Learning Adversarial Markov Decision Processes with Delayed Feedback

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

[1] Tel Aviv University, Israel
[2] Google Research, Israel
lancewicki@mail.tau.ac.il, avivros007@gmail.com, mansour.yishay@gmail.com

Abstract

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

1 Introduction

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

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

This paper makes a substantial step towards closing the major gap on delayed feedback in the RL literature. We consider the challenging adversarial episodic MDP setting where cost functions change arbitrarily between episodes while the transition function remains stationary over time (but unknown to the agent). We present the *adversarial MDP with delayed feedback* model in which the agent observes feedback for episode k only in the end of episode $k + d^k$, where the delays d^k are unknown and not restricted in any way. This model generalizes standard adversarial MDPs (where $d^k = 0 \ \forall k$), and encompasses great challenges that do not arise in standard RL models, e.g., exploration without feedback and latency in policy updates. Adversarial models are extremely important in practice, as they allow dependencies between costs, unlike stochastic models that assume i.i.d samples. This is especially important in the presence of delays (that are also adversarial in our model), since it allows dependencies between costs and delays which are well motivated in practice (Lancewicki et al. 2021).

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

1.1 Related work

Delays in RL. Although delay is a common challenge RL algorithms need to face in practice (Schuitema et al. 2010; Liu, Wang, and Liu 2014; Changuel, Sayadi, and Kieffer 2012; Mahmood et al. 2018), the theoretical literature on the subject is very limited. Previous work only studied *delayed state observability* (Katsikopoulos and Engelbrecht 2003) where the state is observable in delay and the agent picks actions without full knowledge of its current state. This setting is much related to partially observable MDPs (POMDPs) and motivated by scenarios like robotics system delays. Unfortunately, even planning is computationally hard (exponential in the delay d) for

[*]These authors contributed equally.

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

Table 1: Regret bounds comparison (ignoring constant and poly-logarithmic factors) between our algorithms Delayed OPPO (D-OPPO) and Delayed O-REPS (D-O-REPS), and our lower bound under full-information (full) and bandit feedback (bandit). "Known Transition" assumes dynamics are known to the learner in advance, and "Unknown Transition" means that the learner needs to learn the dynamics. "Delayed Cost" assumes only costs are observed in delay, while in "Delayed Trajectory" the trajectory is also observed in delay, together with the costs.

delayed state observability (Walsh et al. 2009).

This paper studies a different setting that we call *delayed feedback*, where the delay only affects the information available to the agent, and not the execution of its policy. Delayed feedback is also an important setting, as it is experienced in recommendation systems and applications where the policy is executed by a different computational unit than the main algorithm (e.g., policy is executed by a robot with limited computational power, while heavy computations are done by the main algorithm on another computer that receives data from the robot in delay). Importantly, unlike delayed state observability, it is not computationally hard to handle delayed feedback, as we show in this paper. 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.

Delays in multi-armed bandit. Delays were extensively studied in MAB recently as a fundamental issue that arises in many real applications (Vernade, Cappé, and Perchet 2017; Pike-Burke et al. 2018; Cesa-Bianchi, Gentile, and Mansour 2018; Zhou, Xu, and Blanchet 2019; Gael et al. 2020; Lancewicki et al. 2021). Our work is most related to the literature on delays in adversarial MAB, starting with Cesa-Bianchi et al. (2016) that showed the optimal regret for MAB with fixed delay d is of order $\sqrt{(A+d)K}$, where A is the number of actions. Even earlier, variable delays were studied by Quanrud and Khashabi (2015) in online learning with full-information feedback, where they showed optimal $\sqrt{K+D}$ regret. More recently, Thune, Cesa-Bianchi, and Seldin (2019); Bistritz et al. (2019); Zimmert and Seldin (2020); György and Joulani (2020) studied variable delays in MAB, proving optimal $\sqrt{AK+D}$ regret. Unlike MDPs, in MAB there is no underlying dynamics, and the only challenge is feedback about the cost arriving in delay.

Regret minimization in stochastic MDPs. There is a vast literature on regret minimization in RL that mostly builds on the optimism in face of uncertainty principle. Most literature focuses on the tabular setting, where the number of states is small (see, e.g., Jaksch, Ortner, and Auer (2010); Azar, Osband, and Munos (2017); Jin et al. (2018); Zanette and Brunskill (2019)). Recently it was extended to function approximation under various assumptions (see, e.g., Jin et al. (2020b); Yang and Wang (2019); Zanette et al. (2020a,b)).

Adversarial MDPs. Early works on adversarial MDPs (Even-Dar, Kakade, and Mansour 2009; Neu, György, and Szepesvári 2010, 2012; Neu et al. 2014) focused on known transitions and used various reductions to MAB. Zimin and Neu (2013) presented O-REPS – a reduction to online linear optimization achieving optimal regret bounds with known dynamics. Later, O-REPS was extended to unknown dynamics (Rosenberg and Mansour 2019a,b; Jin et al. 2020a) obtaining near-optimal regret bounds. Recently, Cai et al. (2020); Shani et al. (2020); He, Zhou, and Gu (2021) proved similar regret for PO methods (that are widely used in practice).

2 Setting

An episodic adversarial MDP is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, p, \{c^k\}_{k=1}^K)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces of sizes S and A, respectively, H is the episode length, $p = \{p_h : \mathcal{S}\times\mathcal{A} \to \Delta_{\mathcal{S}}\}_{h=1}^H$ is the transition function, and $c^k = \{c_h^k : \mathcal{S}\times\mathcal{A} \to [0,1]\}_{h=1}^H$ is the cost function for episode k. For simplicity, $S \geq \max\{A, H^2\}$.

The interaction between the learner and the environment proceeds as follows. At the beginning of episode k, the learner starts in a fixed initial state[1] $s_1^k = s_{\text{init}} \in S$ and picks a policy $\pi^k = \{\pi_h^k : \mathcal{S} \to \Delta_{\mathcal{A}}\}_{h=1}^H$ where $\pi_h^k(a \mid s)$ gives the probability that the agent takes action a at time h given that the current state is s. Then, the policy is executed in the MDP generating a trajectory $U^k = \{(s_h^k, a_h^k)\}_{h=1}^H$, where $a_h^k \sim \pi_h^k(\cdot|s_h^k)$ and $s_{h+1}^k \sim p_h(\cdot|s_h^k, a_h^k)$. With no delays, the learner observes the feedback in the end of the episode, that is, the trajectory U^k and either the entire cost function c^k under *full-information feedback* or the suffered costs $\{c_h^k(s_h^k, a_h^k)\}_{h=1}^H$ under *bandit feedback*. In contrast, with delayed feedback, these are revealed to the learner only in the end of episode $k + d^k$, where the delays $\{d^k\}_{k=1}^K$ are unknown and chosen by an oblivious adversary before the interaction starts. Denote the total delay by $D = \sum_k d^k$ and the maximal delay by $d_{max} = \max_k d^k$. Note that standard adversarial MDPs are a special case in which $d^k = 0\ \forall k$.

For a given policy π, we define its expected cost with respect to cost function c, when starting from state s at

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

time h, as $V_h^\pi(s) = \mathbb{E}\big[\sum_{h'=h}^{H} c_{h'}(s_{h'}, a_{h'}) | s_h = s, \pi, p\big]$, where the expectation is taken over the randomness of the transition function p and the policy π. This is known as the *value function* of π, and we also define the *Q-function* by $Q_h^\pi(s,a) = \mathbb{E}\big[\sum_{h'=h}^{H} c_{h'}(s_{h'}, a_{h'}) | s_h = s, a_h = a, \pi, p\big]$. It is well-known (see Sutton and Barto (2018)) that the value function and Q-function satisfy the Bellman equations:

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

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

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

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

3 Warm-up: a black-box reduction

One simple way to deal with delays (adopted in several MAB and online optimization works, e.g., Weinberger and Ordentlich (2002); Joulani, Gyorgy, and Szepesvári (2013)) is to simulate a non-delayed algorithm and use its regret guarantees. Specifically, we can maintain $d_{max}+1$ instances of the non-delayed algorithm, running the i-th instance on the episodes k such that $k = i \mod (d_{max}+1)$. That is, at the first $d_{max}+1$ episodes, the learner plays the first policy that each instance outputs. By the end of episode $d_{max}+1$, the feedback for the first episode is observed, allowing the learner to feed it to the first instance. The learner would then play the second output of that instance, and so on. Effectively, each instance plays $K/(d_{max}+1)$ episodes, so we can use the regret of the non-delayed algorithm $\widetilde{\mathcal{R}}_K$ in order to bound $\mathcal{R}_K \leq (d_{max}+1)\widetilde{\mathcal{R}}_{K/(d_{max}+1)}$. Plugging in standard adversarial MDP regret bounds (Rosenberg and Mansour 2019a; Jin et al. 2020a), we obtain the following regret for both full-information and bandit feedback:

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

While simple in concept, the black-box reduction suffers from many evident shortcomings. First, it is highly non-robust to variable delays as its regret scales with the *worst-case delay* Kd_{max} which becomes very large even if the feedback from just one episode is missing. One of the major challenges that we tackle in the rest of the paper is to achieve regret bounds that are independent of d_{max} and scale with the *average delay*, i.e., the total delay D which is usually much smaller than worst-case. Second, even if we ignore the problematic dependence in the worst-case delay, this regret bound is still sub-optimal as it suggests a multiplicative relation between d_{max} and A (and S^2) which does not appear in the MAB setting. Our analysis focuses on eliminating this sub-optimal dependence through a clever algorithmic feature that forces exploration and ensures tight near-optimal regret. Finally, the reduction is highly inefficient as it requires running $d_{max}+1$ different algorithms in parallel. Moreover, the $\sqrt{Kd_{max}}$ regret under bandit feedback is only achievable using O-REPS algorithms that are extremely inefficient to implement in practice. In contrast, our algorithm is based on efficient and practical PO methods. Its running time is independent of the delays and it does not require any prior knowledge or parameter tuning (unlike the reduction needs to know d_{max}). In Section 5 we present experiments showing that our algorithm outperforms generic approaches, such as black-box reduction, not only theoretically but also empirically.

Algorithm 1: Delayed OPPO

Input: $\mathcal{S}, \mathcal{A}, H, \eta > 0, \gamma > 0, \delta > 0$.
Initialization: Set $\pi_h^1(a \mid s) = 1/A$ for every (s,a,h).
for $k = 1, 2, \ldots, K$ **do**
  Play episode k with policy π^k.
  Observe feedback from all episodes $j \in \mathcal{F}^k$.
  Compute cost estimators $\hat{c}^j$ and confidence set $\mathcal{P}^k$.
  # Policy Evaluation
  for $j \in \mathcal{F}^k$ **do**
    Set $V_{H+1}^j(s) = 0$ for every $s \in \mathcal{S}$.
    for $h = H, \ldots, 1$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
      $\hat{p}_h^j(\cdot|s,a) \in \arg\min_{p'_h(\cdot|s,a) \in \mathcal{P}_h^k(s,a)} \langle p'_h(\cdot|s,a), V_{h+1}^j \rangle$.
      $Q_h^j(s,a) = \hat{c}_h^j(s,a) + \langle \hat{p}_h^j(\cdot \mid s,a), V_{h+1}^j \rangle$.
      $V_h^j(s) = \langle Q_h^j(s,\cdot), \pi_h^j(\cdot \mid s) \rangle$.
    end for
  end for
  # Policy Improvement
  $\pi_h^{k+1}(a|s) = \frac{\pi_h^k(a|s) \exp\left(-\eta \sum_{j \in \mathcal{F}^k} Q_h^j(s,a)\right)}{\sum_{a' \in \mathcal{A}} \pi_h^k(a'|s) \exp\left(-\eta \sum_{j \in \mathcal{F}^k} Q_h^j(s,a')\right)}$.
end for

4 Delayed OPPO

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

Delayed OPPO is based on the optimistic proximal policy

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

The main difference that Delayed OPPO introduces is performing updates using all the available feedback at the current time step. Furthermore, in Sections 4.1 and 4.2 we equip our algorithm with novel mechanisms that make it robust to all kinds of variable delays without any prior knowledge and enable us to prove tight regret bounds. Importantly, these mechanisms improve existing results even for the fundamental problem of delayed MAB. Even with these algorithmic mechanisms, proving our regret bounds requires careful analysis and new ideas that do not appear in the MAB with delays literature, as we tackle the much more complex MDP environment.

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

$$|p'_h(s' \mid s,a) - \bar{p}_h^k(s' \mid s,a)| \le \epsilon_h^k(s' \mid s,a),$$

where $\epsilon_h^k(s'|s,a) = \widetilde{\Theta}(\sqrt{\bar{p}_h^k(s'|s,a)/n_h^k(s,a)} + 1/n_h^k(s,a))$ is the confidence set radius. Then, the confidence set for episode k is defined by $\mathcal{P}^k = \{\mathcal{P}_h^k(s,a)\}_{s,a,h}$. Under bandit feedback, the computation of Q^j also requires estimating the cost function c^j in state-action pairs that were not visited in that episode. For building these estimates, we utilize optimistic importance-sampling estimators (Jin et al. 2020a) that first optimistically estimate the probability to visit each state s in each time h of episode j by $u_h^j(s) = \max_{p' \in \mathcal{P}^j} \Pr[s_h = s \mid s_1 = s_{\text{init}}, \pi^j, p']$ and then set the estimator to be $\hat{c}_h^j(s,a) = \frac{c_h^j(s,a)\mathbb{I}\{s_h^j = s, a_h^j = a\}}{u_h^j(s)\pi_h^j(a|s)+\gamma}$ with an exploration parameter $\gamma > 0$.

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

Theorem 1. *Running Delayed OPPO with delayed cost feedback and non-delayed trajectory feedback guarantees, with probability* $1-\delta$*, under full-information feedback:*

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

and under bandit feedback:

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

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

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

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

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

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

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

Notice that the regret bound in Theorem 1 overcomes the major problems that we had with the black-box reduction

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

Theorem 2. *Running Delayed OPPO with known transition function guarantees, with probability $1-\delta$, under full-information feedback: $\mathcal{R}_K = \widetilde{O}(H^2\sqrt{K+D})$, and bandit feedback: $\mathcal{R}_K = \widetilde{O}(HS\sqrt{A}K^{2/3} + H^2D^{2/3} + H^2d_{max})$.*

4.1 Handling delayed trajectories

Previously, we analyzed the Delayed OPPO algorithm in the setting where only cost is observed in delay. In this section, we face the *delayed trajectory feedback* setting in which the trajectory of episode k is observed only in the end of episode $k + d^k$ together with the cost. We emphasize that, while the trajectory from episode k is observed in delay, the policy π^k is executed regularly (see discussion in Section 1.1). Delayed trajectory feedback is a unique challenge in MDPs that does not arise in MAB, as no underlying dynamics exist. Next, we provide the first analysis for delays of this kind and present novel ideas which are crucial for obtaining optimal bounds. Some of the ideas in this section are applicable in other regimes and allow, for example, to enhance the famous UCB algorithm for stochastic MAB to be optimal in the presence of delays (see discussion in Section 5). To convey the main ideas, we focus on full-information feedback (for bandit see Appendix B).

The most natural approach to handle delayed trajectory feedback is simply to update the confidence sets once data becomes available, and then investigate the stochastic process describing the way that the confidence sets shrink over time (with the effect of the delays). With a naive analysis of this approach, we can bound term (A) from the proof of Theorem 1 by $H^2S\sqrt{A(K+D)}$. However, as discussed before, this is far from optimal since the total delay should not scale with the number of states and actions.

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

This is a major improvement especially whenever $d_{max} < \sqrt{D}$, and even if not, the second term can be always substituted for $H^{3/2}\sqrt{SAD}$ using the skipping scheme in Section 4.2. However, we still see the undesirable multiplicative relation with S and A. To tighten the regret bound even further we propose a novel algorithmic mechanism to specifically direct wasted exploration. The mechanism, that we call *explicit exploration*, forces the agent to explore until it observes sufficient amount of feedback. Specifically, until we observe feedback for $2d_{max}\log\frac{HSA}{\delta}$ visits to state s at time h, we pick actions uniformly at random in this state. The explicit exploration mechanism directly improves the bound on term (A) by a factor of A (as shown in the following theorem), and is in fact a necessary mechanism for optimistic algorithms in the presence of delays (see Section 5).

Theorem 3. *Running Delayed OPPO with explicit exploration, with delayed cost feedback and delayed trajectory feedback guarantees, with probability $1-\delta$, under full-information feedback:*

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

and under bandit feedback:

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

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

Term (A) that measures the Q-functions estimation error is controlled by the rate at which the confidence sets shrink. Let $\mathbb{I}_h^k(s,a) = \mathbb{I}\{s_h^k = s, a_h^k = a\}$, we can bound it as follows with standard analysis,

$$(A) \lesssim H\sqrt{S}\sum_{s\in S}\sum_{a\in A}\sum_{h=1}^{H}\sum_{k}\frac{\mathbb{I}_h^k(s,a)}{\sqrt{n_h^k(s,a)}}. \tag{2}$$

Now we address the delays. Fix (s,a,h) and denote the number of unobserved visits by $N_h^k(s,a) = (m_h^k(s,a) - n_h^k(s,a))$. Next, we decouple the statistical estimation error and the effect of the delays in the following way,

$$\begin{aligned}\sum_k \frac{\mathbb{I}_h^k(s,a)}{\sqrt{n_h^k(s,a)}} &= \sum_k \frac{\mathbb{I}_h^k(s,a)}{\sqrt{m_h^k(s,a)}}\sqrt{\frac{m_h^k(s,a)}{n_h^k(s,a)}} \\ &\le \sum_k \frac{\mathbb{I}_h^k(s,a)}{\sqrt{m_h^k(s,a)}}\sqrt{1+\frac{N_h^k(s,a)}{n_h^k(s,a)}} \\ &\le \sum_k \frac{\mathbb{I}_h^k(s,a)}{\sqrt{m_h^k(s,a)}} + \sum_k \frac{\mathbb{I}_h^k(s,a)}{\sqrt{m_h^k(s,a)}}\sqrt{\frac{N_h^k(s,a)}{n_h^k(s,a)}}.\end{aligned} \tag{3}$$

The first term is unaffected by delays and bounded by $H^2S\sqrt{AK}$. For the second term, we utilize explicit exploration in the sense that $n_h^k(s,a) \ge d_{max}/A$. Combine this with the observation that $N_h^k(s,a) \le d_{max}$ (since d_{max} is the maximal delay), to obtain the bound $H^2SA\sqrt{K}$. Finally, to get the tight bound (i.e., eliminate the extra $\sqrt{A}$),

we split the second sum into: (1) episodes with $n_h^k(s,a) \geq d_{max}$ where $N_h^k(s,a)/n_h^k(s,a)$ is tightly bounded by 1 (and not A), and (2) episodes with $n_h^k(s,a) < d_{max}$ in which the regret scales as $\sqrt{d_{max}}$ (which is at most $\sqrt{K}$). □

4.2 Large delays and unknown total delay

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

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

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

To address unknown number of episodes and total delay, we design a new *doubling* scheme. Unlike Bistritz et al. (2019) that end up with a worse bound in delayed MAB due to doubling, our carefully tuned mechanism obtains the same regret bounds (as if K and D were known). Moreover, when applied to MAB, our technique confirms the conjecture of Bistritz et al. (2019) that optimal regret with unknown K and D is achievable using a doubling scheme (due to lack of space, we defer the details to Appendix D). Note that K and D are the only parameters that the algorithm requires, since the skipping scheme replaces the need to know d_{max} with the parameter β (which is tuned using D). The doubling scheme manages the tuning of the algorithm's parameters η, γ, β, making it completely parameter-free and eliminating the need for any prior knowledge regarding the delays.

The doubling scheme maintains an optimistic estimate of D and uses it to tune the algorithm's parameters. Every time the estimate doubles, the algorithm is restarted with the new doubled estimate. This ensures that our optimistic estimate is always relatively close to the true value of D and that the number of restarts is only logarithmic, allowing us to keep the same regret bounds. The optimistic estimate of D is computed as follows. Let M^k be the number of episodes with missing feedback at the end of episode k. Notice that $\sum_{k=1}^K M^k \leq D$ because the feedback from episode j was missing in exactly d^j episodes. Thus, at the end of episode k our optimistic estimate is $\sum_{j=1}^k M^j$. So for every episode j with observed feedback, its delay is estimated by exactly d^j, and if its feedback was not observed, then we estimate it as if feedback will be observed in the next episode.

In Appendix D we give the full pseudo-code of Delayed OPPO when combined with doubling, and formally prove that our regret bounds are not damaged by doubling.

5 Additional results and empirical evaluation

Lower bound. For episodic stochastic MDPs, the optimal minimax regret bound is $\widetilde{\Theta}(H^{3/2}\sqrt{SAK})$ (Azar, Osband, and Munos 2017; Jin et al. 2018). As adversarial MDPs generalize the stochastic MDP model, this lower bound also applies to our setting. The lower bound for multi-arm bandits with delays is based on a simple reduction to non-delayed MAB with full-information feedback. Namely, we can construct a non-delayed algorithm for full-information feedback using an algorithm $\mathcal{A}$ for fixed delay d by simply feeding $\mathcal{A}$ with the same cost function for d consecutive rounds. Using the known lower bound for full-information MAB, this yields a $\Omega(\sqrt{dK}) = \Omega(\sqrt{D})$ lower bound which easily translates to a $\Omega(H\sqrt{D})$ lower bound in adversarial MDPs. Combining these two bounds gives a lower bound of $\Omega(H^{3/2}\sqrt{SAK} + H\sqrt{D})$ for all settings, except for full-information feedback with known dynamics where the lower bound is $\Omega(H\sqrt{K+D})$. In light of this lower bound, we discuss the regret of Delayed OPPO and open problems.

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

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

Another important note: even with delayed trajectory feedback, our $\sqrt{D}$ term is still optimal for a wide class of delays – *monotonic delays*. That is, if the sequence of delays is monotonic, i.e., $d^j \leq d^k$ for $j < k$, then the $\sqrt{D}$ term of our regret bound for delayed trajectory feedback is not multiplied by $\sqrt{S}$. This follows because in this case term (A) that handles estimation error of p can be analysed with respect to the *actual* number of visitation, since by the time

we estimate Q^k at the end of episode $k + d^k$ we already have all the feedback for $j < k$. Monotonic delays include the fundamental setting of a fixed delay d.

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

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

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

Stochastic MDP with delayed feedback. Most of the RL literature has focused on stochastic MDPs – a special case of adversarial MDPs where cost $c_h^k(s,a)$ of episode k is sampled i.i.d from a fixed distribution $C_h(s,a)$. Thus, studying the effects of delayed feedback on stochastic MDPs is a natural question. With stochastic costs, OPPO obtains $\sqrt{K}$ regret even under bandit feedback, since we can replace importance-sampling estimators with an empirical average.

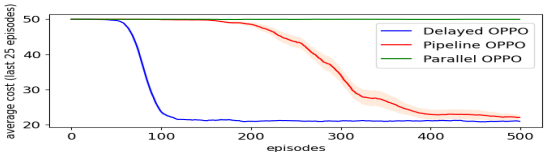

Figure 1: Average cost of delayed algorithms in grid world with geometrically distributed delays.

This means that with stochastic costs and bandit feedback, our Delayed OPPO algorithm obtains the same near-optimal regret bounds as under full-information feedback. However, the $\sqrt{D}$ lower bound heavily relies on adversarial costs, as it uses a sequence of costs that change every d episodes, suggesting that $\sqrt{D}$ dependence might not be necessary.

Indeed, for stochastic cost, delayed versions of optimistic algorithms (e.g., Zanette and Brunskill (2018)) have regret scaling as the estimation error (term (A) in Eq. (2)), which means that our analysis (Section 4.1) proves regret that does not scale with $\sqrt{D}$ but only with H^2SAd_{max}. Again, this can be improved to H^2Sd_{max} using explicit exploration.

Theorem 5. *Running an optimistic algorithm with explicit exploration, with delayed bandit cost feedback and delayed trajectory feedback guarantees, with probability* $1-\delta$*, regret bound of* $\widetilde{O}(H^2S\sqrt{AK} + H^2Sd_{max})$ *in stochastic MDPs.*

This contribution is important, even for the rich literature on delayed stochastic MAB. Lancewicki et al. (2021) show that the (optimistic) UCB algorithm may suffer sub-optimal regret of Ad_{max}. Furthermore, they were able to remove the A factor by an action-elimination algorithm which explores active arms equally. Since optimism is currently the only approach for handling unknown transitions in adversarial MDPs, it was crucial for us to find a novel solution to handle delays in optimistic algorithms. Theorem 5 shows that optimistic algorithms (like UCB) can indeed be "fixed" to handle delays optimally, using explicit exploration.

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

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

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grant number 993/17), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University.

References

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

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

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

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

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

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

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

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

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

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

György, A.; and Joulani, P. 2020. Adapting to Delays and Data in Adversarial Multi-Armed Bandits. *arXiv preprint arXiv:2010.06022*.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *International Conference on Machine Learning*, 1861–1870.

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

He, J.; Zhou, D.; and Gu, Q. 2021. Nearly Optimal Regret for Learning Adversarial MDPs with Linear Function Approximation. *arXiv preprint arXiv:2102.08940*.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(4).

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

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

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

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

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

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

Lancewicki, T.; Segal, S.; Koren, T.; and Mansour, Y. 2021. 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*, 5969–5978. PMLR.

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

Mahmood, A. R.; Korenkevych, D.; Komer, B. J.; and Bergstra, J. 2018. 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.

Maurer, A.; and Pontil, M. 2009. Empirical Bernstein Bounds and Sample Variance Penalization. *stat*, 1050: 21.

Neu, G.; György, A.; and Szepesvári, C. 2010. The Online Loop-free Stochastic Shortest-Path Problem. In *Conference on Learning Theory (COLT)*, 231–243.

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

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

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

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

Rosenberg, A.; Cohen, A.; Mansour, Y.; and Kaplan, H. 2020. Near-optimal Regret Bounds for Stochastic Shortest Path. In *International Conference on Machine Learning*, 8210–8219. PMLR.

Rosenberg, A.; and Mansour, Y. 2019a. Online Convex Optimization in Adversarial Markov Decision Processes. In *International Conference on Machine Learning*, 5478–5486.

Rosenberg, A.; and Mansour, Y. 2019b. Online Stochastic Shortest Path with Bandit Feedback and Unknown Transition Function. In *Advances in Neural Information Processing Systems*, 2209–2218.

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

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

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

Shani, L.; Efroni, Y.; Rosenberg, A.; and Mannor, S. 2020. Optimistic Policy Optimization with Bandit Feedback. In *International Conference on Machine Learning*, 8604–8613. PMLR.

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

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

Vernade, C.; Cappé, O.; and Perchet, V. 2017. Stochastic Bandit Models for Delayed Conversions. In *Conference on Uncertainty in Artificial Intelligence*.

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

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

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

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

Zanette, A.; and Brunskill, E. 2018. Problem dependent reinforcement learning bounds which can identify bandit structure in mdps. In *International Conference on Machine Learning*, 5747–5755.

Zanette, A.; and Brunskill, E. 2019. Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds. In *International Conference on Machine Learning*, 7304–7312.

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

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

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

Zimin, A.; and Neu, G. 2013. 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*, 1583–1591.

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