s) = \min_{p' \in \mathcal{P}^k} q_h^{\pi^k, p'}(s)$.

Policy Evaluation

for j such that $j + d^j = k$ **do**

Observe bandit feedback $\{c_h^j(s_h^j, a_h^j)\}_{h=1}^H$ and set $B_{H+1}^j(s,a) = 0$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$.

for $h = H, H-1, \ldots, 1$ **do**

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

Compute $r_h^j(s,a) = \frac{\pi_h^j(a \mid s)}{\max\{\pi_h^j(a \mid s), \gamma_h^k(a \mid s)\}}$ and $L_h^j = \sum_{h'=h}^H c_{h'}^j(s_{h'}^j, a_{h'}^j)$.

Compute $b_h^j(s) = \sum_{a \in \mathcal{A}} \frac{3H \gamma_h^k(a \mid s) r_h^j(s,a)}{\overline{q}_h^j(s) \pi_h^j(a \mid s) + \gamma}$ and $b_h^j(s) = \sum_{a \in \mathcal{A}} \frac{2H \gamma_h^k(a \mid s) r_h^j(s,a) (\overline{q}_h^j(s,a) - \underline{q}_h^j(s,a))}{\overline{q}_h^j(s) \pi_h^j(a \mid s) + \gamma}$.

Compute $b_h^j(s) = b_h^j(s) + b_h^j(s)$ and $\widehat{Q}_h^j(s,a) = r_h^j(s,a) \frac{\mathbb{I}\{s_h^j = s, a_h^j = a\} L_h^j}{\overline{q}_h^j(s) \pi_h^j(a \mid s) + \gamma}$.

Compute $B_h^j(s,a) = b_h^j(s) + \max_{p' \in \mathcal{P}_h^j(s,a)} \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} p_h^0(s' \mid s,a) \pi_{h+1}^j(a' \mid s') B_{h+1}^j(s',a')$.

end for

end for

end for

Policy Improvement

Define the policy π^{k+1} for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ by:

$$\pi_h^{k+1}(a \mid s) = \frac{\gamma_h^k(a \mid s) \exp\left(\sum_{j: j + d^j = k} (\widehat{Q}_h^j(s,a) - B_h^j(s,a))\right)}{\sum_{a' \in \mathcal{A}} \gamma_h^k(a' \mid s) \exp\left(\sum_{j: j + d^j = k} (\widehat{Q}_h^j(s,a') - B_h^j(s,a'))\right)}.$$

end for

Theorem C.1. Set $\gamma = H - H^2 SAK + H^4 (K + D)^{1/2}$ and $\eta = 2\gamma H$. Running Algorithm 4 in an adversarial MDP $M = (\mathcal{S}, \mathcal{A}, H, p, \{c^k\}_{k=1}^K)$ with unknown transition function and delays $\{d^k\}_{k=1}^K$ guarantees, with probability at least $1 - \delta$,

$$R_K \le O\left(H^3 S \sqrt{AK} \log\frac{HSAK}{\delta} + H^3 \sqrt{K + D} \log\frac{HSAK}{\delta} + H^4 S^2 A d_{\max} + H^4 S^3 A \log^2\frac{HSAK}{\delta} \right).$$

Remark C.2 (Dependence in $d_{\max}$). All our regret bounds contain additive terms that scale linearly with $d_{\max}$. While these are low-order terms when $d_{\max}$ is smaller than $\sqrt{D}$, they may become dominant for large maximal delay. The dependence in $d_{\max}$ can be removed altogether using the *skipping* technique (Thune et al., 2019; Bistritz et al., 2019; Lancewicki

et al., 2022b), i.e., ignoring episodes that their delay is larger than some threshold. In the case of known transitions (Theorem B.1 in Appendix B), we can set $\sigma = \sqrt{D}/H$ and remove the dependence on d_{max} without hurting our original regret bound, i.e., we get the bound $R_K = O(H^2\sqrt{SAK} + H^3\sqrt{K + D})$. However, in the case of unknown transitions (Theorem C.1 in Appendix C), we get a slightly worse regret bound. Specifically, we can set $\sigma = \sqrt{\frac{D}{H^2S^2A}}$ and obtain the regret $R_K = O(H^3S\sqrt{A(K + D)})$. Jin et al. (2022) encounter the same issue in their regret bounds, so it remains an open problem whether the dependence on d_{max} can be removed without hurting the original regret bound in the unknown transitions case.

C.1. The good event

Let $\iota = 10 \log \frac{10KHSA}{\delta}$, H^k be the history of episodes $\{j : j + d^j < k\}$, and H^k be the history of episodes $\{j : j < k\}$. Define the following events:

$$E^p = \left\{ \forall (k, s', s, a, h) : p_h(s' \mid s, a) \quad p_h^k(s' \mid s, a) \quad 4\sqrt{\frac{p_h^k(s' \mid s, a)\log\frac{10HSAK}{\delta}}{\max\{n_h^k(s,a), 1\}}} + 10\frac{\log\frac{10HSAK}{\delta}}{\max\{n_h^k(s,a), 1\}} \right\},$$

$$E^{est} = \left\{ \sum_{h,s,a,k} |q_h^k(s,a) \quad \bar{q}_h^k(s,a)| \quad O\left(\sqrt{H^4S^2AK\log\frac{10KHSA}{\delta}} + H^3S^3A\log^2\frac{10KHSA}{\delta} + H^3S^2Ad_{max}\right) \right\},$$

$$E^d = \left\{ \forall (h,s) : \sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{a} \mathbb{I}\{f \quad k + d^k < j + d^j g\}\, \pi_h^{k+d^k}(a \mid s)(\hat{Q}_h^k(s,a) \quad Q_h^k(s,a)) \quad \frac{10H^2 d_{max}\log\frac{10HS}{\delta}}{} \right\},$$

$$E = \left\{ \sum_{k=1}^{K}\sum_{h,s,a} q_h(s,a)(\hat{Q}_h^k(s,a) \quad Q_h^k(s,a)) \quad \frac{H^2\log\frac{10HSA}{\delta}}{} \right\},$$

$$E^b = \left\{ \sum_{k=1}^{K}\sum_{h,s,a} \frac{q_h(s)\,\pi_h^{k+d^k}(a \mid s)r_h^k(s,a)\mathbb{I}\{f s_h^k = s, a_h^k = ag}{(\bar{q}_h^k(s,a) +)^2} \sum_{k=1}^{K}\sum_{h,s,a} \frac{q_h(s)\,\pi_h^{k+d^k}(a \mid s)r_h^k(s,a)}{\bar{q}_h^k(s,a) +} \quad \frac{H}{2}\ln\frac{10H}{\delta} \right\},$$

$$E^f = \left\{ \sum_{k=1}^{K} E_k\left[\sum_{h,s} q_h(s)\left(\pi_h^{k+d^k}(\cdot \mid s), \hat{Q}_h^k(s,\cdot)\right)\right] \sum_{k=1}^{K}\sum_{h,s} q_h(s)\left(\pi_h^{k+d^k}(\cdot \mid s), \hat{Q}_h^k(s,\cdot)\right) \right.$$
$$\left. \frac{1}{3}\sum_{k=1}^{K}\sum_{h,s} q_h(s)\bar{c}_h^k(s) + \frac{H^2}{}\ln\frac{10}{\delta} \right\},$$

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

Lemma C.3 (The Good Event) Let $G = E^p \cap E^{est} \cap E^d \cap E \cap E^b \cap E^f$ be the good event. It holds that $\Pr[G] \quad 1$. Moreover, under the good event, it holds that $p \in P^k$ and $\bar{q}_h^k(s,a) \quad q_h^k(s,a) \quad \bar{q}_h^k(s,a)$ for every $(k,h,s,a) \in [K] \quad [H] \quad S \quad A$.

Proof. We'll show that each of the events E^p, E^d, E, E^b, E^f holds with probability of at most $=5$ and so by the union bound $\Pr[G] \quad 1$.

Event E^p: Holds with probability $1 \quad =10$ by standard Bernstein inequality (see, e.g., Lemma 2 in Jin et al. (2020a)). As a consequence of event E^p, $p \in P^k$ for all k. In particular, $\bar{q}_h^k(s,a) \quad q_h^k(s,a)$ for all k, h, s and a.

Event E^{est}: Holds with probability $1 \quad =10$ by Jin et al. (2022, Lemma D.12) (see also (Jin et al., 2020a, Lemma 4)) which is a standard techniques (adapted to delays) of summing the the confidence radius on the trajectory.

Event E^d: We show the proof under event E^p which occurs with probability $1 \quad =10$. Fix s and h. Similar to the proof of

Lemma B.2, we apply Lemma E.5. For every (h^0, s^0, a), set $c^k_{h^0}(s^0, a) = \bar{q}^k_{h^0}(s^0, a)$ and

$$z^k_{h^0}(s^0, a) = \mathbb{I}\{s^0 = s, h^0 = h\} \sum_{j=1}^{K} \mathbb{I}\{j \quad k + d^k < j \quad + d^j\} \, \pi^{k+d^k}_h(a \mid s) r^k_h(s, a) Q^j_h(s, a)$$

$$Z^k_{h^0}(s^0, a) = \mathbb{I}\{s^0 = s, h^0 = h\} \sum_{j=1}^{K} \mathbb{I}\{j \quad k + d^k < j \quad + d^j\} \, \pi^{k+d^k}_h(a \mid s) r^k_h(s, a) M^k_h(s, a)$$

$$M^k_h(s, a) = \mathbb{I}\{s^k_h = s, a^k_h = a\} \sum_{h'=1}^{H} c^k_{h'}(s^k_{h'}, a^k_{h'}) + (1 \quad \mathbb{I}\{s^k_h = s, a^k_h = a\}) Q^k_h(s, a).$$

Note that $\mathbb{E}_k[Z^k_{h^0}(s^0, a)] = z^k_{h^0}(s^0, a)$ and

$$\sum_{h^0, s^0, a} \frac{\mathbb{I}\{s^k_h = s, a^k_h = a\} Z^k_h(s, a)}{\bar{q}^k_h(s, a) + } = \sum_{j=1}^{K} \sum_{a} \mathbb{I}\{j \quad k + d^k < j \quad + d^j\} \, \pi^{k+d^k}_h(a \mid s) \hat{Q}^k_h(s, a)$$

$$\sum_{h^0, s^0, a} \frac{q^k_h(s, a) z^k_h(s, a)}{\bar{q}^k_h(s, a)} \sum_{j=1}^{K} \sum_{a} \mathbb{I}\{j \quad k + d^k < j \quad + d^j\} \, \pi^{k+d^k}_h(a \mid s) Q^k_h(s, a);$$

where in the inequality we used the fact that $\bar{q}^k_h(s, a) \quad q^k_h(s, a)$ under the event E^p. Finally, we use Lemma E.5 with $R \quad 2Hd_{\max}$ as in the proof of Lemma B.2. Thus, the event holds for (h, s, a) with probability $1 \quad \frac{}{10HS}$. By taking the union bound over all h and s, E^d holds with probability $1 \quad \frac{}{10}$.

Event E (Lemma C.2 of Luo et al. (2021)): We show the proof under event E^p which occurs with probability $1 \quad =10$. Again, E holds with probability of at least $1 \quad =10$ by applying Lemma E.5 with $c^k_{h^0}(s^0, a) = \bar{q}^k_{h^0}(s^0, a)$, $z^k_h(s, a) = q^k_h(s, a) r^k_h(s, a) Q^k_h(s, a)$ and,

$$Z^k_h(s, a) = q^k_h(s, a) r^k_h(s, a) \quad \mathbb{I}\{s^k_h = s, a^k_h = a\} \sum_{h^0=1}^{H} c^k_{h^0}(s^k_{h^0}, a^k_{h^0}) + (1 \quad \mathbb{I}\{s^k_h = s, a^k_h = a\}) Q^k_h(s, a) \quad .$$

Note that $R \quad H$, $\frac{\mathbb{I}\{s^k_h = s, a^k_h = a\} Z^k_h(s, a)}{\bar{q}^k_h(s, a) + } = q^k_h(s, a) \hat{Q}^k_h(s, a)$ and $\frac{q^k_h(s, a) z^k_h(s, a)}{\bar{q}^k_h(s, a)} \quad q^k_h(s, a) Q^k_h(s, a)$ where similar to before, the inequality holds under event E^p.

Event E^b: We show the proof under event E^p which occurs with probability $1 \quad =10$. Similar to before, E^b holds with probability of at least $1 \quad =10$ by applying Lemma E.5 with $c^k_{h^0}(s^0, a) = \bar{q}^k_{h^0}(s^0, a)$ and $z^k_h(s, a) = Z^k_h(s, a) = \frac{q_h(s) \, \pi^{k+d^k}_h(a|s) r^k_h(s, a)}{\bar{q}^k_h(s, a) + }$.

Event E^f: We show the proof under event E^p which occurs with probability $1 \quad =10$. Let $Y_k = \sum_{h, s} q_h(s) \, \pi^{k+d^k}_h(\mid s) \hat{Q}^k_h(s,)$. Similar to the proof of Lemma B.2, we use a variant of Freedman's inequality (Lemma E.3) to bound $\sum_{k=1}^{K} \mathbb{E}_k[Y_k] \quad \sum_{k=1}^{K} Y_k$.

$$\mathbb{E}_k\left[Y_k^2\right] = \mathbb{E}_k\left[\left(\sum_{h,s,a} q_h(s) \, \pi^{k+d^k}_h(a \mid s) \hat{Q}^k_h(s, a)\right)^2\right]$$

$$\mathbb{E}_k\left[\left(\sum_{h,s,a} q_h(s) \, \pi^{k+d^k}_h(a \mid s)\right)\left(\sum_{h,s,a} q_h(s) \, \pi^{k+d^k}_h(a \mid s) (\hat{Q}^k_h(s, a))^2\right)\right]$$

$$H \, \mathbb{E}_k\left[\sum_{h,s,a} q_h(s) \, \pi^{k+d^k}_h(a \mid s) \frac{r^k_h(s, a)^2 (L^k_h)^2 \mathbb{I}\{s^k_h = s, a^k_h = a\}}{(\bar{q}^k_h(s, a) +)^2}\right]$$

$$H^3 \sum_{h,s,a} q_h(s) \, \pi^{k+d^k}_h(a \mid s) \frac{r^k_h(s, a) q^k_h(s, a)}{(\bar{q}^k_h(s, a) +)^2} \quad H^3 \sum_{h,s,a} \frac{q_h(s) \, \pi^{k+d^k}_h(a \mid s) r^k_h(s, a)}{\bar{q}^k_h(s, a) + };$$

where the first inequality is Cauchy-Schwartz inequality and the last inequality holds under event E. Also, $|Y_k| \le \frac{H^2}{\eta}$. Therefore by Lemma E.3 with probability $1 - \delta = 10$,

$$\sum_{k=1}^{K} E_k[Y_k] - \sum_{k=1}^{K} Y_k \le H \sum_{k=1}^{K} \left(\sum_{h,s,a} \frac{q_h(s)\pi_h^{k+d^k}(a|s) r_h^k(s,a)}{\bar{q}_h^k(s,a) + \gamma} \eta + \frac{H^2}{\eta} \ln \frac{10}{\delta} \right) = \frac{1}{3} \sum_{k=1}^{K} \sum_{h,s} q_h(s) b_h^k(s) + \frac{H^2}{\eta} \ln \frac{10}{\delta}.$$

C.2. Proof of the main theorem

Proof of Theorem C.1. By Lemma C.3, the good event holds with probability of at least $1 - \delta$. As in the previous section, we analyze the regret under the assumption that the good event holds. We start with the following regret decomposition,

$$\begin{aligned}
R_K = \sum_{k=1}^{K} \sum_{h,s} q_h(s) \left(\pi_h^k(\cdot|s) - \pi_h(\cdot|s) \right) Q_h^k(s,\cdot) &= \underbrace{\sum_{k=1}^{K} \sum_{h,s} q_h(s) \left(\pi_h^{k+d^k}(\cdot|s) - \pi_h(\cdot|s) \right) Q_h^k(s,\cdot) - \hat{Q}_h^k(s,\cdot)}_{\text{BIAS}_1} \\
&+ \underbrace{\sum_{k=1}^{K} \sum_{h,s} q_h(s) \left(\pi_h(\cdot|s) - \pi_h(\cdot|s) \right) \hat{Q}_h^k(s,\cdot) - Q_h^k(s,\cdot)}_{\text{BIAS}_2} + \underbrace{\sum_{k=1}^{K} \sum_{h,s} q_h(s) \left(\pi_h^k(\cdot|s) - \pi_h(\cdot|s) \right) B_h^k(s,\cdot)}_{\text{BONUS}} \\
&+ \underbrace{\sum_{k=1}^{K} \sum_{h,s} q_h(s) \left(\pi_h^{k+d^k}(\cdot|s) - \pi_h(\cdot|s) \right) \hat{Q}_h^k(s,\cdot) - B_h^k(s,\cdot)}_{\text{REG}} \\
&+ \underbrace{\sum_{k=1}^{K} \sum_{h,s} q_h(s) \left(\pi_h^k(\cdot|s) - \pi_h^{k+d^k}(\cdot|s) \right) Q_h^k(s,\cdot) - B_h^k(s,\cdot)}_{\text{DRIFT}};
\end{aligned}$$

where the first is by Lemma E.1. BIAS_2 is bounded under event E by $O\left(\frac{H^2}{\eta}\right)$, and DRIFT $O\left(H^5(K+D) + H^5 d_{\max}\right)$ by Lemma B.7. The other three term are bounded in Lemmas C.4 to C.6. Overall,

$$\begin{aligned}
R_K &\le \underbrace{\frac{2}{3} \sum_{k=1}^{K} \sum_{h,s} q_h(s) b_h^k(s) + \sum_{k=1}^{K} \sum_{h,s} q_h(s) b_h^k(s)}_{\text{BAIS}_1} + O\left(H^4(K+D) + \frac{H^4 d_{\max}}{\eta} + \frac{H^2}{\gamma}\right) \\
&+ O\left(\frac{H^2}{\gamma}\right) + O\left(H^2SAK + H^3S\sqrt{AK} + H^4S^3A^2 + H^4S^2Ad_{\max}\right) \underbrace{\sum_{k=1}^{K} \sum_{h,s} q_h(s) b_h^k(s)}_{\text{BONUS}} \\
&+ \underbrace{\frac{H \ln A}{\eta} + \frac{1}{3} \sum_{k,h,s} q_h(s) b_h^k(s)}_{\text{REG}} + O\left(H^5K + \frac{H^3}{\eta}\right) \\
&+ O\left(H^5(K+D) + \frac{H^5 d_{\max}}{\eta}\right)_{\text{DRIFT}} \\
&O\left(\frac{H \ln A}{\eta} + H^2SAK + H^5(K+D) + H^3S\sqrt{AK}\right. \\
&\left. + \frac{H^3}{\eta} + \frac{H^2}{\gamma} + \frac{H^5 d_{\max}}{\eta} + H^4S^2Ad_{\max} + H^4S^3A^2\right).
\end{aligned}$$

For $= \sqrt{\frac{1}{H^2SAK + H^4(K+D)}}$ and $= 2H$ we get:

$$R_K \quad O\left(H^3S\sqrt{AK} + H^3\sqrt{K+D} + H^4S^2Ad_{max} + H^4S^3A^2\right): \qquad \square$$

C.3. Bound on $BIAS_1$

Lemma C.4. Under the good event,

$$BIAS_1 \quad \frac{2}{3}\sum_{k=1}^{K}\sum_{h;s} q_h(s)\bar{b}_h^k(s) + \sum_{k=1}^{K}\sum_{h;s} q_h(s)b_h^k(s) + O\left(H^4(K+D) + \frac{H^4 d_{max}}{} + \frac{H^2}{}\right):$$

Proof. Let $Y_k = \sum_{h;s} q_h(s)\left\langle P_h^{k+d^k}(\cdot\,|\,s), \hat{Q}_h^k(s,\cdot)\right\rangle$. It holds that

$$BIAS_1 = \sum_{k=1}^{K}\sum_{h;s} q_h(s)\left\langle P_h^{k+d^k}(\cdot\,|\,s), Q_h^k(s,\cdot)\right\rangle - \sum_{k=1}^{K} E_k[Y_k] + \sum_{k=1}^{K} E_k[Y_k] - \sum_{k=1}^{K} Y_k:$$

Under event E^f it holds that $\sum_{k=1}^{K} E_k[Y_k] \quad \sum_{k=1}^{K} Y_k \quad \frac{1}{3}\sum_{k=1}^{K}\sum_{h;s} q_h(s)\bar{b}_h^k(s) + \frac{H^2}{}\ln\frac{10}{}$. In addition,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h;s} q_h(s)\left\langle P_h^{k+d^k}(\cdot\,|\,s), Q_h^k(s,\cdot)\right\rangle - \sum_{k=1}^{K} E_k[Y_k] = \\
&= \sum_{k=1}^{K}\sum_{h;s;a} q_h(s)\,P_h^{k+d^k}(a\,|\,s)Q_h^k(s,a)\left(1 - \frac{q_h^k(s,a)r_h^k(s,a)}{\bar{q}_h^k(s,a)+}\right) \\
&= \sum_{k=1}^{K}\sum_{h;s;a} q_h(s)\,P_h^{k+d^k}(a\,|\,s)Q_h^k(s,a)\left(1 - \frac{(q_h^k(s,a)+\;\;+\bar{q}_h^k(s,a) - \underline{q}_h^k(s,a))r_h^k(s,a)}{\bar{q}_h^k(s,a)+}\right)^{!} \\
&\quad + \sum_{k=1}^{K}\sum_{h;s;a} q_h(s)\,P_h^{k+d^k}(a\,|\,s)Q_h^k(s,a)\frac{r_h^k(s,a)}{\bar{q}_h^k(s,a)+} \\
&\quad + \sum_{k=1}^{K}\sum_{h;s;a} q_h(s)\,P_h^{k+d^k}(a\,|\,s)Q_h^k(s,a)\frac{r_h^k(s,a)(\bar{q}_h^k(s,a) - \underline{q}_h^k(s,a))}{\bar{q}_h^k(s,a)+} \\
&= \sum_{k=1}^{K}\sum_{h;s;a} q_h(s)\,P_h^{k+d^k}(a\,|\,s)Q_h^k(s,a)\underbrace{\frac{(\bar{q}_h^k(s,a)+\;\;)(1 - r_h^k(s,a))}{\bar{q}_h^k(s,a)+}}_{(i)} \\
&\quad + \sum_{k=1}^{K}\sum_{h;s;a} q_h(s)\,P_h^{k+d^k}(a\,|\,s)Q_h^k(s,a)\underbrace{\frac{r_h^k(s,a)(q_h^k(s,a) - \underline{q}_h^k(s))}{\bar{q}_h^k(s,a)+}}_{(ii)}{}^{!} \\
&\quad + \sum_{k=1}^{K}\sum_{h;s;a} q_h(s)\,P_h^{k+d^k}(a\,|\,s)Q_h^k(s,a)\underbrace{\frac{r_h^k(s,a)}{\bar{q}_h^k(s,a)+}}_{(iii)} \\
&\quad + \sum_{k=1}^{K}\sum_{h;s;a} q_h(s)\,P_h^{k+d^k}(a\,|\,s)Q_h^k(s,a)\underbrace{\frac{r_h^k(s,a)(\bar{q}_h^k(s,a) - \underline{q}_h^k(s,a))}{\bar{q}_h^k(s,a)+}}_{(iv)}:
\end{aligned}$$

Terms (i) and (iii) are bounded similarly to the known dynamics case:

$$\text{(i)} \quad H \sum_{k=1}^{K} \sum_{h,s,a} q_h(s) \max\{\pi_h^{k+d^k}(a|s), \pi_h^k(a|s)\} \left|\pi_h^k(a|s) - \pi_h^{k+d^k}(a|s)\right| \le H \sum_{k=1}^{K} \sum_{h,s} q_h(s) \left\|\pi_h^{k+d^k}(\cdot|s) - \pi_h^k(\cdot|s)\right\|_1$$

$$\text{(iii)} \quad \sum_{k=1}^{K} \sum_{h,s,a} q_h(s) \frac{\pi_h^{k+d^k}(a|s) r_h^k(s,a)}{\bar{q}_h^k(s,a) + \gamma} H = \frac{1}{3} \sum_{k=1}^{K} \sum_{h,s} q_h(s) \tilde{b}_h^k(s).$$

Finally,

$$\text{(ii)} + \text{(iv)} \quad \sum_{k=1}^{K} \sum_{h,s,a} q_h(s) \frac{2H \pi_h^{k+d^k}(a|s) r_h^k(s,a)(\bar{q}_h^k(s,a) - q_h^k(s,a))}{\bar{q}_h^k(s,a) + \gamma} = \sum_{k=1}^{K} \sum_{h,s} q_h(s) b_h^k(s).$$

In total, using Lemma B.8,

$$\begin{aligned} \text{BIAS}_1 &\le \frac{2}{3} \sum_{k=1}^{K} \sum_{h,s} q_h(s) \tilde{b}_h^k(s) + \sum_{k=1}^{K} \sum_{h,s} q_h(s) b_h^k(s) + H \sum_{h,s} \sum_{k=1}^{K} q_h(s) \left\|\pi_h^{k+d^k}(\cdot|s) - \pi_h^k(\cdot|s)\right\|_1 + \frac{H^2}{\gamma} \\ &\le \frac{2}{3} \sum_{k=1}^{K} \sum_{h,s} q_h(s) \tilde{b}_h^k(s) + \sum_{k=1}^{K} \sum_{h,s} q_h(s) b_h^k(s) + O\left(H^4(K+D) + \frac{H^4 d_{\max}}{\gamma} + \frac{H^2}{\gamma}\right). \end{aligned}$$

□

C.4. Bound on BONUS

Lemma C.5. It holds that

$$\text{BONUS} \le O\left(H^2SAK + H^3S\sqrt{AK} + H^4S^3A^2 + H^4S^2Ad_{\max}\right) \sum_{k=1}^{K} \sum_{h,s} q_h(s) b_h^k(s).$$

Proof. Let p^k be the transition function chosen by the algorithm when calculating B_h^k. It holds that

$$\begin{aligned} &\sum_{h,s} q_h(s) \left\langle \pi_h^k(\cdot|s) - \pi_h(\cdot|s), B_h^k(s,\cdot) \right\rangle = \\ &= \sum_{h,s,a} q_h(s) \pi_h^k(a|s) B_h^k(s,a) - \sum_{h,s,a} q_h(s) \pi_h(a|s) B_h^k(s,a) \\ &= \sum_{h,s,a} q_h(s) \pi_h^k(a|s) B_h^k(s,a) - \sum_{h,s,a} q_h(s) \pi_h(a|s) \left(b_h^k(s) + \sum_{s',a'} p_h^k(s'|s,a) \pi_{h+1}^k(a'|s') B_{h+1}^k(s',a')\right) \\ &\le \sum_{h,s,a} q_h(s) \pi_h^k(a|s) B_h^k(s,a) - \sum_{h,s,a} q_h(s) \pi_h(a|s) \left(b_h^k(s) + \sum_{s',a'} p_h(s'|s,a) \pi_{h+1}^k(a'|s') B_{h+1}^k(s',a')\right) \\ &= \sum_{h,s,a} q_h(s) \pi_h^k(a|s) B_h^k(s,a) - \sum_{h,s} q_h(s) b_h^k(s) - \sum_{h,s,a} q_{h+1}(s) \pi_{h+1}^k(a|s) B_{h+1}^k(s,a) \\ &= \sum_{s,a} q_1(s) \pi_1^k(a|s) B_1^k(s,a) - \sum_{h,s} q_h(s) b_h^k(s) \\ &= \sum_{s,a} q_1^k(s) \pi_1^k(a|s) B_1^k(s,a) - \sum_{h,s} q_h(s) b_h^k(s) \\ &= \sum_{s,a} q_1^{k,p^k}(s) \pi_1^k(a|s) B_1^k(s,a) - \sum_{h,s} q_h(s) b_h^k(s) \\ &= \sum_{h,s} q_h^{k,p^k}(s) b_h^k(s) - \sum_{h,s} q_h(s) b_h^k(s) \\ &\le \sum_{h,s} \bar{q}_h^k(s) b_h^k(s) - \sum_{h,s} q_h(s) b_h^k(s), \end{aligned}$$

where the first two inequalities are by definition of $\bar{q}_h^k$ and the fact that $p \in P^k$ under event E^p; and the last inequality is since, by definition, $\bar{q}_h^k(s)$ maximize the probability to visit s at time h among all the occupancy measures within P^k. Summing over k and bounding the first term above as follows completes the proof:

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h,s} \bar{q}_h^k(s) b_h^k(s) &= \sum_{k=1}^{K}\sum_{h,s,a} \frac{3H\,\bar{q}_h^k(s)\,\pi_h^{k+d^k}(a|s) r_h^k(s,a)}{\bar{q}_h^k(s,a)+\gamma} \\
&\quad + \sum_{k=1}^{K}\sum_{h,s,a} \frac{2H\,\bar{q}_h^k(s)\,\pi_h^{k+d^k}(a|s) r_h^k(s,a)(\bar{q}_h^k(s,a) - \underline{q}_h^k(s,a))}{\bar{q}_h^k(s,a)+\gamma} \\
&\le 3H^2SAK + 2H\sum_{k=1}^{K}\sum_{h,s,a} (\bar{q}_h^k(s,a) - \underline{q}_h^k(s,a)) \\
&\le O\left(H^2SAK + H^3S\sqrt{AK} + H^4S^3A^2 + H^4S^2Ad_{\max}\right),
\end{aligned}$$

where the last is by event E^{est}. □

C.5. Bound on REG

Lemma C.6. For $\eta \le \frac{1}{2H}$ it holds that $\text{REG} \le \frac{H\ln A}{\eta} + \frac{1}{3}\sum_{k,h,s} q_h(s)\bar{b}_h^k(s) + O\left(H^5K + \frac{H^3}{\eta^2}\right)$.

Proof. By Corollary E.7, since $\max_{k,h,s,a} B_h^k(s,a) \le 5H^2$,

$$\begin{aligned}
\text{REG} &\le \frac{H\ln A}{\eta} + 2\sum_{k,h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)\left(\hat{Q}_h^k(s,a) - B_h^k(s,a)\right)^2 + O(H^5K) \\
&\le \frac{H\ln A}{\eta} + 2\sum_{k,h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)\hat{Q}_h^k(s,a)^2 + 2\sum_{k,h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)B_h^k(s,a)^2 + O(H^5K). \qquad (18)
\end{aligned}$$

For the middle term

$$\begin{aligned}
2\sum_{k,h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)\hat{Q}_h^k(s,a)^2 &\le 2\sum_{k,h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)\frac{r_h^k(s,a)^2 H^2 \mathbb{I}\{s_h^k = s, a_h^k = a\}}{(\bar{q}_h^k(s,a)+\gamma)^2} \\
&\le 2H^2\sum_{k,h,s,a} \frac{q_h(s)\,\pi_h^{k+d^k}(a|s) r_h^k(s,a)}{\bar{q}_h^k(s,a)+\gamma} + O\left(\frac{H^3}{\eta^2}\right) \\
&\le \frac{2}{3}H\sum_{k,h,s} q_h(s)\bar{b}_h^k(s) + O\left(\frac{H^3}{\eta^2}\right) \\
&\le \frac{1}{3}\sum_{k,h,s} q_h(s)\bar{b}_h^k(s) + O\left(\frac{H^3}{\eta^2}\right),
\end{aligned}$$

where the second inequality is by E^b and the last is since $\eta \le \frac{1}{2H}$. For the last term in Eq. (18) we use $b_h^k(s) \le 5H$ and therefore $B_h^k(s,a) \le 5H^2$. Thus,

$$\sum_{k,h,s,a} q_h(s)\,\pi_h^{k+d^k}(a|s)B_h^k(s,a)^2 \le 25H^5K.$$

Overall,

$$\text{REG} \le \frac{H\ln A}{\eta} + \frac{1}{3}\sum_{k,h,s} q_h(s)\bar{b}_h^k(s) + O\left(H^5K + \frac{H^3}{\eta^2}\right).$$

□

D. Delay-Adapted Policy Optimization for Adversarial MDP with Linear Q-function

Algorithm 5 Delay-Adapted Policy Optimization with Linear Q-function

Input: feature dimension n, action space $\mathcal{A}$, horizon H, feature map $\{\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n\}_{h=1}^H$, simulator of the environment, horizon H, learning rate $\eta > 0$, exploration parameter $\gamma > 0$, confidence parameter $\delta > 0$.
Initialization: Set approximation parameter $\varepsilon = (HnK)^{-1}$, bonus parameters $\beta_1 = H\sqrt{n}$, $\beta_2 = H\sqrt{n}$, $\beta_r = 2H\sqrt{n}$, $\beta_v = 4H^2$, $\beta_f = H$, $\beta_g = H$ and Geometric Resampling parameters $M = d\frac{24}{\varepsilon^2}\ln\frac{10H^2Kn}{\delta}e$, $N = d\frac{2}{\varepsilon}\log\frac{1}{\varepsilon}e$.
Define $\pi_h^1(a \mid s) = 1/A$ 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 and observe trajectory $\{(s_h^k, a_h^k)\}_{h=1}^H$.
  # Policy Evaluation
  for j such that $j + d^j = k$ **do**
    Observe bandit feedback $\{c_h^j(s_h^j, a_h^j)\}_{h=1}^H$.
    Collect MN trajectories T^j using the simulator and π^j.
    Compute estimated inverse covariance matrix $\{\widehat{\Gamma}_h^{j,\gamma}\}_{h=1}^H = \textsc{Geometric-Resampling}(T^j, M, N, \gamma)$ using the Matrix Geometric Resampling procedure (Neu & Olkhovskaya, 2021) (which samples trajectories of the policy using the simulator).
    Compute Monte-Carlo estimates $L_h^j = \sum_{h'=h}^H c_{h'}^j(s_{h'}^j, a_{h'}^j)$ for every $h \in [H]$.
    Compute estimated Q-function weights $\widehat{w}_h^j = \widehat{\Gamma}_h^{j,\gamma}\phi_h(s_h^j, a_h^j)L_h^j$ for every $h \in [H]$.
    Define delay-adapted ratio $\alpha_h^j(s,a) = \frac{\pi_h^j(a|s)}{\max\{\pi_h^j(a|s), \pi_h^k(a|s)\}}$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
    Define delay-adapted estimated Q-function $\widehat{Q}_h^j(s,a) = r_h^j(s,a) + \phi_h(s,a)^\top \widehat{w}_h^j$ for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
    Define bonus $b_h^j(s,a) = b_h^{j,1}(s) + b_h^{j,2}(s,a) + b_h^{j,v}(s) + b_h^{j,r}(s,a) + b_h^{j,f}(s) + b_h^{j,g}(s,a)$ for every (s,a,h), where:

$$b_h^{j,1}(s) = \frac{1}{\eta}\sum_a r_h^j(s,a)\,\pi_h^{j+d^j}(a \mid s)\,\|\phi_h(s,a)\|_{\widehat{\Gamma}_h^{j,\gamma}}; \qquad b_h^{j,2}(s,a) = \beta_2 r_h^j(s,a)\,\|\phi_h(s,a)\|_{\widehat{\Gamma}_h^{j,\gamma}}$$

$$b_h^{j,v}(s) = \beta_v m^{j+d^j}\sum_a r_h^j(s,a)\,\pi_h^{j+d^j}(a \mid s)\,\|\phi_h(s,a)\|^2_{\widehat{\Gamma}_h^{j,\gamma}}; \qquad b_h^{j,r}(s,a) = \beta_r(1 - r_h^j(s,a))$$

$$b_h^{j,f}(s) = \beta_f \sum_a r_h^j(s,a)\,\pi_h^{j+d^j}(a \mid s)\,\|\phi_h(s,a)\|^2_{\widehat{\Gamma}_h^{j,\gamma}}; \qquad b_h^{j,g}(s,a) = \beta_g r_h^j(s,a)\,\|\phi_h(s,a)\|^2_{\widehat{\Gamma}_h^{j,\gamma}}.$$

  end for
  # Policy Improvement
  Define the policy π^{k+1} for every $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ by:

$$\pi_h^{k+1}(a \mid s) = \frac{\pi_h^k(a \mid s)\exp\left(\eta\sum_{j:j+d^j=k}(\widehat{Q}_h^j(s,a) - \widehat{B}_h^j(s,a))\right)}{\sum_{a' \in \mathcal{A}}\pi_h^k(a' \mid s)\exp\left(\eta\sum_{j:j+d^j=k}(\widehat{Q}_h^j(s,a') - \widehat{B}_h^j(s,a'))\right)};$$

  where $\widehat{B}_h^j(s,a)$ estimates $B_h^j(s,a)$ using the bonus procedure in Algorithm 6.
end for

Theorem D.1. Set $\eta = \sqrt{n/K}$ and $\gamma = \min\{\frac{1}{10Hd_{\max}}, \frac{1}{H(K+D)^{3/4}}\}$ and skip episodes with delay larger than $d_{\max} = D^{1/4}$. Running Algorithm 5 in an adversarial MDP $M = (\mathcal{S}, \mathcal{A}, H, p, \{c^k\}_{k=1}^K)$ with linear Q-function, access to a simulator of the environment, a known features map $\{\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n\}_{h=1}^H$ and delays $\{d^k\}_{k=1}^K$ guarantees, with probability at least $1 - \delta$,

$$R_K \le O\left(H^3 n^{5/4} K^{3/4}\log\frac{KHA}{\delta} + H^2 D^{3/4}\log\frac{KHA}{\delta} + H^5 n\log\frac{KHA}{\delta}\right).$$

Remark D.2. As noted in the main text, Algorithm 6 is not sample efficient, and may require $(KHA)^{O(H)}$ calls to the simulator. However, under the stronger assumption that the MDP is linear (e.g., see assumption 2.1 in Sherman et al. (2023)), we can replace this procedure with the OLSPE procedure of Sherman et al. (2023) which would make our algorithm fully efficient while achieving the same regret guarantee of Theorem D.1.

Algorithm 6 Bonus Procedure (Algorithm 3 in Luo et al. (2021) adapted to non-dilated bonuses)

Input: episode j, horizon h state s, action a, local bonus function b^j, simulator of the environment.
if $\widehat{B}^j_h(s,a)$ was computed before then
  Return the previously computed value $\widehat{B}^j_h(s,a)$.
end if
Compute $\pi^j_h(\cdot\,|\,s)$ and $\pi^{j+d^j}_h(\cdot\,|\,s)$ (which involves recursive calls to Bonus Procedure to compute $\widehat{B}^{j'}_{h'}$ for $j' : j' + d^{j'} < j + d^j$).
Sample $s' \sim p_h(\cdot\,|\,s,a)$ using the simulator.
Compute $\pi^j_{h+1}(\cdot\,|\,s')$ and sample $a' \sim \pi^j_{h+1}(\cdot\,|\,s')$.
Return $b^j_h(s,a) + \widehat{B}^j_{h+1}(s',a')$.

D.1. The good event

Let $\iota = 10\log\frac{10KHA}{\delta}$, $\mathcal{H}^k$ be the history of all episodes $\{j : j + d^j < k\}$ and define $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot\,|\,\mathcal{H}^{k+d^k}]$. Let $\|\cdot\|_{\mathrm{op}}$ be the operator norm. That is, given a matrix $A \in \mathbb{R}^{n\times n}$, $\|A\|_{\mathrm{op}} := \inf\{c \ge 0 : \|Ax\| \le c\|x\|\ \forall x \in \mathbb{R}^n\}$. Define the following events:

$$\mathcal{E} = \left\{\forall k \in [K], h \in [H] : \widehat{\Lambda}^{k,+}_h \succeq \left(I + \Lambda^k_h\right)^{-1}_{\mathrm{op}} \ge \tfrac{1}{2} \text{ and } \widehat{\Lambda}^{k,+}_h \Lambda^k_h{}_{\mathrm{op}} \le 1 + \tfrac{1}{2}\right\},$$

$$\mathcal{E}^b = \left\{\forall h \in [H] : \sum_{k=1}^{K}\sum_{a} \mathbb{E}_{s\sim q_h}\left[\pi^{k+d^k}_h(a\,|\,s)\, m^{k+d^k}\, \widehat{Q}^k_h(s,a)^2\right] \le 2\sum_{k=1}^{K}\sum_{a} \mathbb{E}_{s\sim q_h}\left[\pi^{k+d^k}_h(a\,|\,s)\, m^{k+d^k}\, \mathbb{E}_k\left[\widehat{Q}^k_h(s,a)^2\right]\right] + \frac{4H^2 d_{\max}\log\frac{10H}{\delta}}{\iota^2}\right\},$$

$$\mathcal{E}^f = \left\{\sum_{k=1}^{K}\mathbb{E}_k\left[\sum_h \mathbb{E}_{s\sim q_h}\left[D^{\pi^{k+d^k}_h(\cdot\,|\,s)}, \widehat{Q}^k_h(s,\cdot)\right]\right] \le \sum_{k=1}^{K}\sum_h \mathbb{E}_{s\sim q_h}\left[D^{\pi^{k+d^k}_h(\cdot\,|\,s)}, \widehat{Q}^k_h(s,\cdot)\right] + \sum_{k=1}^{K}\sum_h \mathbb{E}_{s\sim q_h}[b^{k,f}_h(s)] + \frac{H^2\log\frac{10}{\delta}}{\iota} + O(\iota H^2 K)\right\},$$

$$\mathcal{E}^g = \left\{\sum_{k=1}^{K}\sum_h \mathbb{E}_{s\sim q_h}\left[D^{\pi_h(\cdot\,|\,s)}, \widehat{Q}^k_h(s,\cdot)\right] \le \sum_{k=1}^{K}\mathbb{E}_k\left[\sum_h \mathbb{E}_{s\sim q_h}\left[D^{\pi_h(\cdot\,|\,s)}, \widehat{Q}^k_h(s,\cdot)\right]\right] + \sum_{k=1}^{K}\sum_h \mathbb{E}_{s,a\sim q_h}[b^{k,g}_h(s,a)] + \frac{H^2\log\frac{10}{\delta}}{\iota} + O(\iota H^2 K)\right\},$$

$$\mathcal{E}^B = \left\{\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s\sim q_h}\left[D^{\Lambda^k_h(\cdot\,|\,s)\,\pi_h(\cdot\,|\,s)}, \widehat{B}^k_h(s,\cdot)\right] \le \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_k\left[\mathbb{E}_{s\sim q_h}\left[D^{\Lambda^k_h(\cdot\,|\,s)\,\pi_h(\cdot\,|\,s)}, \widehat{B}^k_h(s,\cdot)\right]\right] + O(H^3\sqrt{nK}\ln\tfrac{10}{\delta})\right\}.$$

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

Lemma D.3 (The Good Event) Let $G = \mathcal{E} \cap \mathcal{E}^b \cap \mathcal{E}^f \cap \mathcal{E}^g \cap \mathcal{E}^B$ be the good event. It holds that $\mathbb{P}[G] \ge 1 - \delta$.

Proof. We'll show that each of the events $\mathcal{E}, \mathcal{E}^b, \mathcal{E}^f, \mathcal{E}^g, \mathcal{E}^B$ occures with probability $\ge 1 - \delta/5$. By taking the

union bound we'll get that $\Pr[G] \geq 1 - \cdot$.

Event E : We use GEOMETRICRESAMPLING with $M = \frac{24}{\varepsilon^2 \alpha^2} \ln \frac{10H^2Kn}{\cdot}$ and $N = \frac{2}{\alpha^2} \ln \frac{1}{\cdot}$. Thus, by Lemma D.5, the event holds for each h and k separately with probability of at least $1 - \frac{\cdot}{10HK}$. By taking the union bound over $h \in [H]$ and $k \in [K]$ we get that $\Pr[E] \geq 1 - \cdot/10$.

Event E^b: Fix h. By Lemma D.4 $|\hat{Q}_h^k(s,a)| \leq H$ and thus $\sum_a \mathbb{E}_{s \sim q_h} \pi^{k+d^k}(a \mid s) m^{k+d^k} \left(\hat{Q}_h^k(s,a)\right)^2 \leq \frac{H^2 d_{\max}}{2}$. Also note that π^{k+d^k} is determined by the history H^{k+d^k}. Thus, by Lemma E.4 the event holds with probability $1 - \cdot/(10H)$. The proof is finished by a union bound over h.

Event E^f: Let $Y_k = \sum_h \mathbb{E}_{s \sim q_h} \left\langle \pi_h^{k+d^k}(\cdot \mid s), \hat{Q}_h^k(s,\cdot) \right\rangle$. Similarly to the tabular case, we use a variant of Freedman's inequality (Lemma E.3) to bound $\sum_{k=1}^K X_k := \sum_{k=1}^K \mathbb{E}_k[Y_k] - \sum_{k=1}^K Y_k$.

$$
\begin{aligned}
\mathbb{E}_k\left[X_k^2\right] \leq \mathbb{E}_k\left[Y_k^2\right] &= \mathbb{E}_k\left[\left(\sum_{h,a} \mathbb{E}_{s \sim q_h} \pi_h^{k+d^k}(a \mid s) \hat{Q}_h^k(s,a)\right)^2\right] \\
&= \mathbb{E}_k\left[\sum_h \left(\mathbb{E}_{s \sim q_h, a \sim \pi_h^{k+d^k}(\cdot \mid s)} \hat{Q}_h^k(s,a)\right)^2\right] \\
&\leq H\, \mathbb{E}_k\left[\sum_h \mathbb{E}_{s \sim q_h, a \sim \pi_h^{k+d^k}(\cdot \mid s)} \left|\hat{Q}_h^k(s,a)\right|\right] \\
&\leq H\, \mathbb{E}_k\left[\sum_h \mathbb{E}_{s \sim q_h, a \sim \pi_h^{k+d^k}(\cdot \mid s)} \left(\hat{Q}_h^k(s,a)\right)^2\right] \\
&= H\, \mathbb{E}_k\left[\sum_h \mathbb{E}_{s \sim q_h} \sum_a \pi_h^{k+d^k}(a \mid s) \left(\hat{Q}_h^k(s,a)\right)^2\right] \\
&= H \sum_h \mathbb{E}_{s \sim q_h} \sum_a \pi_h^{k+d^k}(a \mid s) \mathbb{E}_k\left[\left(\hat{Q}_h^k(s,a)\right)^2\right] \\
&\leq H^3 \sum_h \mathbb{E}_{s \sim q_h} \left[\sum_a \pi_h^{k+d^k}(a \mid s) r_h^k(s,a) \left\|\ell_h(s,a)\right\|^2_{\hat{\Sigma}_h^{k;\gamma}}\right] + O(H^4)
\end{aligned}
$$

where the first inequality is since $\|x\|_1^2 \leq n\|x\|_2^2$ for any $x \in \mathbb{R}^n$, the second inequality is by Jensen's inequality, and the last inequality is by Lemma D.10. Also $|Y_k| \leq H^2$. Therefore by Lemma E.3 with probability $1 - \cdot/10$,

$$
\begin{aligned}
\sum_{k=1}^K \mathbb{E}_k[Y_k] - \sum_{k=1}^K Y_k &\leq H \sum_{k=1}^K \sum_h \mathbb{E}_{s \sim q_h} \left[\sum_a \pi_h^{k+d^k}(a \mid s) r_h^k(s,a) \left\|\ell_h(s,a)\right\|^2_{\hat{\Sigma}_h^{k;\gamma}}\right] + \frac{H^2 \log \frac{10}{\cdot}}{\cdot} + O(HK) \\
&= \sum_{k=1}^K \sum_h \mathbb{E}_{s \sim q_h}\left[b_h^{k;f}(s)\right] + \frac{H^2 \log \frac{10}{\cdot}}{\cdot} + O(H^2 K).
\end{aligned}
$$

Event E^g: The proof is similar to event E^f.

Event E^B: Define,

$$
X_k := \sum_{h=1}^H \mathbb{E}_{s \sim q_h}\left[\left\langle \pi_h^k(\cdot \mid s) - \pi_h(\cdot \mid s), \hat{B}_h^k(s,\cdot)\right\rangle\right] - \sum_{h=1}^H \mathbb{E}_k\left[\sum_h \mathbb{E}_{s \sim q_h}\left[\left\langle \pi_h^k(\cdot \mid s) - \pi_h(\cdot \mid s), \hat{B}_h^k(s,\cdot)\right\rangle\right]\right]
$$

and note that $|X_k| \leq O(H^3 \sqrt{n})$. By Azuma–Hoeffding inequality (Lemma E.2) we get that the event holds with probability $1 - \cdot/10$. □

Lemma D.4. For any $(k,h,s,a) \in [K] \times [H] \times S \times A$, it holds that $|\hat{Q}_h^k(s,a)| \le \frac{H}{\gamma}$.

Proof. By Cauchy-Schwartz, Eq. (32) in Lemma E.11 and Assumption 2.1:

$$|\hat{Q}_h^k(s,a)| = |r_h^k(s,a) + \phi_h(s,a)^\top \hat{w}_h^{k,+} (s_h^k, a_h^k) L_h^k| \le H \|\phi(s,a)\|_2 \cdot \|\hat{w}_h^{k,+}(s_h^k, a_h^k)\|_2 \le H \cdot \|\hat{\Sigma}_h^{k,+}\|_{op} (s_h^k, a_h^k) \cdot \|\phi_h^k(s_h^k, a_h^k)\|_2 \le \frac{H}{\gamma}.$$

□

Lemma D.5 (Lemma D.1 in Luo et al. (2021)). Fix a policy π with a covariance matrix $\Sigma_h = \mathbb{E}_{s,a \sim q_h}[\phi(s,a)\phi(s,a)^\top]$. Let $\hat{\Sigma}_h^+$ be the output of GEOMETRIC RESAMPLING(T, M, N, π) with $M \ge \frac{24}{\gamma^2} \ln \frac{nH}{\delta}$ and $N \ge \frac{2}{\gamma^2} \ln \frac{1}{\delta}$ and T are MN trajectories collected with π. Then, $\|\hat{\Sigma}_h^+\|_{op} \le \frac{1}{\gamma}$, and with probability $1-\delta$,

$$\|\hat{\Sigma}_h^+ - (I + \Sigma_h)^{-1}\|_{op} \le \varepsilon_2; \qquad \|\hat{\Sigma}_h^+ \Sigma_h\|_{op} \le 1 + \varepsilon_2.$$

D.2. Proof of the main theorem

Proof of Theorem D.1. By Lemma D.3, the good event holds with probability of at least $1-\delta$. As in the previous section, we analyze the regret under the assumption that the good event holds. We start with the following regret decomposition,

$$\begin{aligned}
R_K &= \sum_{k=1}^K \mathbb{E}_{s \sim q_h}\left[\langle \pi_h^k(\cdot|s) - \pi_h(\cdot|s), Q_h^k(s,\cdot)\rangle\right] = \underbrace{\sum_{k=1}^K \sum_h \mathbb{E}_{s \sim q_h}\left[\langle \pi_h^{k+d^k}(\cdot|s), Q_h^k(s,\cdot) - \hat{Q}_h^k(s,\cdot)\rangle\right]}_{\text{BIAS}_1} \\
&+ \sum_{k=1}^K \sum_h \mathbb{E}_{s \sim q_h}\left[\langle \pi_h(\cdot|s), \hat{Q}_h^k(s,\cdot) - Q_h^k(s,\cdot)\rangle\right] + \sum_{k=1}^K \sum_h \mathbb{E}_{s \sim q_h}\left[\langle \pi_h^k(\cdot|s) - \pi_h(\cdot|s), \hat{B}_h^k(s,\cdot)\rangle\right] \\
&\underbrace{\qquad\qquad}_{\text{BIAS}_2} \qquad\qquad \underbrace{\qquad\qquad}_{\text{BONUS}} \\
&+ \sum_{k=1}^K \sum_h \mathbb{E}_{s \sim q_h}\left[\langle \pi_h(\cdot|s) - \pi_h^{k+d^k}(\cdot|s), \hat{Q}_h^k(s,\cdot) - \hat{B}_h^k(s,\cdot)\rangle\right] \\
&\underbrace{\qquad\qquad}_{\text{REG}} \\
&+ \sum_{k=1}^K \sum_h \mathbb{E}_{s \sim q_h}\left[\langle \pi_h^k(\cdot|s) - \pi_h^{k+d^k}(\cdot|s), \hat{Q}_h^k(s,\cdot) - \hat{B}_h^k(s,\cdot)\rangle\right], \\
&\underbrace{\qquad\qquad}_{\text{DRIFT}}
\end{aligned}$$

where the first is by Lemma E.1. The five terms above are bounded in Lemmas D.6, D.8, D.9, D.11 and D.13. For $< \frac{}{10Hd_{max}}$ and $< (HnK)^{1}$,

$$R_K \quad \underbrace{\sum_{k=1}^{K}\sum_{h}\mathbb{E}_{s\ q_h}\left[b_h^{k;1}(s)+b_h^{k;f}(s)\right]+O\left(-H^3(K+D)^{p}\bar{n}+\frac{H^2}{}+H^{p}\bar{nK}+H^2K\right)}_{\text{BIAS}_1}$$

$$+\underbrace{\sum_{k=1}^{K}\sum_{h}\mathbb{E}_{s;a\ q_h}\left[b_h^{k;2}(s;a)+b_h^{k;r}(s;a)+b_h^{k;g}(s;a)\right]+O\left(\frac{H^2}{}+H^{p}\bar{nK}+H^2K\right)}_{\text{BIAS}_2}$$

$$+\underbrace{\sum_{k=1}^{K}\sum_{h}\mathbb{E}_{s\ q_h}\left[b_h^{k;v}(s)\right]+O\left(\frac{H}{}+H^5nK+\frac{H^2}{}+H^3\right)}_{\text{REG}}+\underbrace{O\left(-H^4\sqrt{\bar{n}}(K+D)\right)}_{\text{DRIFT}}:$$

$$+\underbrace{O\left(\sqrt{-H^2nK}+H^2nK+-H^3\sqrt{\bar{n}}(K+D)+H^3\sqrt{\bar{nK}}\right)+\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{s;a\ q_h}\left[b_h^{k}(s;a)\right]}_{\text{BONUS}}$$

$$O\left(\frac{H}{}+H^5nK+-H^4\sqrt{\bar{n}}(K+D)+\sqrt{-H^2nK}+H^3\sqrt{\bar{nK}}\right):$$

For $= H^{p}\frac{\bar{n}}{K}$ and $= \min\left\{\frac{1}{H(K+D)^{3=4}};\frac{}{10Hd_{max}}\right\}$, we get:

$$R_K \quad O\left(H^3n^{5=4}K^{3=4}+H^2D^{3=4}+H^4nK^{1=4}+H^3\sqrt{\bar{nK}}+H^2d_{max}\sqrt{\bar{K=n}}\right):$$

Finally, by skipping rounds larger than$= D^{1=4}$, we get:$R_K \quad O\left(H^3n^{5=4}K^{3=4}+H^2D^{3=4}+H^5n\right)$: □

D.3. Bound on BIAS_1

Lemma D.6. Under the good event,

$$\text{BIAS}_1 \quad \sum_{k=1}^{K}\sum_{h}\mathbb{E}_{s\ q_h}\left[b_h^{k;1}(s)+b_h^{k;f}(s)\right]+O\left(-H^3\sqrt{\bar{n}}(K+D)+\frac{H^2}{}+H^{p}\bar{nK}+H^2K\right):$$

Proof. We first decompose BIAS_1 as,

$$\text{BIAS}_1=\sum_{k=1}^{K}\sum_{h}\mathbb{E}_{s\ q_h}\left[D_h^{k+d^k}(\ j\ s);Q_h^k(s;\ )\quad\hat{Q}_h^k(s;\ )\right]$$

$$=\underbrace{\sum_{k=1}^{K}\sum_{h}\mathbb{E}_{s\ q_h}\left[D_h^{k+d^k}(\ j\ s);\mathbb{E}_k\left[Q_h^k(s;\ )\right]\quad\hat{Q}_h^k(s;\ )\right]}_{(A)}$$

$$+\underbrace{\sum_{k=1}^{K}\sum_{h}\mathbb{E}_{s\ q_h}\left[D_h^{k+d^k}(\ j\ s);\mathbb{E}_k\left[\hat{Q}_h^k(s;\ )\right]\quad\hat{Q}_h^k(s;\ )\right]}_{(B)};$$

where the (B) $\sum_{k,h} \mathbb{E}_{s \sim q_h}[b_h^{k,f}(s)] + \frac{H^2}{} + O(H^2K)$ by event E^f. For (A) we use Lemma D.7,

$$
\begin{aligned}
(A) &= \sum_{k=1}^{K} \sum_{h,a} \mathbb{E}_{s \sim q_h}\left[\pi_h^{k+d^k}(a \mid s)\, \phi_h(s,a)^\top \left(\Lambda + \Phi_h^k\right)^{-1} \phi_h^k \right] \\
&\quad + \sum_{k=1}^{K} \sum_{h,a} \mathbb{E}_{s \sim q_h}\left[\pi_h^{k+d^k}(a \mid s)(1 - r_h^k(s,a))\, \phi_h(s,a)^\top \left(\Lambda + \Phi_h^k\right)^{-1} \phi_h^k \right] + O(H^2K) \\
&= \underbrace{\sum_{k=1}^{K} \sum_{h,a} \mathbb{E}_{s \sim q_h}\left[r_h^k(s,a)\, \pi_h^{k+d^k}(a \mid s)\, \phi_h(s,a)^\top \left(\Lambda + \Phi_h^k\right)^{-1} \phi_h^k \right]}_{\text{(i)}} \\
&\quad + \underbrace{\sum_{k=1}^{K} \sum_{h,a} \mathbb{E}_{s \sim q_h}\left[\pi_h^{k+d^k}(a \mid s)(1 - r_h^k(s,a))\, \phi_h(s,a)^\top \left(\Lambda + \Phi_h^k\right)^{-1} \phi_h^k \right]}_{\text{(ii)}} \\
&\quad + \underbrace{\sum_{k=1}^{K} \sum_{h,a} \mathbb{E}_{s \sim q_h}\left[\pi_h^{k+d^k}(a \mid s)(1 - r_h^k(s,a))\, \phi_h(s,a)^\top \left(\Lambda + \Phi_h^k\right)^{-1} \phi_h^k \right]}_{\text{(iii)}} + O(H^2K).
\end{aligned}
$$

Using Cauchy-Schwarz inequality,

$$
\begin{aligned}
\text{(i)} &\le \sum_{h,a,k} \mathbb{E}_{s \sim q_h}\left[r_h^k(s,a)\, \pi_h^{k+d^k}(a \mid s) \|\phi_h(s,a)\|_{(\Lambda + \Phi_h^k)^{-1}} \|\phi_h^k\|_{(\Lambda + \Phi_h^k)^{-1}} \right] \\
&\le H\sqrt{n} \sum_{h,a,k} \mathbb{E}_{s \sim q_h}\left[r_h^k(s,a)\, \pi_h^{k+d^k}(a \mid s) \|\phi_h(s,a)\|_{(\Lambda + \Phi_h^k)^{-1}} \right] \\
&\le H\sqrt{n} \sum_{h,k} \mathbb{E}_{s \sim q_h}\left[\sum_{a} r_h^k(s,a)\, \pi_h^{k+d^k}(a \mid s) \|\phi_h(s,a)\|_{\widehat{\Lambda}_h^{k,+}} \right] + O(HK\sqrt{n}) \\
&= \sum_{h,k} \mathbb{E}_{s \sim q_h}\left[b_h^{k,1}(s) \right] + O(HK\sqrt{n});
\end{aligned}
$$

where the second inequality is since $\|\phi_h^k\|_{(\Lambda + \Phi_h^k)^{-1}} = \sqrt{\phi_h^{k\top} (\Lambda + \Phi_h^k)^{-1} \phi_h^k} \le \sqrt{\|\phi_h^k\|_2^2} \le H\sqrt{n}$ by Eq. (31) of Lemma E.11 and Assumption 2.1, and the third is by Cauchy–Schwarz and Event in a similar way to Eq. (21). For term (ii), note that

$$
\|\phi_h(s,a)^\top (\Lambda + \Phi_h^k)^{-1} \phi_h^k\| \le \|\phi_h(s,a)\|_2 \, \|(\Lambda + \Phi_h^k)^{-1} \phi_h^k\|_2 \le \frac{1}{\lambda} \|\phi_h^k\|_2 \le \frac{H\sqrt{n}}{\lambda}; \tag{19}
$$

where the first inequality is by Cauchy–Schwarz, and the second inequality is by Eq. (32) in Lemma E.11. Therefore,

$$
\begin{aligned}
\text{(ii)} &\le H\sqrt{n} \sum_{h,a,k} \mathbb{E}_{s \sim q_h}\left[\pi_h^{k+d^k}(a \mid s)(1 - r_h^k(s,a)) \right] \\
&= H\sqrt{n} \sum_{h,a,k} \mathbb{E}_{s \sim q_h}\left[\pi_h^{k+d^k}(a \mid s) \frac{\max\{\pi_h^{k+d^k}(a \mid s), \pi_h^k(a \mid s)\} - \pi_h^k(a \mid s)}{\max\{\pi_h^{k+d^k}(a \mid s), \pi_h^k(a \mid s)\}} \right] \\
&\le H\sqrt{n} \sum_{h,a,k} \mathbb{E}_{s \sim q_h}\left[\max\{\pi_h^{k+d^k}(a \mid s), \pi_h^k(a \mid s)\} - \pi_h^k(a \mid s) \right] \\
&\le H\sqrt{n} \sum_{h,k} \mathbb{E}_{s \sim q_h}\left[\|\pi_h^{k+d^k}(\cdot \mid s) - \pi_h^k(\cdot \mid s)\|_1 \right] \\
&\le O\left(H^3\sqrt{n}(K + D) \right);
\end{aligned}
$$

where the lsat is by Lemma D.12. Similarly, using Eq. (33) in Lemma E.11,

$$\left\| \phi_h(s,a)^\top \left(I + \Sigma_h^k\right)^{-1} \Sigma_h^k \right\| \le \left\| \phi_h(s,a)^\top \left(I + \Sigma_h^k\right)^{-1} \right\|_2 \le \left\| \left(I + \Sigma_h^k\right)^{-1} \Sigma_h^k \right\|_2 \le H\sqrt{n}. \tag{20}$$

Thus,(iii) is bounded in the same way as(ii). □

Lemma D.7. Under the good event, for every$(k,h,s,a) \in [K] \times [H] \times S \times A$, it holds that

$$\mathbb{E}_k\left[Q_h^k(s,a) - \widehat{Q}_h^k(s,a)\right] \le \phi_h(s,a)^\top \left(I + \Sigma_h^k\right)^{-1} \theta_h^k + (1 - r_h^k(s,a))\, \phi_h(s,a)^\top \left(I + \Sigma_h^k\right)^{-1} \Sigma_h^k + O(H).$$

Proof. By the de nition of$Q_h^k(s,a)$, $\widehat{Q}_h^k(s,a)$ and$\widehat{\theta}_h^k$,

$$\begin{aligned}
\mathbb{E}_k\left[Q_h^k(s,a) - \widehat{Q}_h^k(s,a)\right] &= \phi_h(s,a)^\top \theta_h^k - r_h^k(s,a)\mathbb{E}_k\left[\widehat{\theta}_h^k\right] \\
&= \phi_h(s,a)^\top \theta_h^k - r_h^k(s,a)\mathbb{E}_k\left[\widehat{\theta}_h^{k,+}\right] - \mathbb{E}_k\left[\phi_h(s_h^k,a_h^k) L_h^j\right] \\
&= \phi_h(s,a)^\top \theta_h^k - r_h^k(s,a)\left(I + \Sigma_h^k\right)^{-1}\mathbb{E}_k\left[\phi_h(s_h^k,a_h^k) L_h^j\right] + O(H) \\
&= \phi_h(s,a)^\top \theta_h^k - r_h^k(s,a)\left(I + \Sigma_h^k\right)^{-1}\mathbb{E}_k\left[\phi_h(s_h^k,a_h^k)\mathbb{E}\left[L_h^j \mid s_h^k,a_h^k\right]\right] + O(H) \\
&= \phi_h(s,a)^\top \theta_h^k - r_h^k(s,a)\left(I + \Sigma_h^k\right)^{-1}\mathbb{E}_k\left[\phi_h(s_h^k,a_h^k) Q_h^j(s_h^k,a_h^k)\right] + O(H) \\
&= \phi_h(s,a)^\top \theta_h^k - r_h^k(s,a)\left(I + \Sigma_h^k\right)^{-1}\mathbb{E}_k\left[\phi_h(s_h^k,a_h^k)\phi_h(s_h^k,a_h^k)^\top\right]\theta_h^k + O(H) \\
&= \phi_h(s,a)^\top \theta_h^k - r_h^k(s,a)\left(I + \Sigma_h^k\right)^{-1}\Sigma_h^k\theta_h^k + O(H) \\
&= \phi_h(s,a)^\top \left(I + \Sigma_h^k\right)^{-1}\left(I + \Sigma_h^k\right)\theta_h^k - r_h^k(s,a)\left(I + \Sigma_h^k\right)^{-1}\Sigma_h^k\theta_h^k + O(H) \\
&= \phi_h(s,a)^\top \left(I + \Sigma_h^k\right)^{-1}\theta_h^k + (1 - r_h^k(s,a))\,\phi_h(s,a)^\top \left(I + \Sigma_h^k\right)^{-1}\Sigma_h^k\theta_h^k + O(H)
\end{aligned}$$

where the third equality is since,

$$\begin{aligned}
&\phi_h(s,a)^\top r_h^k(s,a)\left(\mathbb{E}_k\left[\widehat{\theta}_h^{k,+}\right] - \left(I + \Sigma_h^k\right)^{-1}\right)\mathbb{E}_k\left[\phi_h(s_h^k,a_h^k) L_h^j\right] \\
&\quad \le \left\|\phi_h(s,a)\right\|_2 \left\|\left(\mathbb{E}_k\left[\widehat{\theta}_h^{k,+}\right] - \left(I + \Sigma_h^k\right)^{-1}\right)\mathbb{E}_k\left[\phi_h(s_h^k,a_h^k) L_h^j\right]\right\|_2 \\
&\quad \le H\left\|\widehat{\theta}_h^{k,+} - \left(I + \Sigma_h^k\right)^{-1}\right\|_{op} \left\|\mathbb{E}_k\left[\phi_h(s_h^k,a_h^k)\right]\right\|_2 \le 2H,
\end{aligned} \tag{21}$$

where the rst inequality is by Cauchy–Schwarz, the second inequality is by Assumption 2.1 and the last is by event Eand Assumption 2.1. In the same way we have,

$$\phi_h(s,a)^\top \left(I + \Sigma_h^k\right)^{-1}\theta_h^k + (1 - r_h^k(s,a))\,\phi_h(s,a)^\top \left(I + \Sigma_h^k\right)^{-1}\Sigma_h^k \le \mathbb{E}_k\left[Q_h^k(s,a) - \widehat{Q}_h^k(s,a)\right] + O(H). \quad \square$$

D.4. Bound on$\textsc{Bias}_2$

Lemma D.8. Under the good event,

$$\textsc{Bias}_2 \le \sum_{k=1}^K \sum_h \mathbb{E}_{s,a\sim q_h}\left[b_h^{k,2}(s,a) + b_h^{k,r}(s,a) + b_h^{k,g}(s,a)\right] + O\left(\frac{H^2}{\eta} + H\sqrt{nK} + H^2\eta K\right).$$

Proof. Similarly to$\textsc{Bias}_1$, we rst decompose,

$$\begin{aligned}
\textsc{Bias}_2 &= \sum_{k=1}^K \sum_h \mathbb{E}_{s\sim q_h}\left[\left\langle \phi_h(\cdot,s), \widehat{Q}_h^k(s,\cdot) - Q_h^k(s,\cdot)\right\rangle\right] \\
&= \underbrace{\sum_{k=1}^K \sum_h \mathbb{E}_{s\sim q_h}\left[\left\langle \phi_h(\cdot,s), \mathbb{E}_k\left[\widehat{Q}_h^k(s,\cdot) - Q_h^k(s,\cdot)\right]\right\rangle\right]}_{(A)} + \underbrace{\sum_{k=1}^K \sum_h \mathbb{E}_{s\sim q_h}\left[\left\langle \phi_h(\cdot,s), \widehat{Q}_h^k(s,\cdot) - \mathbb{E}_k\left[\widehat{Q}_h^k(s,\cdot)\right]\right\rangle\right]}_{(B)};
\end{aligned}$$

where (B) $\sum_{k,h} \mathbb{E}_{s,a\sim q_h}[b_h^{k,g}(s,a)] + \frac{H^2}{} + O(H^2K)$ by event E^g, and for (A) we use again Lemma D.7,

$$
\begin{aligned}
(A) &= \sum_{k=1}^{K}\sum_{h,a} \mathbb{E}_{s\sim q_h}\left[\pi_h(a|s)|\ell_h(s,a)^\top (I+\Sigma_h^k)^{-1}\ell_h^k|\right] \\
&\quad + \sum_{k=1}^{K}\sum_{h,a} \mathbb{E}_{s\sim q_h}\left[\pi_h(a|s)(r_h^k(s,a)-1)\ell_h(s,a)^\top (I+\Sigma_h^k)^{-1}\ell_h^k\right] + O(H^2K) \\
&= \sum_{k=1}^{K}\sum_{h,a} \mathbb{E}_{s\sim q_h}\Big[r_h^k(s,a)\pi_h(a|s)|\ell_h(s,a)^\top (I+\Sigma_h^k)^{-1}\ell_h^k|\Big] \\
&\quad \underbrace{\qquad\qquad\qquad\qquad}_{(i)} \\
&\quad + \sum_{k=1}^{K}\sum_{h,a} \mathbb{E}_{s\sim q_h}\left[\pi_h(a|s)(1-r_h^k(s,a))|\ell_h(s,a)^\top (I+\Sigma_h^k)^{-1}\ell_h^k|\right] \\
&\quad \underbrace{\qquad\qquad\qquad\qquad}_{(ii)} \\
&\quad + \sum_{k=1}^{K}\sum_{h,a} \mathbb{E}_{s\sim q_h}\left[\pi_h(a|s)(r_h^k(s,a)-1)\ell_h(s,a)^\top (I+\Sigma_h^k)^{-1}\ell_h^k\right] + O(H^2K) \\
&\quad \underbrace{\qquad\qquad\qquad\qquad}_{(iii)}
\end{aligned}
$$

Similar to (i) in Lemma D.6

$$
\begin{aligned}
(i) &\le \sum_{h,a} \mathbb{E}_{s\sim q_h}\left[r_h^k(s,a)\pi_h(a|s)\|\ell_h(s,a)\|_{(I+\Sigma_h^k)^{-1}}\|\ell_h^k\|_{(I+\Sigma_h^k)^{-1}}\right] \\
&\le H\sqrt{n}\sum_{h,a} \mathbb{E}_{s\sim q_h}\left[r_h^k(s,a)\pi_h(a|s)\|\ell_h(s,a)\|_{(I+\Sigma_h^k)^{-1}}\right] \\
&\le H\sqrt{n}\sum_{h} \mathbb{E}_{s\sim q_h}\left[\sum_{a}\pi_h(a|s)r_h^k(s,a)\|\ell_h(s,a)\|_{(\hat{\Sigma}_h^{k,+})^{-1}}\right] + O(H\sqrt{n}) \\
&= \sum_{h} \mathbb{E}_{s,a\sim q_h}\left[b_h^{k,2}(s,a)\right] + O(H\sqrt{n}).
\end{aligned}
$$

Using Eq. (19) and Eq. (20),

$$
\begin{aligned}
(ii) &\quad H\sqrt{n}\sum_{h,a} \mathbb{E}_{s\sim q_h}\left[\pi_h(a|s)(1-r_h^k(s,a))\right] = \frac{1}{2}\sum_{h} \mathbb{E}_{s,a\sim q_h}\left[b_h^{k,r}(s,a)\right] \\
(iii) &\quad H\sqrt{n}\sum_{h,a} \mathbb{E}_{s\sim q_h}\left[\pi_h(a|s)(1-r_h^k(s,a))\right] = \frac{1}{2}\sum_{h} \mathbb{E}_{s,a\sim q_h}\left[b_h^{k,r}(s,a)\right]:
\end{aligned}
$$

□

D.5. Bound on REG

Lemma D.9. Under the good event, for $\le \frac{}{10Hd_{max}}$,

$$
\mathrm{REG} \le \sum_{k=1}^{K}\sum_{h} \mathbb{E}_{s\sim q_h}\left[b_h^{k,v}(s)\right] + O\left(H^5nK + \frac{H\log A}{} + \frac{H^2}{} + H^3\right):
$$

Proof. Note that $(\widehat{Q}^k_h(s,a) - \widehat{B}^k_h(s,a)) \le 1$. Thus, using lemma Lemma E.8 for each (h,s),

$$\sum_{k=1}^{K} \left\langle q^{k+d^k}_h(\cdot|s) - \pi_h(\cdot|s), \widehat{Q}^k_h(s,\cdot) - \widehat{B}^k_h(s,\cdot) \right\rangle$$
$$\le \frac{\log A}{\eta} + \sum_{k=1}^{K}\sum_{a} q^{k+d^k}(a|s)\, m^{k+d^k} \left(\widehat{Q}^k_h(s,a) - \widehat{B}^k_h(s,a)\right)^2$$
$$\le \frac{\log A}{\eta} + 2\sum_{k=1}^{K}\sum_{a} q^{k+d^k}(a|s)\, m^{k+d^k} \left(\widehat{Q}^k_h(s,a)\right)^2 + 2\sum_{k=1}^{K}\sum_{a} q^{k+d^k}(a|s)\, m^{k+d^k} \left(\widehat{B}^k_h(s,a)\right)^2. \quad (22)$$

Note that $b^k_h(s,a) \le O(H\sqrt{n})$. Thus $\widehat{B}^k_h(s,a) \le O(H^2\sqrt{n})$, and the last sum can be bounded by,

$$\sum_{k,a} q^{k+d^k}(a|s)\, m^{k+d^k} \left(\widehat{B}^k_h(s,a)\right)^2 \le O(H^4 n) \sum_{k,a} q^{k+d^k}(a|s)\, m^{k+d^k} = O(H^4 n)\sum_{k=1}^{K} m^{k+d^k} \le O(H^4 nK). \quad (23)$$

Taking the expectation with respect to q_h on the first sum in (22), by event E^b,

$$2\sum_{k=1}^{K}\sum_{a} \mathbb{E}_{s\sim q_h}\, q^{k+d^k}(a|s)\, m^{k+d^k} \left(\widehat{Q}^k_h(s,a)\right)^2$$
$$\le 4\sum_{k=1}^{K}\sum_{a} \mathbb{E}_{s\sim q_h}\, q^{k+d^k}(a|s)\, m^{k+d^k}\, \mathbb{E}_k\left(\widehat{Q}^k_h(s,a)\right)^2 + \frac{8H^2 d_{\max}}{2}$$
$$\le 4\sum_{k=1}^{K}\sum_{a} \mathbb{E}_{s\sim q_h}\, q^{k+d^k}(a|s)\, m^{k+d^k}\, \mathbb{E}_k\left(\widehat{Q}^k_h(s,a)\right)^2 + \frac{H}{4}, \quad (24)$$

where the last is since $\eta < \frac{1}{10Hd_{\max}}$. Finally, by Lemma D.10,

$$4\sum_{a} q^{k+d^k}(a|s)\, m^{k+d^k}\, \mathbb{E}_k\left(\widehat{Q}^k_h(s,a)\right)^2 \le 4H^2 \sum_{a} q^{k+d^k}(a|s)\, m^{k+d^k} r^k_h(s,a)\, \|\phi_h(s,a)\|^2_{(\Lambda^{k,+}_h)^{-1}} + O(H^2)$$
$$= b^{k,v}_h(s) + O(H^2). \quad (25)$$

Combining the above with Eqs. (22) to (25), summing over k and taking the expectation completes the proof. □

Lemma D.10. Under event E, for every $(k,h,s,a) \in [K]\times[H]\times S\times A$, it holds that

$$\mathbb{E}_k\left(\widehat{Q}^k_h(s,a)\right)^2 \le H^2 r^k_h(s,a)\, \|\phi_h(s,a)\|^2_{(\Lambda^{k,+}_h)^{-1}} + O(H^2).$$

Proof. By the definition of $\widehat{Q}^k_h(s,a)$ and Λ^k_h,

$$\mathbb{E}_k\left(\widehat{Q}^k_h(s,a)\right)^2 = \mathbb{E}_k\left[r^k_h(s,a)^2\, \phi_h(s,a)^\top \Lambda^{k,+}_h \Lambda^k_h \Lambda^{k,+}_h \phi_h(s,a)\right]$$
$$\le H^2 \mathbb{E}_k\left[r^k_h(s,a)^2\, \phi_h(s,a)^\top \Lambda^{k,+}_h\, \phi_h(s^k_h,a^k_h)\, \phi_h(s^k_h,a^k_h)^\top \Lambda^{k,+}_h \phi_h(s,a)\right]$$
$$= H^2\, r^k_h(s,a)^2\, \phi_h(s,a)^\top \Lambda^{k,+}_h\, \mathbb{E}_k\left[\phi_h(s^k_h,a^k_h)\, \phi_h(s^k_h,a^k_h)^\top\right] \Lambda^{k,+}_h \phi_h(s,a)$$
$$= H^2\, r^k_h(s,a)^2\, \phi_h(s,a)^\top \Lambda^{k,+}_h\, \Lambda^k_h\, \Lambda^{k,+}_h \phi_h(s,a)$$
$$\le H^2 r^k_h(s,a)\, \phi_h(s,a)^\top \Lambda^{k,+}_h\, \Lambda^k_h\, \Lambda^{k,+}_h \phi_h(s,a). \quad (26)$$

We rewrite,

$$\begin{aligned}
{}_h(s;a)^\top \hat{}^{k;+}_h \; {}^k_h \hat{}^{k;+}_h \; {}_h(s;a) = & \underbrace{{}_h(s;a)^\top \hat{}^{k;+}_h \; {}^k_h \; \hat{}^{k;+}_h \; (I + {}^k_h)^{-1} \; {}_h(s;a)}_{\text{(i)}} \\
& + \underbrace{{}_h(s;a)^\top \; \hat{}^{k;+}_h \; (I + {}^k_h)^{-1} \; {}^k_h (I + {}^k_h)^{-1} \; {}_h(s;a)}_{\text{(ii)}} \\
& + \underbrace{{}_h(s;a)^\top (I + {}^k_h)^{-1} \; {}^k_h (I + {}^k_h)^{-1} \; {}_h(s;a)}_{\text{(iii)}} : \qquad (27)
\end{aligned}$$

We now bound each of the above as follows:

$$\begin{aligned}
\text{(i)} \quad & {}_h(s;a) \hat{}^{k;+}_h \; {}^k_h \; {}_2 \quad \hat{}^{k;+}_h \; (I + {}^k_h)^{-1} \; {}_h(s;a) \; {}_2 \qquad \text{(Cauchy–Schwarz)} \\
& \| {}_h(s;a)\|_2 \; \hat{}^{k;+}_h \; {}^k_h \; {}_{\text{op}} \quad \hat{}^{k;+}_h \; (I + {}^k_h)^{-1} \; {}_{\text{op}} \| {}_h(s;a)\|_2 \\
& \hat{}^{k;+}_h \; {}^k_h \; {}_{\text{op}} \quad \hat{}^{k;+}_h \; (I + {}^k_h)^{-1} \; {}_{\text{op}} \qquad (\| {}_h(s;a)\| \quad 1) \\
& (1 +)2 ; \qquad (28)
\end{aligned}$$

where the last inequality is by event E_t. Similarly,

$$\begin{aligned}
\text{(ii)} \quad & \| {}_h(s;a)\|_2 \quad \hat{}^{k;+}_h \; (I + {}^k_h)^{-1} \; {}^k_h (I + {}^k_h)^{-1} \; {}_h(s;a) \; {}_2 \\
& \hat{}^{k;+}_h \; (I + {}^k_h)^{-1} \; {}_{\text{op}} \; {}^k_h (I + {}^k_h)^{-1} \; {}_h(s;a) \; {}_2 \\
& 2 \; {}^k_h (I + {}^k_h)^{-1} \; {}_h(s;a) \; {}_2 \\
& 2 \| {}_h(s;a)\|_2 \quad 2 ; \qquad (29)
\end{aligned}$$

where the third inequality is by event E_t and the forth inequality uses Eq. (33) in Lemma E.11. Finally, using Eq. (35) in Lemma E.11,

$$\begin{aligned}
\text{(iii)} \quad & {}_h(s;a)^\top (I + {}^k_h)^{-1} \; {}_h(s;a) \\
& = {}_h(s;a)^\top \; (I + {}^k_h)^{-1} \; \hat{}^{k;+}_h \quad {}_h(s;a) + \| {}_h(s;a)\|^2_{\hat{}^{k;+}_h} \\
& \| {}_h(s;a)\|_2 \; (I + {}^k_h)^{-1} \; \hat{}^{k;+}_h \quad {}_h(s;a) \; {}_2 + \| {}_h(s;a)\|^2_{\hat{}^{k;+}_h} \\
& (I + {}^k_h)^{-1} \; \hat{}^{k;+}_h \quad {}_{\text{op}} + \| {}_h(s;a)\|^2_{\hat{}^{k;+}_h} \\
& 2 + \| {}_h(s;a)\|^2_{\hat{}^{k;+}_h} : \qquad (30)
\end{aligned}$$

Combining Eqs. (26) to (30) completes the proof. □

D.6. Bound on DRIFT

Lemma D.11. If $\frac{1}{H^2 \sqrt{n}}$ and $\frac{\sqrt{n}}{4Hd_{\max}}$ then, $\text{DRIFT} \quad O \quad -H^4 \sqrt{n}(K + D)$:

Proof. Note that if $< \frac{\sqrt{n}}{4Hd_{\max}}$ then $b^k_h(s;a) \quad O \; (H \sqrt{n})$ and thus $B^k_h(s;a) \quad O \; (H^2 \sqrt{n})$. Now, using Lemma D.12,

$$\begin{aligned}
\text{DRIFT} = & \sum_{k=1}^{K} \sum_h E_{s \; q_h} \left[D \; {}^k_h(\; | s) \quad {}^{k+d^k}_h(\; | s); Q^k_h(s;) \quad \hat{B}^k_h(s;) \right] \\
& O \; (H^2 \sqrt{n}) \sum_{k=1}^{K} \sum_h E_{s \; q_h} \left[\| \; {}^k_h(\; | s) \quad {}^{k+d^k}_h(\; | s)\|_1 \right] \quad O \quad -H^4 \sqrt{n}(K + D) : \qquad \square
\end{aligned}$$

Lemma D.12. If $\eta \le \frac{1}{H^2\sqrt{n}}$ and $\gamma \le \frac{\sqrt{n}}{4Hd_{\max}}$ then for each (h,s): $\sum_{k=1}^{K} \|q_h^{k+d^k}(\cdot|s) - q_h^k(\cdot|s)\|_1 \le O(\eta H(K+D))$.

Proof. We apply lemma Lemma E.9 for each (k,h,s,a) with $\ell(a) = \sum_{j:k \le j+d^j < k+d^k} (\hat{Q}_h^j(s,a) - \hat{B}_h^j(s,a))$ and $M = \sum_{j:k \le j+d^j < k+d^k} \frac{10H}{\sqrt{n}}$ since $|\hat{Q}_h^k(s,a)| \le \frac{H}{\gamma}$ by Lemma D.4, and $b_h^k(s,a) \le 6H^2\sqrt{n}$ whenever $\gamma \le \frac{\sqrt{n}}{4Hd_{\max}}$ which implies that $\hat{B}_h^k(s,a) \le 6H^2\sqrt{n} \le \frac{6H}{\gamma}$. We get that,

$$\begin{aligned}
\|q_h^{k+d^k}(\cdot|s) - q_h^k(\cdot|s)\|_1 &= \sum_a |q_h^{k+d^k}(a|s) - q_h^k(a|s)| \\
&\le \sum_a \sum_{a'} q_h^{k+d^k}(a|s) q_h^k(a'|s) \sum_{j:k \le j+d^j < k+d^k} |\hat{Q}_h^j(s,a') - \hat{Q}_h^j(s,a)| + |\hat{B}_h^j(s,a') - \hat{B}_h^j(s,a)| + \frac{10H}{\sqrt{n}} \\
&\quad + \sum_a q_h^k(a|s) \sum_{j:k \le j+d^j < k+d^k} |\hat{Q}_h^j(s,a) - \hat{Q}_h^j(s,a)| + |\hat{B}_h^j(s,a) - \hat{B}_h^j(s,a)| + \frac{10H}{\sqrt{n}} \\
&\le \sum_a q_h^{k+d^k}(a|s) \sum_{a'} q_h^k(a'|s) \sum_{j:k \le j+d^j < k+d^k} \frac{20H}{\gamma} + \sum_a q_h^k(a|s) \sum_{j:k \le j+d^j < k+d^k} \frac{20H}{\gamma} \\
&= \sum_{j:k \le j+d^j < k+d^k} \frac{40H}{\gamma}.
\end{aligned}$$

Summing the above over k and applying Lemma E.10 completes the proof. □

D.7. Bound on BONUS

Lemma D.13. Under the good event,

$$\text{BONUS} \le \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{s,a\sim q_h^k}[b_h^k(s,a)] + O\left(\eta H^2 nK + \frac{H^2 nK}{\gamma} + \eta H^3\sqrt{n}(K+D) + \frac{H^3\sqrt{nK}}{\gamma}\right).$$

Proof. Under event E^B,

$$\begin{aligned}
\sum_{h,k} \mathbb{E}_{s\sim q_h}\left[\langle q_h^k(\cdot|s) - q_h^{k+d^k}(\cdot|s), \hat{B}_h^k(s,\cdot)\rangle\right] &\overset{(i)}{=} \sum_{h,k} \mathbb{E}_k\left[\mathbb{E}_{s\sim q_h}\left[\langle q_h^k(\cdot|s) - q_h^{k+d^k}(\cdot|s), \hat{B}_h^k(s,\cdot)\rangle\right]\right] + O(H^3\sqrt{nK}) \\
&\overset{(ii)}{=} \sum_{h,k} \mathbb{E}_{s,a\sim q_h^k}[b_h^k(s,a)] - \sum_{h,k} \mathbb{E}_{s,a\sim q_h}[b_h^k(s,a)] + O(H^3\sqrt{nK}),
\end{aligned}$$

where the equality is by Lemma D.14. Recall that $b_h^k(s,a) = b_h^{k,1}(s) + b_h^{k,2}(s,a) + b_h^{k,v}(s) + b_h^{k,r}(s,a) + b_h^{k,f}(s) + b_h^{k,g}(s,a)$. Now, the expectation $\mathbb{E}_{s,a\sim q_h^k}[\cdot]$ of each of the bonus terms is bounded in Lemmas D.15 to D.20. □

Lemma D.14. For any k, let $b_h^k(s,a)$ be a loss function determined by the history $\mathcal{H}^{k+d^k}$, and let $\hat{B}_h^k(s,a)$ be a randomized bonus function such that, for every $(h,s,a) \in [H] \times S \times A$,

$$\mathbb{E}_k[\hat{B}_h^k(s,a)] = b_h^k(s,a) + \sum_{s'} P_h(s'|s,a)\sum_{a'} q_{h+1}^k(a'|s')\mathbb{E}_k[\hat{B}_{h+1}^k(s',a')]$$

Then, $\mathbb{E}_k\left[\sum_h \mathbb{E}_{s\sim q_h}\left[\langle q_h^k(\cdot|s) - q_h^{k+d^k}(\cdot|s), \hat{B}_h^k(s,\cdot)\rangle\right]\right] = \sum_h \mathbb{E}_{s,a\sim q_h^k}[b_h^k(s,a)] - \sum_h \mathbb{E}_{s,a\sim q_h}[b_h^k(s,a)]$.

Proof. Fix some $h \in [H]$,

$$\begin{aligned}
\mathbb{E}_k\left[\mathbb{E}_{s\sim q_h}\left[\pi^k_h(\cdot|s) - \pi_h(\cdot|s), \widehat{B}^k_h(s,\cdot)\right]\right] &= \mathbb{E}_k\left[\mathbb{E}_{s\sim q_h}\left[\sum_a \pi^k_h(a|s)\widehat{B}^k_h(s,a)\right]\right] - \mathbb{E}_{s,a\sim q_h}\left[\mathbb{E}_k[\widehat{B}^k_h(s,a)]\right] \\
&= \mathbb{E}_k\left[\mathbb{E}_{s\sim q_h}\left[\sum_a \pi^k_h(a|s)\widehat{B}^k_h(s,a)\right]\right] - \mathbb{E}_{s,a\sim q_h}\left[b^k_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\mathbb{E}_{a'\sim\pi^k_{h+1}(\cdot|s')}\mathbb{E}_k[\widehat{B}^k_{h+1}(s',a')]\right] \\
&= \mathbb{E}_k\mathbb{E}_{s\sim q_h}\left[\sum_a \pi^k_h(a|s)\widehat{B}^k_h(s,a)\right] - \mathbb{E}_{s'\sim q_{h+1}}\mathbb{E}_{a'\sim\pi^k_{h+1}(\cdot|s')}\mathbb{E}_k[\widehat{B}^k_{h+1}(s',a')] - \mathbb{E}_{s,a\sim q_h}[b^k_h(s,a)] \\
&= \mathbb{E}_k\mathbb{E}_{s\sim q_h}\left[\sum_a \pi^k_h(a|s)\widehat{B}^k_h(s,a)\right] - \mathbb{E}_k\mathbb{E}_{s\sim q_{h+1}}\left[\sum_a \pi^k_{h+1}(a|s)\widehat{B}^k_{h+1}(s,a)\right] - \mathbb{E}_{s,a\sim q_h}[b^k_h(s,a)].
\end{aligned}$$

Summing over h we get,

$$\begin{aligned}
\sum_h \mathbb{E}_k\left[\mathbb{E}_{s\sim q_h}\left[\pi^k_h(\cdot|s) - \pi_h(\cdot|s), \widehat{B}^k_h(s,\cdot)\right]\right] &= \mathbb{E}_k\mathbb{E}_{s\sim q_1}\left[\sum_a \pi^k_1(a|s)\widehat{B}^k_1(s,a)\right] - \sum_h \mathbb{E}_{s,a\sim q_h}[b^k_h(s,a)] \\
&= \mathbb{E}_k\mathbb{E}_{s,a\sim q^k_1}\left[\widehat{B}^k_1(s,a)\right] - \sum_h \mathbb{E}_{s,a\sim q_h}[b^k_h(s,a)].
\end{aligned}$$

For last,

$$\begin{aligned}
\mathbb{E}_{s,a\sim q^k_1}\left[\widehat{B}^k_1(s,a)\right] &= \mathbb{E}_{s,a\sim q^k_1}[b^k_1(s,a)] + \mathbb{E}_{s'\sim P_1(\cdot|s,a)}\mathbb{E}_{a'\sim\pi^k_2(\cdot|s')}\mathbb{E}_k[\widehat{B}^k_2(s',a')] \\
&= \mathbb{E}_{s,a\sim q^k_1}[b^k_1(s,a)] + \mathbb{E}_{s,a\sim q^k_2}\mathbb{E}_{s'\sim P_1(\cdot|s,a)}\mathbb{E}_{a'\sim\pi^k_2(\cdot|s')}\mathbb{E}_k[\widehat{B}^k_{h+1}(s',a')] \\
&= \mathbb{E}_{s,a\sim q^k_1}[b^k_1(s,a)] + \mathbb{E}_{s,a\sim q^k_2}\left[\widehat{B}^k_2(s,a)\right] = \cdots = \sum_{h=1}^H \mathbb{E}_{s,a\sim q^k_h}[b^k_h(s,a)]. \qquad \square
\end{aligned}$$

Lemma D.15. Under the good event, $\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{s\sim q^k_h}\left[b^{k,1}_h(s)\right] \le O\left(\sqrt{(1+\gamma)H^2nK}\right)$.

Proof. For any k and h,

$$\begin{aligned}
\mathbb{E}_{s\sim q^k_h}\left[b^{k,1}_h(s)\right] &= \sum_a \mathbb{E}_{s\sim q^k_h}\left[r^k_h(s,a)\pi^{k+d^k}_h(a|s)\|\phi_h(s,a)\|_{\Lambda^{k,+}_h}\right] \le \sum_a \mathbb{E}_{s\sim q^k_h}\left[\pi^k_h(a|s)\|\phi_h(s,a)\|_{\Lambda^{k,+}_h}\right] \\
&= \mathbb{E}_{(s,a)\sim q^k_h}\left[\sqrt{\phi_h(s,a)^\top \Lambda^{k,+}_h \phi_h(s,a)}\right] \le \sqrt{\mathbb{E}_{(s,a)\sim q^k_h}\left[\phi_h(s,a)^\top \Lambda^{k,+}_h \phi_h(s,a)\right]} \\
&= \sqrt{\mathbb{E}_{(s,a)\sim q^k_h}\left[\mathrm{tr}\left(\phi_h(s,a)^\top \Lambda^{k,+}_h \phi_h(s,a)\right)\right]} = \sqrt{\mathbb{E}_{(s,a)\sim q^k_h}\left[\mathrm{tr}\left(\Lambda^{k,+}_h \phi_h(s,a)\phi_h(s,a)^\top\right)\right]} \\
&= \sqrt{\mathrm{tr}\left(\Lambda^{k,+}_h \mathbb{E}_{(s,a)\sim q^k_h}\left[\phi_h(s,a)\phi_h(s,a)^\top\right]\right)} = \sqrt{\mathrm{tr}\left(\Lambda^{k,+}_h \Sigma^k_h\right)},
\end{aligned}$$

where the second inequality is Jensen's inequality, and in the last three equalities we use the cyclic property of trace, the linearity of trace and the de nition of Σ^k_h. Finally,

$$\mathrm{tr}\left(\Lambda^{k,+}_h \Sigma^k_h\right) = \sum_{i=1}^n e_i^\top \Lambda^{k,+}_h \Sigma^k_h e_i \le \sum_{i=1}^n \|e_i\|_2 \left\|\Lambda^{k,+}_h \Sigma^k_h e_i\right\|_2 \le \sum_{i=1}^n \left\|\Lambda^{k,+}_h \Sigma^k_h\right\|_{\mathrm{op}} \le \sum_{i=1}^n (1+2\gamma) = n(1+2\gamma);$$

where the first inequality is Cauchy–Schwarz and the last inequality is by $\square$

Lemma D.16. Under the good event, $\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{s,a\sim q^k_h}\left[b^{k,2}_h(s,a)\right] \le O\left(\sqrt{(1+\gamma)H^2nK}\right)$.

Proof. For any h and k,

$$\mathbb{E}_{s,a\sim q_h^k}\left[b_h^{k,2}(s,a)\right] = \gamma^2\mathbb{E}_{s,a\sim q_h^k}\left[r_h^k(s,a)\mathbb{1}_h(s,a)\mathbb{1}_{\hat{h},h}^{k,\tau}\right] \le \gamma^2\mathbb{E}_{s,a\sim q_h^k}\left[\mathbb{1}_h(s,a)\mathbb{1}_{\hat{h},h}^{k,\tau}\right] \le \gamma^2\frac{1}{n(1+2\varepsilon)},$$

where the last is as in the proof of Lemma D.15. □

Lemma D.17. Under the good event, $\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{s,a\sim q_h^k}\left[b_h^{k,r}(s,a)\right] \le O\left(\gamma H^3\sqrt{\bar{n}(K+D)}\right)$.

Proof. For any s and h,

$$\begin{aligned}
\mathbb{E}_{s,a\sim q_h^k}\left[b_h^{k,r}(s,a)\right] &= \gamma\mathbb{E}_{s,a\sim q_h^k}\left[\frac{1-r_h^k(s,a)}{\mu_h^k(s,a)}\right] = \gamma\mathbb{E}_{s\sim q_h^k}\left[\sum_a \pi_h^k(a\mid s)(1-r_h^k(s,a))\right] \\
&= \gamma\mathbb{E}_{s\sim q_h^k}\left[\sum_a \pi_h^k(a\mid s)\frac{\max\{\mu_h^k(a\mid s),\mu_h^{k+d^k}(a\mid s)\}-\mu_h^k(a\mid s)}{\max\{\mu_h^k(a\mid s),\mu_h^{k+d^k}(a\mid s)\}}\right] \\
&\le \gamma\mathbb{E}_{s\sim q_h^k}\left[\sum_a \max\{\mu_h^k(a\mid s),\mu_h^{k+d^k}(a\mid s)\}-\mu_h^k(a\mid s)\right] \\
&\le \gamma\mathbb{E}_{s\sim q_h^k}\left[\left\|\mu_h^{k+d^k}(\cdot\mid s)-\mu_h^k(\cdot\mid s)\right\|_1\right].
\end{aligned}$$

Finally, taking the sum and applying Lemma D.12 completes the proof. □

Lemma D.18. Under the good event, $\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{s,a\sim q_h^k}\left[b_h^{k,v}(s)\right] \le O\left((1+2\varepsilon)H^2nK\right)$.

Proof. For any s and h,

$$\begin{aligned}
\mathbb{E}_{s\sim q_h^k}\left[b_h^{k,v}(s)\right] &= \gamma m^{k+d^k}\mathbb{E}_{s\sim q_h^k}\left[\sum_a \mu_h^{k+d^k}(a\mid s)r_h^k(s,a)\mathbb{1}_h(s,a)\mathbb{1}_{\hat{h},h}^{k,\tau}\right] \\
&\le \gamma m^{k+d^k}\mathbb{E}_{s\sim q_h^k}\left[\sum_a \pi_h^k(a\mid s)\mathbb{1}_h(s,a)\mathbb{1}_{\hat{h},h}^{k,\tau}\right] \\
&= \gamma m^{k+d^k}\mathbb{E}_{s,a\sim q_h^k}\left[\mathbb{1}_h(s,a)\mathbb{1}_{\hat{h},h}^{k,\tau}\right] \le \gamma m^{k+d^k}n(1+2\varepsilon).
\end{aligned}$$

Summing over h and k and noting that $\sum_k m^{k+d^k} \le K$ completes the proof. □

Lemma D.19. Under the good event, $\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{s,a\sim q_h^k}\left[b_h^{k,f}(s)\right] \le O\left(\gamma(1+2\varepsilon)HnK\right)$.

Proof. Similarly to Lemma D.18, $\mathbb{E}_{s\sim q_h^k}\left[b_h^{k,f}(s)\right] \le \gamma n(1+2\varepsilon)$. □

Lemma D.20. Under the good event, $\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{s,a\sim q_h^k}\left[b_h^{k,g}(s,a)\right] \le O\left((1+2\varepsilon)H^2nK\right)$.

Proof. Again, similarly to Lemma D.18, $\mathbb{E}_{s,a\sim q_h^k}\left[b_h^{k,g}(s,a)\right] \le \gamma n(1+2\varepsilon)$. □

E. Auxiliary Lemmas

Lemma E.1 (Value Difference Lemma (Even-Dar et al., 2009)) For any two policies π and π',

$$V_1^{\pi}(s_{\text{init}}) - V_1^{\pi'}(s_{\text{init}}) = \sum_{h=1}^{H} \mathbb{E}_{s \sim q_h^{\pi}} \left[\sum_{a} \pi_h(a \mid s) Q_h^{\pi'}(s,a) - \sum_{a} \pi'_h(a \mid s) Q_h^{\pi'}(s,a) \right].$$

Lemma E.2 (Azuma–Hoeffding inequality) Let $\{X_t\}_{t=1}^{T}$ be a real valued martingale difference sequence adapted to a filtration $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ (i.e., $\mathbb{E}[X_t \mid \mathcal{F}_t] = 0$). If $|X_t| \le R$ a.s. then with probability at least $1-\delta$,

$$\sum_{t=1}^{T} X_t \le R\sqrt{T \ln \frac{1}{\delta}}.$$

Lemma E.3 (A special form of Freedman's Inequality, Theorem 1 of Beygelzimer et al. (2011)) Let $\{X_t\}_{t=1}^{T}$ be a real valued martingale difference sequence adapted to a filtration $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ (i.e., $\mathbb{E}[X_t \mid \mathcal{F}_t] = 0$). If $|X_t| \le R$ a.s. then for any $\delta \in (0, 1/R)$, $T \in \mathbb{N}$ it holds with probability at least $1-\delta$,

$$\sum_{t=1}^{T} X_t \le \sum_{t=1}^{T} \mathbb{E}[X_t^2 \mid \mathcal{F}_t] + \frac{\log(1/\delta)}{\delta}.$$

Lemma E.4 (Consequence of Freedman's Inequality, e.g., Lemma E.2 in (Cohen et al., 2021)) Let $\{X_t\}_{t=1}^{T}$ be a sequence of random variables, supported in $[0, R]$, and adapted to a filtration $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$. For any T, with probability $1-\delta$,

$$\sum_{t=1}^{T} X_t \le 2\sum_{t=1}^{T} \mathbb{E}[X_t \mid \mathcal{F}_t] + 4R \log \frac{1}{\delta}.$$

Lemma E.5 (Lemma A.2 of Luo et al. (2021)) Given a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$, let $z_h^k(s,a) \in [0,R]$ and $q_h^k(s,a) \in [0,1]$ be sequences of $\mathcal{F}_k$-measurable functions. If $Z_h^k(s,a) \in [0,R]$ is a sequence of random variables such that $\mathbb{E}[Z_h^k(s,a) \mid \mathcal{F}_k] = z_h^k(s,a)$ then with probability $1-\delta$,

$$\sum_{k=1}^{K} \sum_{h,s,a} \frac{\mathbb{I}\{s_h^k = s, a_h^k = a\} Z_h^k(s,a)}{q_h^k(s,a) + \gamma} \le \sum_{k=1}^{K} \sum_{h,s,a} \frac{q_h^k(s,a) z_h^k(s,a)}{q_h^k(s,a) + \gamma} + \frac{RH}{2} \ln \frac{H}{\delta}.$$

Lemma E.6 (Lemma 9 of Thune et al. (2019)) Let $\eta > 0$, variables delays $\{d^k\}_{k=1}^{K}$, and loss vectors $\hat{\ell}^k \in [0, 1/\eta)^A$ for all $k \in [K]$. Define,

$$\pi^1(a) = \frac{1}{A}; \qquad \pi^{k+1}(a) = \frac{\pi^k(a) e^{-\eta \sum_{j: j + d^j = k} \hat{\ell}^j(a)}}{\sum_{a' \in A} \pi^k(a') e^{-\eta \sum_{j: j + d^j = k} \hat{\ell}^j(a')}}.$$

Then, for any $\pi \in \Delta_A$:

$$\sum_{k=1}^{K} \sum_{a \in A} (\pi^{k+d^k}(a) - \pi(a)) \hat{\ell}^k(a) \le \frac{\ln A}{\eta} + \sum_{k=1}^{K} \sum_{a \in A} \eta \pi^{k+d^k}(a) \hat{\ell}^k(a)^2.$$

Corollary E.7. Let $\eta > 0$, variables delays $\{d^k\}_{k=1}^{K}$, and loss vectors $\hat{\ell}^k \in [-M, 1/\eta)^A$ for all $k \in [K]$. Define,

$$\pi^1(a) = \frac{1}{A}; \qquad \pi^{k+1}(a) = \frac{\pi^k(a) e^{-\eta \sum_{j: j + d^j = k} \hat{\ell}^j(a)}}{\sum_{a' \in A} \pi^k(a') e^{-\eta \sum_{j: j + d^j = k} \hat{\ell}^j(a')}}.$$

Then, for any $\pi \in \Delta_A$:

$$\sum_{k=1}^{K} \sum_{a \in A} (\pi^{k+d^k}(a) - \pi(a)) \hat{\ell}^k(a) \le \frac{\ln A}{\eta} + 2\sum_{k=1}^{K} \sum_{a \in A} \eta \pi^{k+d^k}(a) \hat{\ell}^k(a)^2 + 2\eta KM^2.$$

Proof. Note that,

$$\pi^1(a) = \frac{1}{A} \quad ; \quad \pi^{k+1}(a) = \frac{\pi^k(a) e^{\eta \sum_{j:j+d^j=k} (\ell^j(a) + M)}}{\sum_{a' \in A} \pi^k(a') e^{\eta \sum_{j:j+d^j=k} (\ell^j(a') + M)}}.$$

The statement now follows immediately by applying Lemma E.6 on the losses $\hat{\ell}^k + M$. □

Lemma E.8. Let $\eta > 0$, variables delays $\{d^k\}_{k=1}^K$, and loss vectors $\ell^k \in [-M, 1)^A$ for all $k \in [K]$. Define,

$$\pi^k(a) = \frac{\exp\left(-\eta \sum_{j:j+d^j<k} \ell^j(a)\right)}{\sum_{a'} \exp\left(-\eta \sum_{j:j+d^j<k} \ell^j(a')\right)};$$

where the empty sum is zero. If $\sum_{j:j+d^j=k} \ell^k(a) > -1$ for all $k \in [K]$, then,

$$\sum_{k=1}^K \sum_a \pi^{k+d^k}(a) \left(\pi^k(a) - \ell^k(a)\right) \le \frac{\ln A}{\eta} + \eta \sum_{k=1}^K \sum_a \pi^{k+d^k}(a) m^{k+d^k} \left(\ell^k(a)\right)^2;$$

where $m^k = |\{j : j + d^j = k\}|$.

Proof. The proof is based in part on the proof of Thune et al. (2019, Lemma 9). Denote $W^k = \sum_a \exp\left(-\eta \sum_{j:j+d^j<k} \ell^k(a)\right)$. We have that

$$\begin{aligned}
\frac{W^{k+1}}{W^k} &= \frac{\sum_a \exp\left(-\eta \sum_{j:j+d^j<k} \ell^j(a)\right) \exp\left(-\eta \sum_{j:j+d^j=k} \ell^j(a)\right)}{W^k} \\
&= \sum_a \pi^k(a) \exp\left(-\eta \sum_{j:j+d^j=k} \ell^j(a)\right) \\
&\le \sum_a \pi^k(a) \left(1 - \eta \sum_{j:j+d^j=k} \ell^j(a) + \eta^2 \left(\sum_{j:j+d^j=k} \ell^j(a)\right)^2\right) \\
&\le \sum_a \pi^k(a) \left(1 - \eta \sum_{j:j+d^j=k} \ell^j(a) + \eta^2 m^k \sum_{j:j+d^j=k} \left(\ell^j(a)\right)^2\right) \\
&= 1 - \eta \sum_{j:j+d^j=k} \sum_a \pi^k(a) \ell^j(a) + \eta^2 \sum_{j:j+d^j=k} \sum_a \pi^k(a) m^k \left(\ell^j(a)\right)^2 \\
&\le \exp\left(-\eta \sum_{j:j+d^j=k} \sum_a \pi^k(a) \ell^j(a) + \eta^2 \sum_{j:j+d^j=k} \sum_a \pi^k(a) m^k \left(\ell^j(a)\right)^2\right);
\end{aligned}$$

where the first inequality is since $e^x \le 1 + x + x^2$ for $x \ge -1$, the second inequality is since $\|x\|_1^2 \le n\|x\|_2^2$ for any $x \in \mathbb{R}^n$, and the last inequality is since $1 + x \le e^x$. Telescoping the ratio above for $k = 1, \ldots, K$ we get,

$$\begin{aligned}
\frac{W^{K+1}}{W^1} &\le \exp\left(-\eta \sum_{k=1}^K \sum_{j:j+d^j=k} \sum_a \pi^k(a) \ell^j(a) + \eta^2 \sum_{k=1}^K \sum_{j:j+d^j=k} \sum_a \pi^k(a) m^k \left(\ell^j(a)\right)^2\right); \\
&\le \exp\left(-\eta \sum_{k=1}^K \sum_a \pi^{k+d^k}(a) \ell^k(a) + \eta^2 \sum_{k=1}^K \sum_a \pi^{k+d^k}(a) m^{k+d^k} \left(\ell^k(a)\right)^2\right);
\end{aligned}$$

where we used that for all j, $j + d^j \leq K$ (we can assume w.l.o.g that all the missing feedback is observed in the end of the interaction). On the other hand,

$$\begin{aligned}
\frac{W^{K+1}}{W^1} &\geq \frac{\sum_a \exp\left(-\eta \sum_{j:j+d^j \quad K} \ell^j(a)\right)}{A} \\
&\geq \frac{\max_a \exp\left(-\eta \sum_{j:j+d^j \quad K} \ell^j(a)\right)}{A} \\
&\geq \frac{\exp\left(-\min_a \eta \sum_{j:j+d^j \quad K} \ell^j(a)\right)}{A} \\
&\geq \frac{\exp\left(-\eta \sum_{j:j+d^j \quad K} \sum_a \pi \ (a)\ell^j(a)\right)}{A} \\
&= \frac{\exp\left(-\eta \sum_{k=1}^K \sum_a \pi \ (a)\ell^k(a)\right)}{A},
\end{aligned}$$

where again we used that for all j, $j + d^j \leq K$ for the last equality. Combining the last two inequalities taking $\ln$ on both sides and rearranging the terms we get,

$$\sum_{k=1}^K \sum_a \left(\pi^{k+d^k}(a) - \pi \ (a)\right) \ell^k(a) \leq \frac{\ln A}{\eta} + \eta \sum_{k=1}^K \sum_a \pi^{k+d^k}(a) m^{k+d^k} \left(\ell^k(a)\right)^2. \qquad \square$$

Lemma E.9 (Lemma 1 of Cesa-Bianchi et al. (2019) adapted to negative losses)**.** *Let $\pi, \tilde{\pi} \in \Delta_A$ and $\ell \in [-M, \infty)^A$ such that*

$$\tilde{\pi}(a) = \frac{\pi(a) e^{\ \eta \ell(a)}}{\sum_{a^0} \pi(a^0) e^{\ \eta \ell(a^0)}}.$$

It holds that,

$$-\eta \pi(a) \left(\ell(a) + M\right) \leq \tilde{\pi}(a) - \pi(a) \leq \eta \tilde{\pi}(a) \sum_{a^0} \pi(a^0) \left(\ell(a^0) + M\right).$$

Proof. By the condition in the lemma,

$$\begin{aligned}
\tilde{\pi}(a) &= \frac{\pi(a) \exp\left(-\eta \ell(a)\right)}{\sum_{a^0} \pi(a^0) \exp\left(-\eta \ell(a^0)\right)} \\
&= \frac{\pi(a) \exp\left(-\eta (\ell(a) + M)\right)}{\sum_{a^0} \pi(a^0) \exp\left(-(\ell(a^0) + M)\right)} \\
&\geq \frac{\pi(a) \exp\left(-\eta (\ell(a) + M)\right)}{\sum_{a^0} \pi(a^0)} \\
&= \pi(a) \exp\left(-\eta (\ell(a) + M)\right) \\
&\geq \pi(a) \left(1 - \eta (\ell(a) + M)\right),
\end{aligned}$$

where in the first inequality we use the fact that $\ell(a^0) + M \geq 0$ and so the exponent at the denominator ≤ 1; and the second inequality is by $e^{\ x} \geq 1 - x$. Thus,

$$\tilde{\pi}(a) - \pi(a) \geq -\eta \pi(a) \left(\ell(a) + M\right).$$

Similarly,

$$\begin{aligned}
\tilde{\pi}(a) &= \frac{\pi(a) \exp\left(-\eta \ell(a)\right)}{\sum_{a^0} \pi(a^0) \exp\left(-\eta \ell(a^0)\right)} \\
&= \frac{\pi(a) \exp\left(-\eta (\ell(a) + M)\right)}{\sum_{a^0} \pi(a^0) \exp\left(-\eta (\ell(a^0) + M)\right)} \\
&\leq \frac{\pi(a)}{\sum_{a^0} \pi(a^0) \exp\left(-\eta (\ell(a^0) + M)\right)},
\end{aligned}$$

where the inequality is since $\ell(a) + M \geq 0$. Thus,

$$\begin{aligned}\tilde{\pi}(a) - \pi(a) &\leq \tilde{\pi}(a)\left(1 - \sum_{a^0}\pi(a^0)\exp\left(-\eta(\ell(a^0)+M)\right)\right)\\ &= \tilde{\pi}(a)\sum_{a^0}\pi(a^0)\left(1-\exp\left(-\eta(\ell(a^0)+M)\right)\right)\\ &\leq \eta\tilde{\pi}(a)\sum_{a^0}\pi(a^0)\left(\ell(a^0)+M\right).\end{aligned}$$

□

Lemma E.10 (Thune et al. (2019))**.** *Let $\{d^k\}_{k=1}^K$ be a sequence of non-negative delays such that $\sum_{k=1}^K d^k = D$. Then,*

$$\sum_{k=1}^{K}\sum_{i=1}^{K}\mathsf{I}\{k \leq i + d^i < k + d^k\} \leq D + K.$$

Proof. The proof appears as part of the proof of Theorem 1 in Thune et al. (2019) or as a separate lemma in Jin et al. (2022, Lemma C.7) □

Lemma E.11. *If A is a positive semi-definite (PSD) matrix and $\gamma > 0$, then for any vector $x \in \mathsf{R}^d$,*

$$x^T(A+\gamma I)^{-1}x \leq \frac{1}{\gamma}\|x\|_2^2 \tag{31}$$

$$x^T(A+\gamma I)^{-2}x \leq \frac{1}{\gamma^2}\|x\|_2^2 \tag{32}$$

$$x^T(A+\gamma I)^{-1}Ax \leq \|x\|_2^2 \tag{33}$$

$$x^T(A+\gamma I)^{-1}A^2(A+\gamma I)^{-1}x \leq \|x\|_2^2 \tag{34}$$

$$x^T(A+\gamma I)^{-1}A(A+\gamma I)^{-1}x \leq x^T(A+\gamma I)^{-1}x \tag{35}$$

Proof. By the spectral decomposition $A = U^T DU$ where U is orthogonal matrix and $D = diag(\lambda_1, ..., \lambda_d)$ with $\lambda_i \geq 0$. Note that,

$$A + \gamma I = U^T(D+\gamma I)U \Longrightarrow (A+\gamma I)^{-1} = U^T(D+\gamma I)^{-1}U.$$

Denote $y = Ux$. Then,

$$x^T(A+\gamma I)^{-1}x = y^T(D+\gamma I)^{-1}y = \sum_{i=1}^{d}\frac{1}{\lambda_i+\gamma}y_i^2 \leq \sum_{i=1}^{d}\frac{1}{\gamma}y_i^2 = \frac{1}{\gamma}\|y\|_2^2 = \frac{1}{\gamma}\|x\|_2^2,$$

which establishes Eq. (31). Eq. (32) is done similarly by noting that $(A+\gamma I)^{-2} = U^T(D+\gamma I)^{-2}U$ since $U^TU = I$. For Eq. (33),

$$(A+\gamma I)^{-1}A = U^T(D+\gamma I)^{-1}DU.$$

Hence,

$$x^T(A+\gamma I)^{-1}Ax = y^T(D+\gamma I)^{-1}Dy = \sum_{i=1}^{d}\frac{\lambda_i}{\lambda_i+\gamma}y_i^2 \leq \sum_{i=1}^{d}y_i^2 = \|y\|_2^2 = \|x\|_2^2.$$

Eq. (34) is done similarly. For Eq. (35),

$$(A+\gamma I)^{-1}A(A+\gamma I)^{-1} = U^T(D+\gamma I)^{-1}D(D+\gamma I)^{-1}U.$$

Thus,

$$\begin{aligned} x^T(A+\gamma I)^{-1}A(A+\gamma I)^{-1}x = y^T(D+\gamma I)^{-1}D(D+\gamma I)^{-1}y &= \sum_{i=1}^{d}\frac{\lambda_i}{(\lambda_i+\gamma)^2}y_i^2 \\ &\le \sum_{i=1}^{d}\frac{1}{\lambda_i+\gamma}y_i^2 = y^T(D+\gamma I)^{-1}y \\ &= x^TU^T(D+\gamma I)^{-1}Ux = x^T(A+\gamma I)^{-1}x. \end{aligned}$$

□

F. DAPPO Implementation Details and Additional Experiments

F.1. DAPPO Implementation Details

Our experiments are based on the implementation of PPO from the Stable-Baselines3 library (Raffin et al., 2021). All of the implementation details remain identical to the original implementation (including the architecture of the Deep Neural Networks and the default hyper-parameters), except for the two following modifications: (i) The objective of DAPPO, and (ii) We mimic learning with delayed feedback by withholding feedback from the algorithm for d steps.

PPO maintains a policy network π^θ and a value network V^ϕ. In each round k, the algorithm collects a rollout $(s_1^k, a_1^k, r_1^k, ..., s_H^k, a_H^k, r_H^k)$ of length $H = 2048$ (notice that for the experiments we switched from costs to rewards). Note that the rollout is of length H, regardless of the number of episodes it takes to fill the rollout buffer to be of that length. That is, if for example the environment has a termination state and the episode ends before time H, we start a new episode and keep filling the buffer until we reach H interactions of the policy with the environment. That way, we can emulate fixed finite horizon MDPs as in our setting, even if the environment is not of fixed horizon. To this end, we treat each rollout of length H as a single episode. Apart from the state, action, reward and next state, the rollout buffer also stores the probability to take the chosen action $\pi^{\theta^k}(a_h^k \mid s_h^k)$ for each $h \in [H]$.

Now, since we want to simulate delayed feedback, we do not use the buffer of round k to update π^{θ^k}. Instead, we store this buffer and load the buffer from round $k-d$, where d is the delay in terms of episodes and not in terms of timesteps. I.e., if the delay in terms of timesteps is $\tilde{d}$, then $d = \lfloor \tilde{d}/H \rfloor$. At this point, the policy network objectives for DPPO and DAPPO are $L_D^k(\theta)$ and $L_{DA}^k(\theta)$, respectively, where,

$$L_D^k(\theta) = \sum_{h=1}^{H} \min\left\{ g_h^k(\theta)\hat{A}_h^{k\ d}, \text{clip}_{1\ \epsilon}\left(g_h^k(\theta)\right)\hat{A}_h^{k\ d} \right\};$$

$$L_{DA}^k(\theta) = \sum_{h=1}^{H} \min\left\{ R_h^k(\theta)\hat{A}_h^{k\ d}, \text{clip}_{1\ \epsilon}\left(R_h^k(\theta)\right)\hat{A}_h^{k\ d} \right\},$$

for $g_h^k(\theta) = \frac{\pi^{\ }(a_h^{k\ d} js_h^{k\ d})}{\pi^{k\ d}(a_h^{k\ d} js_h^{k\ d})}$ and $R_h^k(\theta) = \frac{\pi^{\ }(a_h^{k\ d} js_h^{k\ d})}{\max\{\pi^{k\ d}(a_h^{k\ d} js_h^{k\ d}), \pi^{k\ d}(a_h^{k\ d} js_h^{k\ d})\}}$. $\hat{A}_h^{k\ d} = L_h^{k\ d} - V^\phi(s_h^{k\ d})$ is an estimate of the advantage function, where $L_h^{k\ d}$ is the realized cost-to-go from $(s_h^{k\ d}, a_h^{k\ d})$ until the first termination state in the rollout buffer. Note that $L_D^k(\theta)$ is computed solely based on the parameters θ and on data stored in rollout buffer, and does not require any modification to the the original algorithm. On the other hand, DAPPO computes, in addition, the probabilities of the last policy π^{θ^k} over the trajectory of $\pi^{\theta^{k\ d}}$ (which has a relatively small computational cost).

The value network is trained simply by optimizing the mean-squared error (MSE) loss $\sum_{h=1}^{H}(L_h^{k\ d} - V^\phi(s_h^{k\ d}))^2$. Finally, the optimization of both the policy network and the value network is done using the Adam optimizer (Kingma & Ba, 2014) with learning rate $\eta = 0.0003$, batch size of 64 and for 10 epochs over the rollout buffer (these parameters remain unchanged from the original implementation of Raffin et al. (2021)).

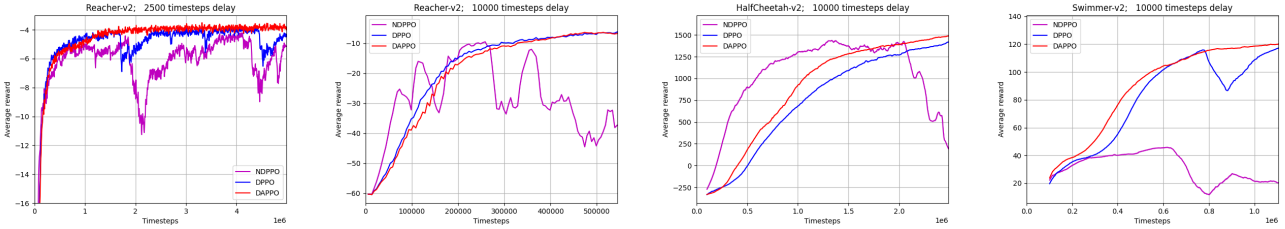

Figure 3. **Instability of NDPPO:** The learning curve of NDPPO vs DAPPO and DPPO under various settings. Plots show reward over a single run. x-axis is the number of timesteps until massive drop in performance (or up to 5M)

.

F.2. Additional Experiments – Instability of NDPPO

We conducted experiments to show that NDPPO is an unstable algorithm. Note that $\pi^{\theta^{k-d}}(a_h^{k-d} \mid s_h^{k-d})$ is not likely to be very small since a_h^{k-d} was sampled from $\pi^{\theta^{k-d}}(\cdot \mid s_h^{k-d})$. On the other hand, $\pi^{\theta^k}(a_h^{k-d} \mid s_h^{k-d})$ may be effectively 0 if the probability to choose a_h^{k-d} has decreased dramatically between time $k-d$ and time k. This emphasizes that NDPPO (described in Section 5) does not only optimize over a biased objective, but is also highly unstable since the gradient of $L_{ND}^k(\theta)$ is inversely proportional to $\pi^{\theta^k}(a_h^{k-d} \mid s_h^{k-d})$. To demonstrate this phenomena, we present in Fig. 3 a few runs of NDPPO (compared to DPPO and DAPPO) under various settings. In some cases, such as in the SWIMMER-V2 environment, NDPPO is not able to improve the policy due to the large bias of its estimator. In other cases, such as the REACHER-V2 or HALFCHEETAH-V2 environments, the learning curve initially behaves similarly to DAPPO and DPPO, and sometimes even slightly better. This is due to the fact that $\pi^{\theta^k}(a_h^{k-d} \mid s_h^{k-d})$ is likely to be smaller then $\pi^{\theta^{k-d}}(a_h^{k-d} \mid s_h^{k-d})$, leading to larger updates (compared to DPPO and DAPPO). However, at some point, the NDPPO's learning curve becomes much noisier due to the dramatic updates whenever $\pi^{\theta^k}(a_h^{k-d} \mid s_h^{k-d}) \approx 0$. This may results in an unrecoverable drop in performance even when the delay is small (as presented in Fig. 3), and in general, it gives an unstable algorithm with huge variance. DAPPO naturally avoids this issue by taking the maximum between $\pi^{\theta^k}(a_h^{k-d} \mid s_h^{k-d})$ and $\pi^{\theta^{k-d}}(a_h^{k-d} \mid s_h^{k-d})$.

F.3. Additional Experiments – Drop in Performance as Delay Length Increases

We conducted experiments to exhibit the drop in performance that occurs when delay length increases. Fig. 4 compares the training curves of DPPO vs DAPPO with different lengths of $\tilde{d} \in \{10000, 25000, 50000, 75000, 100000\}$, alongside the training curve of PPO without delay. As expected, when the delay is relatively small (e.g., $\tilde{d} = 10000$), there is no significant difference between learning with or without delayed feedback. As the delay becomes larger, the performance of all algorithms drops (but at different rates).

One exception is `HalfCheetah-v2` where even for small delay performance drops. However, this is mainly due to the fact that the performance across different seeds is very noisy. Specifically, in 3 out of the 5 seeds, PPO without delays converges to a local maxima which has average reward of ≈ 1500 (similar to DPPO and DAPPO with sufficiently large delay). In the two other seeds it converges to a much better policy, which explains the large std in these graphs.

Whenever the delay becomes sufficiently large ($\sim 25K - 50K$), there is a massive drop in the performance. This emphasize the great challenge that online algorithms need to face in the presence of delays. Namely, the algorithm updates its current policy based on estimated advantage function of a very different policy than the current one. This is also the point where the way we estimate the advantage function becomes important and the difference between DPPO and DAPPO becomes much more significant.