

# Control in Stochastic Environment with Delays: A Model-based Reinforcement Learning Approach

Anonymous Author(s)

Affiliation

Address

email

## 1 Appendix

### 2 1.1

**Theorem 1.** Assume a discrete-time MDP with an infinite time horizon. The Markovian movement is deterministic, i.e., for arbitrary  $(s, a) \in \mathcal{S} \times \mathcal{A}$ ,  $t \geq 0$ , there exists a  $s' \in \mathcal{S}$  such that  $P(S_{t+1} = s' \mid S_t = s, A_t = a) = 1$  for all  $t = 0, 1, \dots$ . Assume a  $d$ -step delay exists in the observation, for an arbitrary augmented state

$$I_t = (s_{t-d}, a_{t-d}, \dots, a_{t-1}),$$

the policy function of the SMBS method

$$\pi_1(I_t) = \arg \max_{a \in \mathcal{A}} \frac{1}{M} \sum_{i=1}^M q^*(s_t^{(i)}, a). \quad (1)$$

is equivalent to the following optimal policy:

$$\pi_{opt}(I_t) = \arg \max_{a \in \mathcal{A}} \tilde{q}^*(I_t, a), \quad (2)$$

where  $\tilde{q}^*$  denotes the optimal  $Q$ -function for the AMDP.

*Proof.* For the augmented state  $I_t$ , if the movement is deterministic, there exists a sequence of states (one trajectory)  $\tau' = (s'_{t-d+1}, \dots, s'_t) \in \mathcal{S}^d$  such that  $P(S_{t-d+i} = s'_{t-d+i} \mid I_t) = 1$  for  $i = 1, 2, \dots, d$ . Therefore,  $\pi_1$  can be written as

$$\pi_1(I_t) = \arg \max_{a \in \mathcal{A}} \frac{1}{M} \sum_{i=1}^M q^*(s_t^{(i)}, a) = \arg \max_a q^*(s'_t, a)$$

We show the two policy functions are equivalent using mathematical induction. We denote the optimal  $Q$ -functions, the optimal value functions and the reward function for  $k$ -step delayed AMDP as  $q_k^*$ ,  $v_k^*$  and  $r_k$  respectively, denote the  $k$ -step augmented state as  $I_t^{(k)} = (s_{t-d}, a_{t-d}, \dots, a_{t-d+k-1})$  for  $k = 1, 2, \dots, d$ . Note  $\tilde{q}^* = q_d^*$ .

Using the connection between the optimal value function and the optimal  $Q$ -function in [1], for  $k = 1, 2, \dots, d-1$ , we have

$$q_k(I_t^{(k)}, a_{t-d+k}) = \mathbb{E}[r_k(I_t^{(k)}, a_t) + \gamma v_k^*(I_{t+1}^{(k)}) \mid I_t^{(k)}, a_{t-d+k}] \quad (3)$$

Due to the deterministic movement,  $I_{t+1}^{(k)}$  is deterministic because  $P(I_{t+1}^{(k)} = (s'_{t-d+1}, a_{t-d+1}, \dots, a_{t-d+k}) \mid I_t^{(k)}, a_{t-d+k}) = 1$ . Therefore, (3) can be simplified as

$$q_k^*(I_t^{(k)}, a_{t-d+k}) = r_k(I_t^{(k)}, a_{t-d+k}) + \gamma v_k^*(I_{t+1}^{(k)}) \quad (4)$$

Because  $v_k^*(I_{t+1}^{(k)}) = \max_a q_k^*(I_{t+1}^{(k)}, a)$ , we can further have

$$q_k^*(I_t^{(k)}, a_{t-d+k}) = r_k(I_t^{(k)}, a_{t-d+k}) + \gamma \max_a q_k^*(I_{t+1}^{(k)}, a) \quad (5)$$

It can be seen from (5) that  $q_k^*$  satisfies the Bellman equation of the  $(k+1)$ -step AMDP (equation 3.19 in [1]):

$$v_{k+1}(I_t^{(k+1)}) = \max_a \mathbb{E}[r_{k+1}(I_t^{(k+1)}, a) + \gamma v_{k+1}(I_{t+1}^{(k+1)}) \mid I_t^{(k+1)}, a],$$

16 where  $I_{t+1}^{(k+1)}$  denotes  $(I_{t+1}^{(k)}, a)$ . Please note that we use the asynchronized reward function in [2],  
 17 hence  $r_k(I_t^{(k)}, a) = r_{k+1}(I_{t+1}^{(k)}, a') = r(s_{t-d}, a_{t-d})$  where  $r$  is the reward function of the non-  
 18 delayed MDP.

19 Therefore,  $q_k^*$  is the solution of the optimal problem defined on the  $(k+1)$ -step AMDP. Due to the  
 20 uniqueness of this solution, we have  $v_{k+1}^* = q_k^*$ . Combining this conclusion with (4), we obtain

$$q_d^*(I_t, a_t) = \sum_{j=0}^{d-1} \gamma^j r(s'_{t-d+j}, a_{t-d+j}) + \gamma^d q^*(s'_t, a_t) \quad (6)$$

Apply  $\arg \max$  operator on  $a_t$  on both sides of (6), we have

$$\pi_{\text{opt}}(I_t) = \arg \max_{a_t \in \mathcal{A}} q_d^*(I_t, a_t) = \arg \max_{a_t \in \mathcal{A}} q^*(s'_t, a_t) = \pi_1(I_t)$$

21 This concludes that two policy functions are equivalent.

22

□

## 23 1.2

24 **Theorem 2.** Assume a discrete-time MDP with a positive reward function and a finite discrete action  
 25 space  $\mathcal{A}$ . For any  $a \in \mathcal{A}$  and augmented state  $I_t \in \mathcal{I}$ , assume the random variable  $q^*(s_t, a)$  has  
 26 mean  $\bar{Q}(a)$  and variance  $\hat{Q}(a)^2$ . Then, for  $\delta > 0$ , we have

$$P\left(\max_{a \in \mathcal{A}} \bar{Q}_M(a) \leq \frac{1}{|\mathcal{A}|} \mathbb{E}[V^*(s) \mid I_t] - \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a)\right) \leq \frac{|\mathcal{A}|}{\delta^2}$$

*Proof.* Denote  $\bar{Q}(a) = \mathbb{E}_s[q^*(s, a) \mid I]$ , thus  $\bar{Q}_M(a)$  is the sample estimates of  $\bar{Q}(a)$ . For any  $a \in \mathcal{A}$ , by Chebyshev inequality, we have

$$P\left(|\bar{Q}_M(a) - \bar{Q}(a)| > \delta \sqrt{\text{Var}[\bar{Q}_M(a)]}\right) \leq \frac{1}{\delta^2}$$

where  $\hat{Q}_M(a)$  is the standard deviation of  $\bar{Q}_M(a)$ . We can replace the variance term as  $\hat{Q}(a)^2/M$ , thus

$$P\left(|\bar{Q}_M(a) - \bar{Q}(a)| > \frac{\delta}{\sqrt{M}} \hat{Q}(a)\right) \leq \frac{1}{\delta^2}$$

We may further have, for any  $a \in \mathcal{A}$ ,

$$P\left(|\bar{Q}_M(a) - \bar{Q}(a)| > \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a)\right) \leq \frac{1}{\delta^2}.$$

27 Further, since the event sets have the following relationship,

$$\begin{aligned} & \{\omega \mid \max_{a \in \mathcal{A}} |\bar{Q}_M(a) - \bar{Q}(a)| > \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a)\} \\ &= \{\omega \mid |\bar{Q}_M(a) - \bar{Q}(a)| > \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a), \exists a \in \mathcal{A}\} \\ &= \bigcup_{a \in \mathcal{A}} \{\omega \mid |\bar{Q}_M(a) - \bar{Q}(a)| > \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a)\}. \end{aligned}$$

We have

$$P\left(\max_{a \in \mathcal{A}} |\bar{Q}_M(a) - \bar{Q}(a)| > \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a)\right) \leq \sum_{a \in \mathcal{A}} P\left(|\bar{Q}_M(a) - \bar{Q}(a)| > \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a)\right) \leq \frac{|\mathcal{A}|}{\delta^2}.$$

Since

$$\max_{a \in \mathcal{A}} |\bar{Q}_M(a) - \bar{Q}(a)| > \left| \max_{a \in \mathcal{A}} \bar{Q}_M(a) - \max_{a \in \mathcal{A}} \bar{Q}(a) \right|,$$

28 we have,

$$\begin{aligned} & P\left(\left| \max_{a \in \mathcal{A}} \bar{Q}_M(a) - \max_{a \in \mathcal{A}} \bar{Q}(a) \right| > \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a)\right) \\ & \leq P\left(\max_{a \in \mathcal{A}} |\bar{Q}_M(a) - \bar{Q}(a)| > \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a)\right) \\ & \leq \frac{|\mathcal{A}|}{\delta^2}. \end{aligned}$$

Using lemma 3 in [3],

$$\max_{a \in \mathcal{A}} \bar{Q}(a) \geq \frac{1}{|\mathcal{A}|} \mathbb{E}[V^*(s) \mid I],$$

29 We have

$$\begin{aligned} \frac{|\mathcal{A}|}{\delta^2} & \geq P\left(\left| \max_{a \in \mathcal{A}} \bar{Q}_M(a) - \max_{a \in \mathcal{A}} \bar{Q}(a) \right| > \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a)\right) \\ & = P\left(\max_{a \in \mathcal{A}} \bar{Q}_M(a) < \max_{a \in \mathcal{A}} \bar{Q} - \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a)\right) + P\left(\max_{a \in \mathcal{A}} \bar{Q}_M(a) > \max_{a \in \mathcal{A}} \bar{Q} + \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a)\right) \\ & \geq P\left(\max_{a \in \mathcal{A}} \bar{Q}_M(a) < \frac{1}{|\mathcal{A}|} \mathbb{E}[V^*(s) \mid I] - \frac{\delta}{\sqrt{M}} \max_{a \in \mathcal{A}} \hat{Q}(a)\right) \end{aligned}$$

30 This concludes the proof.

31

□

## 32 References

- 33 [1] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- 34 [2] K. V. Katsikopoulos and S. E. Engelbrecht. Markov decision processes with delays and asyn-  
35 chronous cost collection. *IEEE transactions on automatic control*, 48(4):568–574, 2003.
- 36 [3] M. Agarwal and V. Aggarwal. Blind decision making: Reinforcement learning with delayed ob-  
37 servations. In *Proceedings of the International Conference on Automated Planning and Schedul-*  
38 *ing*, volume 31, pages 2–6, 2021.