

# Blind Decision Making: Reinforcement Learning with Delayed Observations

Mridul Agarwal, Vaneet Aggarwal

## Abstract

Reinforcement learning typically assumes that the state update from the previous actions happens instantaneously, and thus can be used for making future decisions. However, this may not always be true. When the state update is not available, the decision taken is partly in the blind since it cannot rely on the current state information. This paper proposes an approach, where the delay in the knowledge of the state can be used, and the decisions are made based on the available information which may not include the current state information. One approach could be to include the actions after the last-known state as a part of the state information, however, that leads to an increased state-space making the problem complex and slower in convergence. The proposed algorithm gives an alternate approach where the state space is not enlarged, as compared to the case when there is no delay in the state update. Evaluations on the basic RL environments further illustrate the improved performance of the proposed algorithm.

## I. INTRODUCTION AND RELATED WORK

A reinforcement learning (RL) agent models the world into states, which come from the set  $\mathcal{S}$ . The agent, at time  $t$  and in state  $s_t$ , chooses an action  $a_t \in \mathcal{A}$ . After choosing an action, the agent receives a reward  $r_t$  and moves to state  $s_{t+1}$  at the next time step  $t + 1$ . The goal of the agent is to find a policy to select an action that maximizes the total cumulative reward the agent collects over  $T$  time steps, where  $T$  may be unbounded [16]. Applications of RL are continuously increasing in domains that can be formulated using state, action, and rewards. Many such applications include cloud scheduling [2], robot manipulation [5], financial trading [6], and microgrid management [11]. However, many of these works assume that the state update is immediately observed after the action is played, which may not be true in general applications. This paper proposes an algorithm and evaluates it when the state updates are not immediately available.

We note that in many applications, the state update information is not available instantaneously. As an example, micro-grid control may have stochastic delays because of the communication link and these delays may have adverse impact on the system [12]. The authors of [4] studied the impact of delays in video scheduling for mobile devices. For a mobile device, channel state information might be delayed since the streaming decisions are made many seconds before the content is played (to reduce the rebuffering events), which impacts the video streaming algorithms using RL based control. 360-degree video streaming further adds a challenge by introducing different qualities in different tiles in a frame, and thus the head movement prediction can be used for efficient streaming [7]. However, the head position while viewing a frame is not available for the streaming algorithm, which makes the decision before the content is played and is thus making decisions without the current state information. Similarly, cloud-based scheduling will observe stragglers only after a certain time threshold, and network delays may cause financial losses in trading. Delays can also arise from processing delays on input images or computationally complex data processing. The issue of such delay in the availability of state information limits the use of RL for practical applications [13].

As studied by [10], delays may be of three types, 1) observation delays, where observations for state updates are delayed, 2) action delays, where actions may take effect after a certain delay, and 3) cost/reward delays, where rewards are not obtained till certain time after an action is taken in any state. They show that action delays are equivalent to observation delays. If observations are delayed then the agent plays action with knowledge of the last observed state and the actions happened in the meantime. Similarly, if actions are delayed, the agent schedules actions for the future with the same information. In case of delays in rewards updates, algorithms train using mini-batches where a mini-batch consists of state observations, actions taken, and available rewards received for some duration. We note that even though the current state information is not known, the decisions still have to be made. In this paper, we assume only observation or action delays. We further assume that rewards are provided to the agent along with state updates.

The authors of [1] showed that for Markov Decision Process (MDP) where each observation is delayed by  $d$  steps, an equivalent MDP can be constructed with an augmented state space where previous  $d$  actions are appended to the currently known state. Thus, the new state space  $\mathcal{S}' = \mathcal{S} \times \mathcal{S}^d$  with  $s'_t = (s_{t-d}, a_{t-d}, a_{t-(d-1)}, \dots, a_{t-1})$ . This formulation restores the problem structure back to an MDP, and any RL algorithm can now be applied to the new MDP. We point to three limitations of this approach.

The first limitation is that this approach does not scale to stochastic delays. In order to resolve this,

[10] proposed a new solution by assuming that the maximum delay is bounded. They assume that if the delays are more than a threshold  $n$ , then the algorithm freezes, and would not take any action. However, many real-time systems might not enjoy this liberty of freezing the execution of the algorithm.

The second limitation is that the expansion of the state space might not be efficient for implementation because of the increased storage complexity and exponentially larger time to converge to the optimal solution. In order to alleviate this, the authors of [18] proposed an algorithm to play action  $a_t$  which is optimal for the state in which the agent is most likely in, or  $a_t = \arg \max_{a_t} Q(\arg \max_{s_t} p_{s_t}(s_t|s'_t), a_t)$ . It was assumed that the probability of not being in the most likely state is bounded by  $\epsilon$ , where  $\epsilon$  is small enough. Note that for Markov Decision Process (MDP) where state distributions are not concentrated, this assumption might not hold.

The third limitation for augmenting the state space is that the regret analysis of the MDP shows that the regret of this algorithm scales as  $|\mathcal{A}|^{d/2}$ , where  $d$  is the delay in the availability of the state information [8]. For large  $d$ , the gap may be large enough for the approach to have significantly decreased performance.

Having pointed out the limitations in the prior works, we propose a solution that aims to alleviate these issues. The proposed algorithm, called *Expectation-Maximization Q-learning* (EMQL), takes an action that maximizes the expected gain of true MDP across all possible states conditioned over the last known state  $s_{t-d}$ , and actions taken till time  $t$ . EMQL for delayed reinforcement learning has the following properties:

- **Space efficient:** The proposed algorithm does not use an augmented MDP to determine the action. It, however, uses the fact the current true state comes from the probability distribution generated by augmented MDP.
- **Robust under deviation from most likely state:** The algorithm selects the action that maximizes the expected value  $Q$  function. Thus, even though the distribution is not concentrated around a single state, the distribution is efficiently utilized.
- **Handles stochastic delays:** The algorithm works well with stochastic delays, as well as missed information.

EMQL is evaluated on Frozen Lake ( $8 \times 8$  grid), and Cart Pole environments of OpenAI Gym platform [3]. The results for delayed settings are compared with Extended MDP formulation of [1], MBS algorithm given by [18], and dQ algorithm proposed by [15] respectively. The metric of comparison is the total reward, collected in each episode.

The rest of the paper is organized as follows. Section II defines the problem under consideration. Section

IV presents the proposed algorithm. Evaluation and comparison results are presented in section V. Section VI concludes the paper with a brief discussion.

## II. FORMULATION

We consider a Markov Decision Process  $\mathcal{M}$ , with set of states denoted by  $\mathcal{S}$ , and set of actions denoted by  $\mathcal{A}$ . At time  $t$ , the environment is in state  $s_t \in \mathcal{S}$ . The definitions are mostly consistent with those in [16], [14]

At any time  $t$ , the agent chooses action  $a_t \in \mathcal{A}$  based on its knowledge about the current state. On playing the action  $a_t$ , environment rewards the agent with  $R_t$ , which is random variable conditioned on environment state  $s_t$ , and action chosen by agent at time  $t$ . The maximum reward the agent can receive at any time step is  $R_{max}$ . The goal of the agent is to maximize the discounted cumulative rewards it receives. The discount factor  $\gamma \in [0, 1)$  denotes the importance of future rewards.

$$\mathcal{R} = \sum_{t=0}^{\infty} \gamma^t R_t \quad (1)$$

Expected reward when action  $a$  is taken in a state  $s$  is defined as  $r(s, a)$

$$r(s, a) = \mathbb{E} [R_t | s_t = s, a_t = a] \quad (2)$$

The probability distribution of next state  $s_{t+1}$  conditioned on current state  $s_t$  and action  $a_t$  is denoted by  $p(s_t, a_t, s_{t+1})$ . Shorthand notation by dropping the subscripts is denoted as,

$$p(s, a, s') = P(s_{t+1} = s' | s_t = s, a_t = a) \quad (3)$$

Agent uses a policy  $\pi$  to select an action.  $\pi$  is defined as the probability distribution over actions given the state.

$$\pi(a|s) = \mathbb{P}(a_t = a | s_t = s) \quad (4)$$

The value function  $V^\pi(s)$  of a state  $s$  is defined as the expected value of sum of discounted rewards which agent can receive over time starting from state  $s$  and choosing actions according to the policy  $\pi$ .

$$V^\pi(s) = \mathbb{E}_\pi \left[ \sum_{t=t_0}^{\infty} \gamma^{t-t_0} R_t \middle| s_{t_0} = s \right] \quad (5)$$

This makes the maximum possible value of  $V^\pi(s)$  as  $\frac{R_{max}}{1-\gamma}$ . Similarly action-value function  $Q^\pi(s, a)$  is defined as the expected cumulative rewards which agent receives in state  $s$  on taking action  $a$  and then following policy  $\pi$ ,

$$Q^\pi(s) = \mathbb{E}_\pi \left[ \sum_{t=t_0}^{\infty} \gamma^{t-t_0} R_t \mid s_{t_0} = s, a_{t_0} = a \right] \quad (6)$$

For both value function  $V^\pi(s)$  and action-value function  $Q^\pi(s, a)$  the expectation is taken over the states which are distributed according to the transition dynamics of the MDP  $\mathcal{M}$  and actions which are distributed according to the policy  $\pi$ . We use only  $\pi$  in the subscript for expectation as we can only control the policy. Optimal policy  $\pi^*$  is defined as the policy which maximizes the value function for all states.

$$V^*(s) = V^{\pi^*}(s) \quad (7)$$

$$= \sup_{\pi} V^\pi(s) \quad \forall s \in \mathcal{S} \quad (8)$$

The optimal policy  $\pi^*$  gives an optimal value function  $V^*$  and an optimal Q-function  $Q^*$ , which are related as

$$Q^*(s, a) = \mathbb{E}_\pi [r(s_t, a_t)] + \gamma \mathbb{E} [V^*(s_{t+1}) \mid s_t = s, a_t = a] \quad (9)$$

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a) \quad \forall s \in \mathcal{S} \quad (10)$$

In practical RL algorithms, agent deploys some strategy to calculate an estimate of Q-function  $\hat{Q}$ . Some common strategies are discussed in [16]. The agent selects the action greedily based on the estimate of Q-function  $\hat{Q}$  as

$$a_t = \arg \max_{a_t \in \mathcal{A}} \hat{Q}(s_t, a_t) \quad (11)$$

The delay  $\Delta$  is a random variable denoting the delay of the system. All realizations of delay  $d$  are assumed to be a non-negative integer. At time  $t$ , the last known state for the agent is  $s_{t-d}$ . The actions played in  $d$  time steps are  $a_{t-d}, a_{t-(d-1)}, \dots, a_{t-1}$ .

We assume that at the beginning of any episode, all the delayed observations of previous episodes are available. This also means that the observations from any of the previous episodes are not corrupting the observations received in current episode.

### III. PROPOSED POLICY AND BOUNDS

For a system with  $d$  delays, we construct an extended MDP  $\widetilde{\mathcal{M}}$  which has state  $\tilde{s}_t$  as  $(s_{t-d}, a_{t-d}, a_{t-(d-1)}, \dots, a_{t-1})$ . The two MDPs  $\widetilde{\mathcal{M}}$ , and  $\mathcal{M}$  share the same action space  $\mathcal{A}$ , so we will not change the notation for actions. For everything else we will put a tilde over the variables for the augmented MDP. The policy  $\tilde{\pi}$  now selects an action based on  $\tilde{s}$  or the tuple  $(s_{t-d}, a_{t-d}, a_{t-(d-1)}, \dots, a_{t-1})$ . The corresponding Q-function for a policy  $\tilde{\pi}$  over  $\widetilde{\mathcal{M}}$  becomes,

$$\begin{aligned} \widetilde{Q}^{\tilde{\pi}}(\tilde{s}, a) &= \mathbb{E}_{\tilde{\pi}} [R_t | \tilde{s}_t = \tilde{s}, a_t = a] + \\ &\quad \gamma \sum_{\tilde{s}_{t+1} \in \widetilde{\mathcal{S}}} \widetilde{V}^{\tilde{\pi}}(\tilde{s}_{t+1}) \mathbb{P}[\tilde{s}_{t+1} | \tilde{s}_t = \tilde{s}, a_t = a] \end{aligned} \quad (12)$$

Using this construction we present the key lemma based on which we construct our policy.

*Lemma 1:* Expected reward obtained by agent in augmented state  $\tilde{s}_t$  by taking an action  $a$ , is related to the true state  $s_t$  of environment as

$$\tilde{r}(\tilde{s}_t, a) = \sum_{s \in \mathcal{S}} r(s, a) p(s | \tilde{s}_t) \quad (13)$$

*Proof 1:* Reward  $R_t$  generated by the environment is oblivious to the state maintained by the agent. The expected reward for the agent is  $\tilde{r}(\tilde{s}_t, a)$

$$\tilde{r}(\tilde{s}, a) = \mathbb{E} [R_t | \tilde{s}_t = \tilde{s}, a_t = a] \quad (14)$$

$$\begin{aligned} &= \sum_{s \in \mathcal{S}} \mathbb{E} [R_t | s_t = s, \tilde{s}_t = \tilde{s}, a_t = a] \mathbb{P}(s_t = s | \tilde{s}_t = \tilde{s}) \\ &= \sum_{s \in \mathcal{S}} \mathbb{E} [R_t | s_t = s, a_t = a] \mathbb{P}(s_t = s | \tilde{s}_t = \tilde{s}) \end{aligned} \quad (15)$$

Equality (15) follows from the fact that reward does not depend on the state maintained by agent, but on the environment's state.

Lemma 1 states that the expected reward received on taking action  $a$  in state  $\tilde{s}$  is the expected reward received by taking action  $a$  in the unobserved state conditioned on  $\tilde{s}$ . Based on Lemma 1, a myopic policy, which maximizes immediate expected return for the agent, selects greedy action  $a_t(\tilde{s})$  as

$$a_t = \arg \max_{a \in \mathcal{A}} (\mathbb{E}_s [R(s, a) | \tilde{s}_t]) \quad (16)$$

Inspired by the myopic policy, we now propose a policy for working with delayed state updates.

The agent would be able to maximize its expected discounted cumulative rewards if it has oracle access

which could return the optimal action for unobserved current state  $s_t$  of the environment. However, since such oracle access is not available, we settle for a policy that assumes that the state at the next time step will be available. This policy maximizes the sum of immediate reward and the expected value of the next state under the optimal policy for  $\mathcal{M}$ , or

$$\tilde{\pi}(a_t|\tilde{s}_t) = \begin{cases} 1, & \text{if } a_t = \arg \max_{a \in \mathcal{A}} \mathbb{E}[Q^*(s, a)|\tilde{s}_t], \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

The following theorem provides bounds on minimum value an augmented state  $\tilde{s} \in \tilde{\mathcal{S}}$  would fetch for the agent. That is, value function using the policy defined in Equation (17) ensures the minimum value given in Theorem 1.

*Theorem 1:* If the agent follows policy as given in (17), for an augmented MDP  $\tilde{\mathcal{M}}$ , then the value of each state  $\tilde{s} \in \tilde{\mathcal{S}}$  satisfies the following lower bound,

$$\tilde{V}^{\tilde{\pi}}(\tilde{s}) \geq \mathbb{E}_{s|\tilde{s}} [V^{\pi^*}(s)] - \frac{R_{max}}{(1-\gamma)^2} \left(1 - \frac{1}{|\mathcal{A}|}\right) \quad (18)$$

where  $\pi^*(a_t|s_t)$  is oracle aided policy which gives the optimal action for true MDP  $\mathcal{M}$ .

*Proof 2:* We first mention and prove the lemmas required for the proof, and then continue to the final proof. Lemma 2 relates the value function for  $\mathcal{M}$  and  $\tilde{\mathcal{M}}$  under same policy  $\tilde{\pi}$ . Lemma 3 relates expected value using the oracle aided policy and the state-action value using oracle aided policy.

*Lemma 2:* Value function of a policy  $\tilde{\pi}(a|\tilde{s})$  for augmented MDP  $\tilde{\mathcal{M}}$  is related to value function for same policy under true MDP  $\mathcal{M}$  is related as

$$\tilde{V}^{\tilde{\pi}}(\tilde{s}) = \mathbb{E}_s [V^{\tilde{\pi}}(s)|\tilde{s}] \quad (19)$$

where  $V^{\tilde{\pi}}(s)$  is value function for policy  $\tilde{\pi}$  with MDP  $\mathcal{M}$ .

*Proof 3:* Note that the value function for the augmented MDP  $\tilde{\mathcal{M}}$  is expected cumulative discounted rewards collected by the agent when it starts from state  $\tilde{s}_t$ , and follows a policy  $\tilde{\pi}$ . Thus from the definition of  $\tilde{V}^{\tilde{\pi}}(\tilde{s})$ , we have

$$\tilde{V}^{\tilde{\pi}}(\tilde{s}) = \mathbb{E}_{\tilde{\pi}} \left[ \sum_{k=0}^{\infty} \gamma^{t+k} R_{t+k} | \tilde{s}_t = \tilde{s} \right] \quad (20)$$

$$= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}_{\tilde{\pi}} [R_{t+k} | \tilde{s}_t = \tilde{s}] \quad (21)$$

$$= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}_{\tilde{\pi}} [R_{t+k} | s_{t+k}, a_{t+k}] \mathbb{P} [s_{t+k} | \tilde{s}_t, \tilde{\pi}] \quad (22)$$

$$= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}_{\tilde{\pi}} [R_{t+k} | s_{t+k}, a_{t+k}] \sum_{s_t \in \mathcal{S}} \mathbb{P} [s_{t+k} | s_t, \tilde{\pi}] \mathbb{P} [s_t | \tilde{s}_t] \quad (23)$$

$$= \sum_{s_t \in \mathcal{S}} \left( \sum_{k=0}^{\infty} \gamma^k \mathbb{E}_{\tilde{\pi}} [R_{t+k} | s_{t+k}, a_{t+k}] \mathbb{P} [s_{t+k} | s_t, \tilde{\pi}] \right) \mathbb{P} [s_t | \tilde{s}_t]$$

$$= \sum_{s_t \in \mathcal{S}} V^{\tilde{\pi}}(s) \mathbb{P} [s_t | \tilde{s}_t] \quad (24)$$

Equation (22) follows from Lemma 1. Equation (23) follows the fact that if the agent knew the true state and followed the policy  $\tilde{\pi}$ , then the environment state evolution probabilities would remain the same.

*Lemma 3:* For the optimal policy  $\pi^*$ , and distribution over initial state  $s$ ,  $\mu$ ,

$$\arg \max_a \mathbb{E}_{s \sim \mu} [Q^*(s, a)] \geq \frac{1}{|\mathcal{A}|} \mathbb{E}_{s \sim \mu} [V^*(s)] \quad (25)$$

*Proof 4:* Let initial state follow some distribution  $\mu$ , or  $\mathbb{P} [s_t = s] \sim \mu$ . Then we have,

$$\mathbb{E}_{s \sim \mu} [V^*(s)] = \mathbb{E}_{s \sim \mu} \left[ \max_a Q^*(s, a) \right] \quad (26)$$

$$\leq \mathbb{E}_{s \sim \mu} \left[ \sum_{a \in \mathcal{A}} Q^*(s, a) \right] \quad (27)$$

$$= \sum_{a \in \mathcal{A}} \mathbb{E}_{s \sim \mu} [Q^*(s, a)] \quad (28)$$

$$\leq |\mathcal{A}| \max_{a \in \mathcal{A}} \mathbb{E}_{s \sim \mu} [Q^*(s, a)] \quad (29)$$

We can now use Lemma 2, Lemma 6.1 from [9] and Lemma 3 to find the minimum value of  $\tilde{V}^{\tilde{\pi}}(\tilde{s}) - \mathbb{E}_s [V^{\pi^*}(s) | \tilde{s}]$ . Mathematically we have,

$$\tilde{V}^{\tilde{\pi}}(\tilde{s}) - \mathbb{E}_s [V^{\pi^*}(s) | \tilde{s}] = \mathbb{E}_s [V^{\tilde{\pi}}(s) | \tilde{s}] - \mathbb{E}_s [V^{\pi^*}(s) | \tilde{s}] \quad (30)$$

$$= \mathbb{E}_{s | \tilde{s}, s_{t+1}, a_t \sim \tilde{\pi}} \left[ \sum_{k=0}^{\infty} \gamma^k Q^{\pi^*}(s_{t+k}, a_{t+k}) - V^{\pi^*}(s_{t+k}) \right] \quad (31)$$

$$= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}_{s | \tilde{s}, s_{t+1}, a_t \sim \tilde{\pi}} [Q^{\pi^*}(s_{t+k}, a_{t+k})]$$

$$- \mathbb{E}_{s | \tilde{s}, s_{t+1}, a_t \sim \tilde{\pi}} [V^{\pi^*}(s_{t+k})] \quad (32)$$

$$\geq \sum_{k=0}^{\infty} \gamma^k \frac{1}{|\mathcal{A}|} \mathbb{E}_{s | \tilde{s}, s_{t+1}, a_t \sim \tilde{\pi}} [V^{\pi^*}(s_{t+k})]$$

$$- \mathbb{E}_{s | \tilde{s}, s_{t+1}, a_t \sim \tilde{\pi}} [V^{\pi^*}(s_{t+k})] \quad (33)$$

$$\geq \sum_{k=0}^{\infty} \gamma^k \left( \frac{1}{|\mathcal{A}|} - 1 \right) \mathbb{E}_{s|\bar{s}, s_{t+1}, a_t \sim \tilde{\pi}} [V^{\pi^*}(s_{t+k})] \quad (34)$$

$$\geq \sum_{k=0}^{\infty} \gamma^k \left( \frac{1}{|\mathcal{A}|} - 1 \right) \frac{R_{max}}{1-\gamma} = \left( \frac{1}{|\mathcal{A}|} - 1 \right) \frac{R_{max}}{(1-\gamma)^2} \quad (35)$$

Equation (30) follows from Lemma 2, and Equation (31) comes from using Lemma 6.1 from [9]. Equation (33) comes from the fact that  $a_{t+k}$  is chosen from the policy defined in Equation (17) for all  $k \geq 0$ , and Lemma 3. Equation (35) follows from maximum possible value of  $V^{\pi}(s)$ .

Theorem 1 states that the proposed policy can suffer a maximum degradation of  $\left( \frac{1}{|\mathcal{A}|} - 1 \right) \frac{R_{max}}{(1-\gamma)^2}$  only from the expected optimal value of unobserved state conditioned on the extended state.

Now, the task that remains is to find  $Q^*$  for true MDP. We assume that the delayed state observations can be identified using timestamp or index header. This is a common engineering principle in communication networks to deal with asynchronous packets [17], and hence it is a valid assumption. This allows to find optimal Q-function for true MDP  $\mathcal{M}$ . We next provide a detailed algorithm for the policy described in this section.

#### IV. ALGORITHM

We now utilize Equation (17) to construct *Expectation Maximization Q Learning* (EMQL) algorithm (described in Algorithm 1) which is space efficient and which can handle stochastic delays. We note that to calculate the expected Q-value function, we require an algorithm to calculate and store the Q values for each state-action pair. Also the algorithm requires to estimate and store the state transition probabilities to calculate the expected Q value for Equation (17). For this, we divide our algorithm into two parts. First part (Algorithm 1) keeps a track of visited state-action pairs, and observed next states and rewards to calculate the Q-value table and the probability transition matrix. Second part is an auxiliary algorithm (Algorithm 2) which actually implements the policy of Equation (17) using the Q-value table and probability transition matrix of the first part.

##### A. Algorithm Construction

The algorithm takes the state space  $\mathcal{S}$ , action space  $\mathcal{A}$ , discount factor  $\gamma$ , and exploration factor  $\epsilon_t$  as inputs. Since, we consider a model based algorithm, we maintain variables corresponding to number of times a state-action pair was visited ( $N(\cdot, \cdot)$ ), rewards obtained for the state-action pair ( $R(\cdot, \cdot)$ ), and the counter for next state from a state-action pair to calculate the estimates of probability transitions ( $P(\cdot, \cdot, \cdot)$ ).

Before beginning any episode the algorithm calculates the estimated probability transitions ( $\hat{p}(\cdot, \cdot, \cdot)$ ) and the expected rewards ( $\hat{r}(\cdot, \cdot)$ ) according the equations (38) and (39) respectively. It then updates the Q-function ( $Q(\cdot, \cdot)$ ) for the true MDP as per the equation (40). The Q-function learned is for  $\mathcal{M}$ , which converges faster because of the smaller state space.

In each episode, the algorithm follows an  $\epsilon$ -greedy approach for exploration. It generates a random number  $\mathbb{X}$  from uniform distribution over  $[0, 1]$ . If the random number generated is less than  $\epsilon_t$ , it plays action randomly uniformly from the action space  $\mathcal{A}$ . Else the algorithm uses the auxiliary algorithm *Get\_EMQL\_Action* to determine the action to be played using the proposed policy of Equation (17) in Section III. After playing an action, if a new observation is available to the algorithm, it updates the following parameters - the number of times state  $s$  was visited and action  $a$  was taken in  $s$   $N(s, a)$ , number of times  $s'$  was observed as the next state  $P(s, a, s')$ , and reward observed for the state action pair  $R(s, a)$ .

*Get\_EMQL\_Action* is presented in Algorithm 2, and is described in subsection IV-B. If the current state is available, the expected Q-function becomes the true Q function. Thus Algorithm 1 evaluates Q-value table and probability estimates, and Algorithm 2 computes the expected Q-value for Equation (17). Last line of Algorithm 2 then returns the action which maximizes the expected Q-value for the proposed policy. We further note that as the estimates of transition probabilities and Q-values are improved by the Algorithm 1, the gap between the value of the proposed policy and the expected optimal value of the unobserved state can be bounded by Theorem 1.

### B. Complexity

At each time step  $t$ , Algorithm 2 computes the expected value of the Q-function whenever an action needs to be taken. This requires  $\mathcal{O}(d|\mathcal{S}|^2 + \log(|\mathcal{A}|))$  computations. For our algorithm, we calculate the probability vector  $\bar{\mathbf{p}}$  which is the conditional probability distribution of the states given the last known state  $s_{t-d}$ , and the sequence of actions  $a_{t-d}, \dots, a_{t-1}$ . For time  $t-d$ , the true state is known and conditional probability becomes

$$\bar{\mathbf{p}}_{t-d}(s) = \begin{cases} 1, & s = s_{t-d}, \\ 0, & otherwise \end{cases} \quad (36)$$

Then, for each next time step, the probability vector is updated using the following recursion equation.

$$\bar{\mathbf{p}}_{t-k} = (\hat{p}(\cdot, a_{t-(d-k)}, \cdot))^T \bar{\mathbf{p}}_{t-k-1} \quad \forall s \in \mathcal{S}, \forall 1 \leq k < d \quad (37)$$

where  $\hat{p}(\cdot, a_{t-k-1}, \cdot)$  is the state transition matrix of MDP  $\mathcal{M}$  induced by action  $a_{t-k-1}$ .  $\hat{p}(\cdot, a_{t-k-1}, \cdot)$  is obtained using Equation (38). Since we do this update  $d$  times, and each matrix multiplication costs  $\mathcal{O}(|\mathcal{S}|^2)$ , the total complexity to compute the state probability at each time step  $t$  becomes  $\mathcal{O}(d|\mathcal{S}|^2)$ . Fetching the maximum element cost extra  $\mathcal{O}(\log(|\mathcal{A}|))$ . The overall complexity at any time step  $t$  thus becomes  $\mathcal{O}(d|\mathcal{S}|^2 + \log(|\mathcal{A}|))$ .

---

**Algorithm 1** *Expectation Maximization Q-Learning (EMQL)*


---

**Input:**  $\mathcal{S}, \mathcal{A}, \gamma, \epsilon_t$

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

**for**  $a \in \mathcal{A}$  **do**

$\hat{p}(s, a, s') = 0, \hat{r}(s, a) = 0$

$P(s, a, s') = 0, N(s, a) = 0$

$R(s, a) = 0, Q(s, a) = 0$

**end for**

**end for**

**while** 1 **do**

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

**for**  $a \in \mathcal{A}$  **do**

      Update probability and reward estimates

$$\hat{p}(s, a, s') = \frac{P(s, a, s')}{\max(1, N(s, a))} \quad (38)$$

$$\hat{r}(s, a) = \frac{R(s, a)}{\max(1, N(s, a))} \quad (39)$$

**end for**

$V(s) = \max_a Q(s, a)$

**end for**

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

**for**  $a \in \mathcal{A}$  **do**

    Update Q function

$$Q(s, a) = \hat{r}(s, a) + \sum_{s' \in \mathcal{S}} \gamma V(s') \hat{p}(s, a, s') \quad (40)$$

**end for**

**end for**

$t = 0$

**for**  $t = 1, 2, \dots$  **do**

$t+ = 1$

  Known STATE  $s_{t-d}$

**if**  $\mathbb{X} \sim U(0, 1) < \epsilon_t$  **then**

    play  $a_t$  randomly uniFORMly from  $\mathcal{A}$

**else**

    play  $a_t = \text{Get\_EMQL\_Action}(s_{t-d}, a_{t-d}, \dots, a_{t-1})$

**end if**

**if** Observation  $(s_{t-(d-1)}, r_{t-(d-1)})$  available **then**

$R(s_{t-d}, a_{t-d}) + = r_{t-(d-1)}$

$P(s_{t-d}, a_{t-d}, s_{t-(d-1)}) + = 1$

$N(s_{t-d}, a_{t-d}) + = 1$

**end if**

**end for**

**end while**

---

---

**Algorithm 2** *Get\_EMQL\_Action*


---

**Input:**  $\mathcal{S}, \mathcal{A}, \hat{p}, Q, (s_{t-d}, a_{t-d}, \dots, a_{t-1})$   
**Output:** Estimated greedy action  $a_t$   
 $\bar{\mathbf{p}} = [0, \dots, 0]$ , vector of length  $|\mathcal{S}|$   
 $\bar{\mathbf{p}}[s_{t-d}] = 1$   
**for**  $0 \leq k < d$  **do**  
     $\bar{\mathbf{p}} = (\hat{p}(:, a_{t-(d-k)}, :))^T \bar{\mathbf{p}}$   
**end for**  
 $\bar{Q} = \bar{\mathbf{p}}^T Q$   
Return  $\arg \max_a \bar{Q}$

---

## V. EVALUATION

We evaluate our algorithm EMQL on OpenAI Gym platform [3]. We consider the standard Frozen Lake ( $8 \times 8$  grid), and Cart Pole problem of OpenAI Gym which is as per the description in [16]. Frozen Lake environment has a discrete state space and Cart Pole environment has continuous state space with discrete action space.

We compare our algorithm with Extended MDP formulation by [1], Model Based Simulation (MBS) algorithm of [18], and dQ algorithm of [15] for constant delays. We also compare the proposed EMQL algorithm with MBS algorithm for stochastic delays. The metric of comparison is total cumulative reward accumulated at the end of each episode averaged over last 50 episodes. Exploration factor is time dependent and is chosen as  $\epsilon_t = \frac{\log(|\mathcal{S}| \sum_{s,a} N(s,a) + 1)}{\sum_a N(s,a) + 1} \quad \forall s, a$ . This choice of exploration factor is same across all simulations. For stochastic delays, creating an augmented MDP is not feasible as delays can be arbitrarily large. We considered 50 iterations of Frozen Lake environment, and 20 iterations of Cart Pole environment. Each iteration is trained over 1000 episodes.

For constant delays, we chose delays in the range of  $d \in \{2, 4\}$ . For stochastic delays, each observation was independently delayed by delays generated using a geometric distribution with parameter  $p$ . The expected delay for this distribution is  $\frac{1}{1-p}$ . We note that, this may create asynchronous observations as delay  $d_1$  of observation at  $t_1$  may be higher than delay  $d_2$  of observation at  $t_2$ , where  $t_1 + d_1 > t_2 + d_2$ . The issue of asynchronous delays can be dealt by introducing time stamps in observations.

### A. Results

Simulations results based for both Frozen Lake and Cart Pole environment are presented in Figure 1 for constant delays. For stochastic delays the results are presented in Figure 2. In both figures, median of rewards in each iteration is plotted along with the top and bottom quantiles.

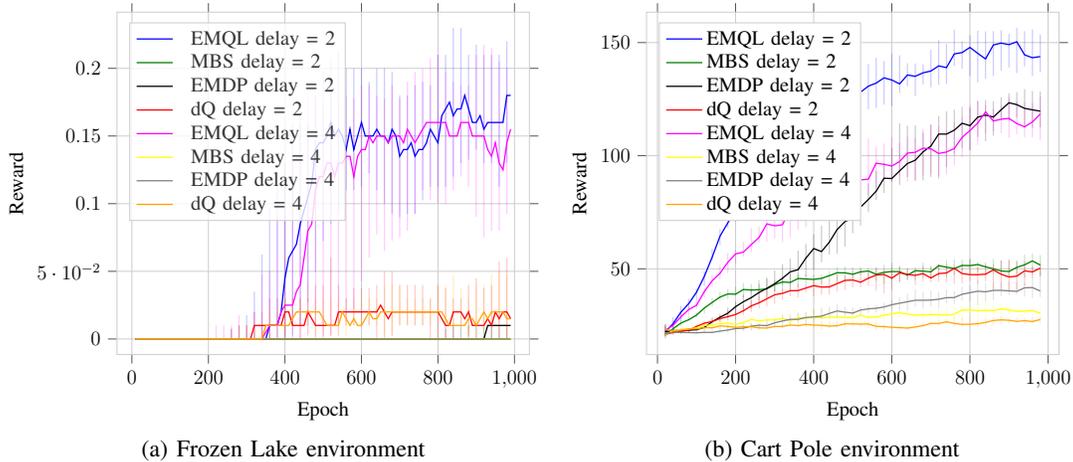


Fig. 1: Reward accumulated by EMQL, MBS, EMDP, and dQ-learning algorithm with constant delays for Frozen Lake and Cart Pole environments. EMQL achieves higher average reward per episodes compared to other algorithms.

As observed in Figure 1, total reward per episode is higher for EMQL algorithm compared to all other algorithms (MBS, dQ, EMDP) in the presence of constant delays. As delay increases, gap between the rewards also increase between the two algorithms. For MBS algorithm, this can be reasoned as - with large values of delays the state with largest likelihood might have lower probability of occurring.

For dQ algorithm, the reason for increasing difference in accumulated rewards can be credited to the memoryless algorithm. Also, Extended MDP algorithm is slow in convergence because of larger state space which grows exponentially. This reduction in convergence speed is visible in Figure 1b where delay of 4 time steps cause a significant drop in performance of Extended MDP algorithm.

For stochastic delays, the difference between the rewards accumulated by EMQL, and MBS algorithm increases when the probability of delay increase. For Frozen Lake environment, the proposed EMQL algorithm doesn't suffer much degradation even by increasing the expected delay from 1 unit to 3 units. However, MBS algorithm is not able to achieve similar performance for small expected delays. In Cart Pole environment, EMQL algorithm beats the MBS algorithm significantly even when the stochastic delays are geometric distributed with expected delay of 1 unit in Figure 2b.

## VI. CONCLUSION

We considered the problem of delays in observation updates for a reinforcement learning agent. The current state of the environment is not immediately available to the agent. We proved that the expected immediate rewards generated for MDP with delays is same as expected immediate rewards generated

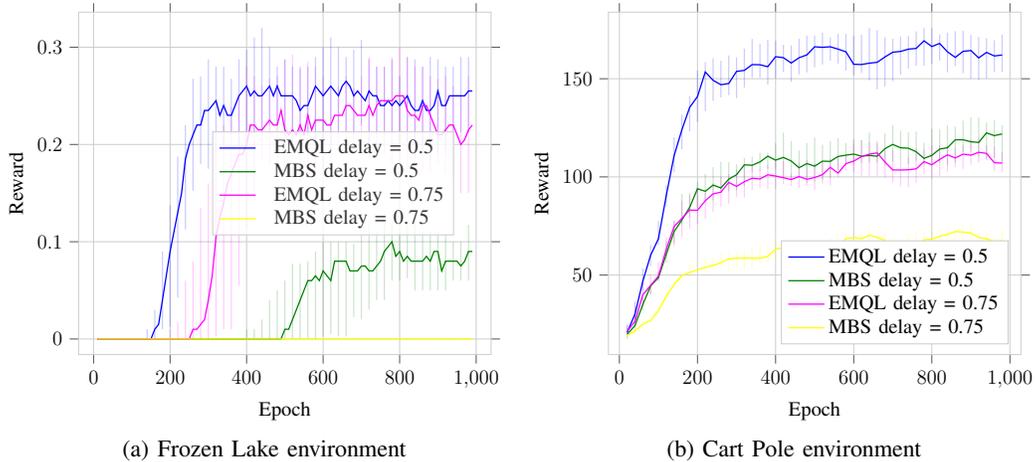


Fig. 2: Reward accumulated by EMQL, MBS algorithms with stochastic delays for Frozen Lake and Cart Pole environments. EMQL achieves higher average reward per episodes compared to MBS algorithm when expected delays are high.

for corresponding extended MDP without delays. We proposed a new policy which can handle stochastic delays by optimizing on optimal Q-function of the true MDP. We then provided a lower bound on the value function for all states following the proposed policy. Based on this policy, we proposed a new algorithm, *Expectation Maximization Q-Learning* (EMQL), which is robust under constant, and stochastic delays. Using the knowledge of latest available state, sequence of actions, estimated transition probabilities, and reward distributions of the underlying MDP, we determine the best action which maximizes the expected reward for the unobserved state. Evaluations demonstrate the improvement over existing algorithms under constant, and stochastic delays.

## REFERENCES

- [1] Eitan Altman and Philippe Nain. *Closed-loop control with delayed information*, volume 20. ACM, 1992.
- [2] Hamid Arabnejad, Claus Pahl, Pooyan Jamshidi, and Giovanni Estrada. A comparison of reinforcement learning techniques for fuzzy cloud auto-scaling. In *Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGrid '17*, pages 64–73, Piscataway, NJ, USA, 2017. IEEE Press.
- [3] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *CoRR*, abs/1606.01540, 2016.
- [4] Nesrine Changuel, Bessem Sayadi, and Michel Kieffer. Online learning for qoe-based video streaming to mobile receivers. In *2012 IEEE Globecom Workshops*, pages 1319–1324. IEEE, 2012.
- [5] Ignasi Clavera, Anusha Nagabandi, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *International Conference on Learning Representations*, 2019.

- [6] Yue Deng, Feng Bao, Youyong Kong, Zhiqian Ren, and Qionghai Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE transactions on neural networks and learning systems*, 28(3):653–664, 2017.
- [7] Arnob Ghosh, Vaneet Aggarwal, and Feng Qian. A rate adaptation algorithm for tile-based 360-degree video streaming. *arXiv preprint arXiv:1704.08215*, 2017.
- [8] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4868–4878, 2018.
- [9] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274, 2002.
- [10] Konstantinos V Katsikopoulos and Sascha E Engelbrecht. Markov decision processes with delays and asynchronous cost collection. *IEEE transactions on automatic control*, 48(4):568–574, 2003.
- [11] Elizaveta Kuznetsova, Yan-Fu Li, Carlos Ruiz, Enrico Zio, Graham Ault, and Keith Bell. Reinforcement learning for microgrid energy management. *Energy*, 59:133–146, 2013.
- [12] Shichao Liu, Xiaoyu Wang, and Peter Xiaoping Liu. Impact of communication delays on secondary frequency control in an islanded microgrid. *IEEE Transactions on Industrial Electronics*, 62(4):2021–2031, 2015.
- [13] A Rupam Mahmood, Dmytro Korenkevych, Brent J Komer, and James Bergstra. Setting up a reinforcement learning task with a real-world robot. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4635–4640. IEEE, 2018.
- [14] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- [15] Erik Schuitema, Lucian Buşoniu, Robert Babuška, and Pieter Jonker. Control delay in reinforcement learning for real-time dynamic systems: a memoryless approach. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3226–3231. IEEE, 2010.
- [16] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [17] Jean Walrand and Shyam Parekh. Communication networks: a concise introduction. *Synthesis Lectures on Communication Networks*, 3(1):1–192, 2010.
- [18] Thomas J Walsh, Ali Nouri, Lihong Li, and Michael L Littman. Learning and planning in environments with delayed feedback. *Autonomous Agents and Multi-Agent Systems*, 18(1):83, 2009.