Markov Decision Processes

Rupam Mahmood

January 22, 2020

MDPs review

Probability of an outcome or a sequence of experience: $P(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1, A_1 = a_1, R_2 = r_2, \cdots)$

history (everything before S_t): $H_t = (S_0, A_0, R_1, S_1, A_1, R_2, \cdots, S_{t-1}, A_{t-1}, R_t)$

Probability of sequence up to S_{t+1}: $P(H_t = h, S_t = s, A_t = a, R_{t+1} = r, S_{t+1} = s')$

$$= P(R_{t+1} = r, S_{t+1} = s' \mid S_t = s, A_t = a) P(A_t = a \mid S_t = s) P(H_t = h, S_t = s)$$

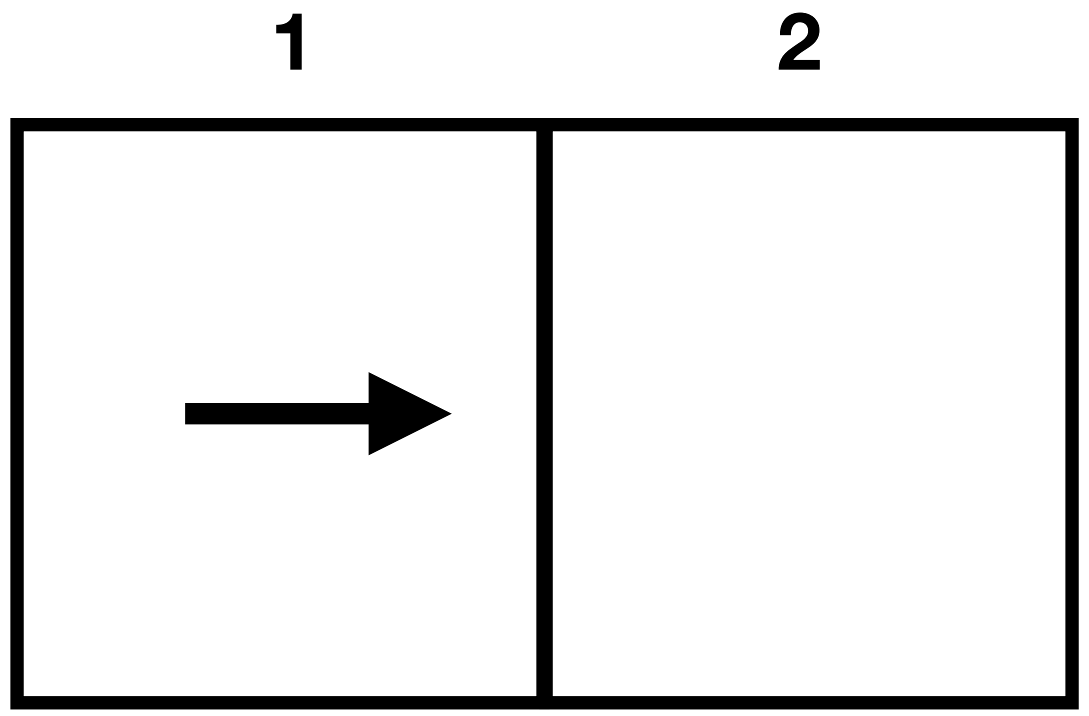
shorthands:

$$P(R_{t+1} = r, S_{t+1} = s' \mid S_t = s, A_t = a) = p(r, s' \mid s, a); \quad P(A_t = a \mid S_t = s) = \pi(a \mid s)$$

$$= p(r, s' \mid s, a)\pi(a \mid s) p(\bar{r}, s \mid \bar{s}, \bar{a}) \pi(\bar{a} \mid \bar{s}) \cdots$$

previous reward state and action

Example 1: An MDP



State S is the location and the orientation: (1, →)

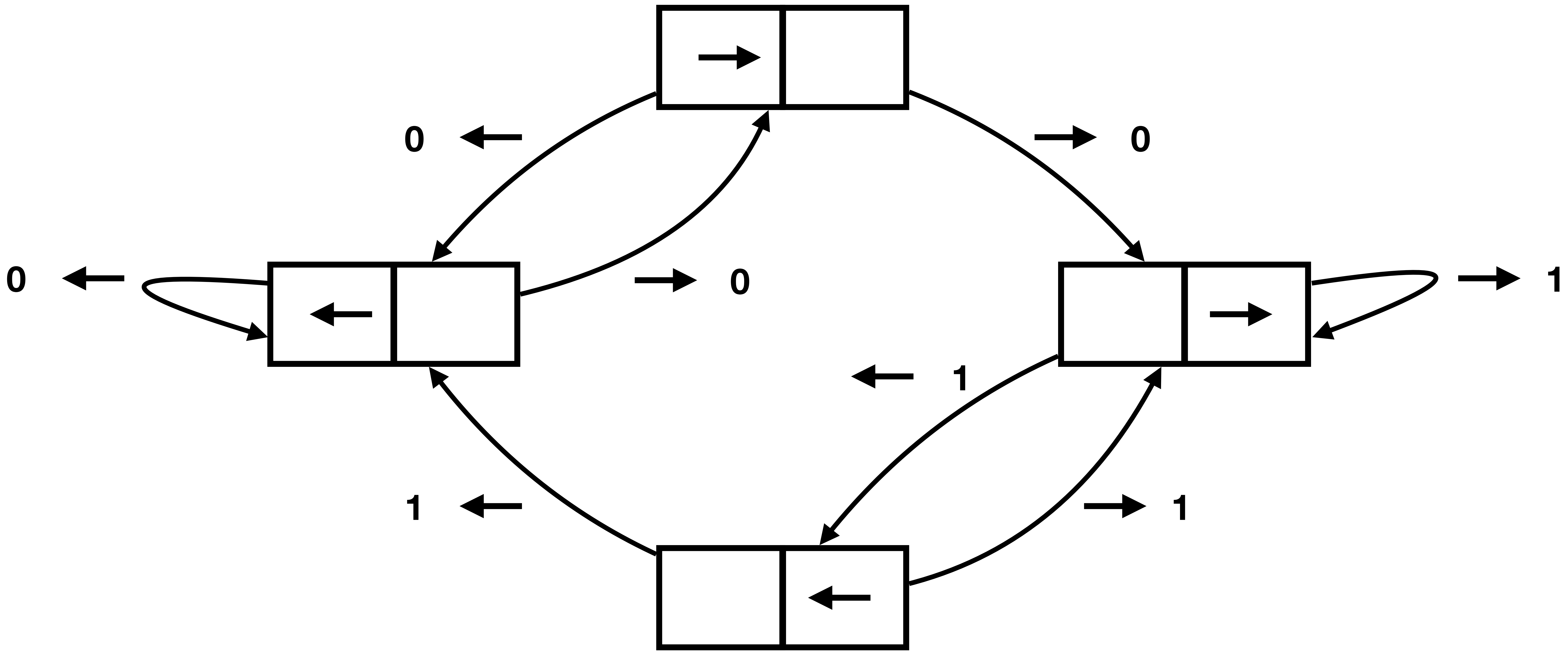
Action is ← or →

Reward is +1 for any action at location 2, and 0 otherwise

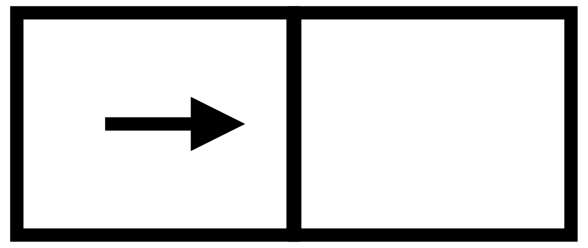
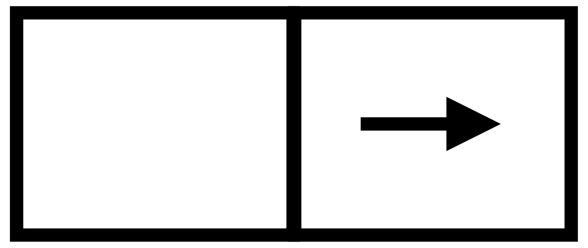
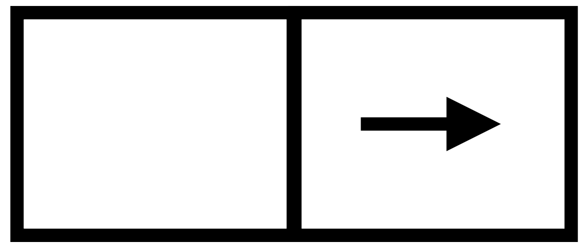
$$P(S_{t+1} = (2, \rightarrow) \mid S_t = (1, \rightarrow), A_t = \rightarrow) = 1$$

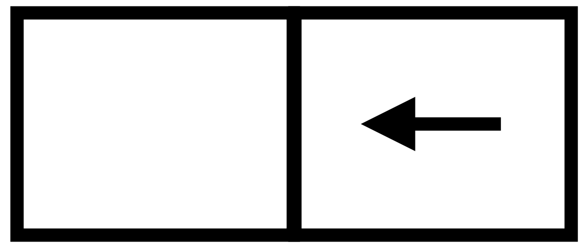
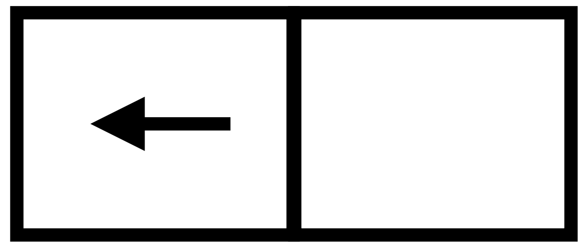
$$P(S_{t+1} = (2, \leftarrow) \mid S_t = (2, \rightarrow), A_t = \leftarrow) = 1$$

Example 1 (continued): The state-transition diagram

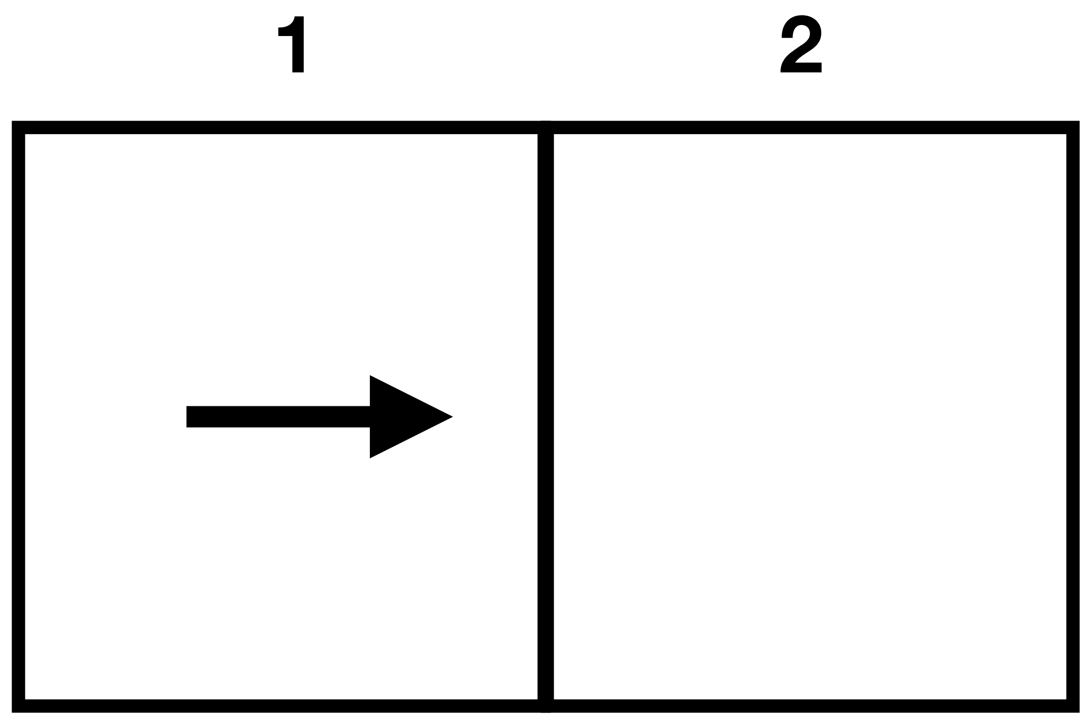


Example 1 (continued): A sample sequence

S_0	A_0	R_1	S_1	A_1	R_2	S_2
	$\rightarrow$	0		$\rightarrow$	1	

A_2	R_3	S_3	A_3	R_4	S_5
$\leftarrow$	1		$\leftarrow$	1	

Example 2: Not an MDP



State O is just the location: 1

Action is ← or →

Reward is +1 for any action at location 2, and 0 otherwise

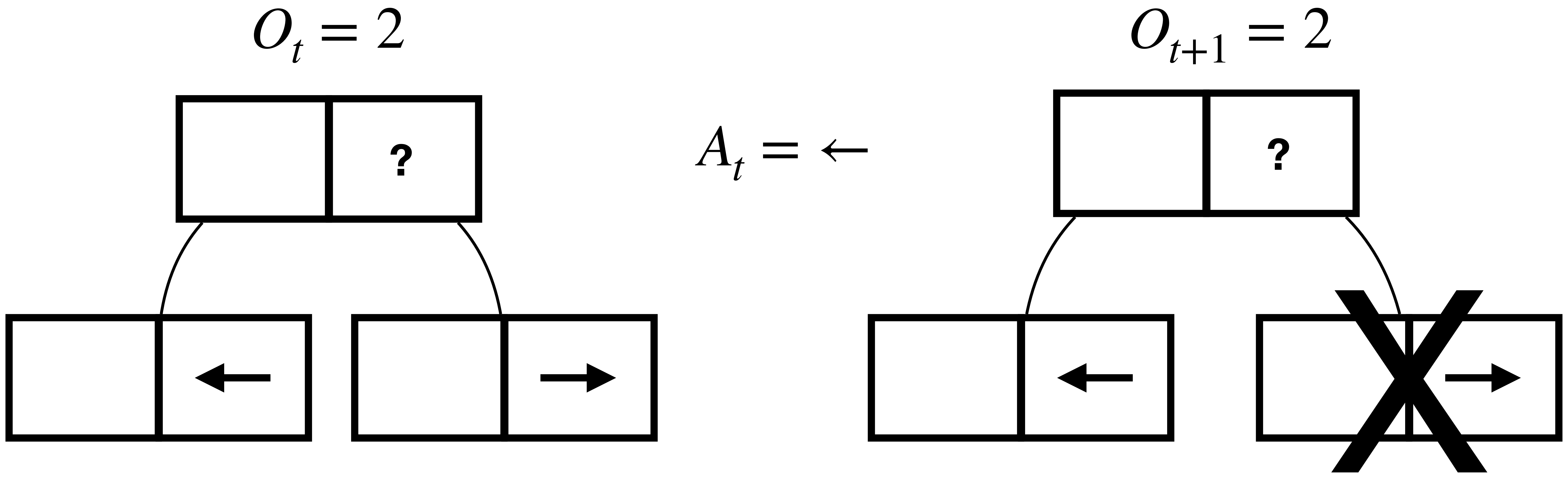
Show that: $P(O_{t+1} = 2 \mid O_t = 2, A_t = \leftarrow) \neq P(O_{t+1} = 2 \mid O_t = 2, A_t = \leftarrow, R_t = 0)$

That is state O is not Markov

Example 2: Not an MDP (continued)

Show that: $P(O_{t+1} = 2 \mid O_t = 2, A_t = \leftarrow) \neq P(O_{t+1} = 2 \mid O_t = 2, A_t = \leftarrow, R_t = 0)$

$$P(O_{t+1} = 2 \mid O_t = 2, A_t = \leftarrow)$$

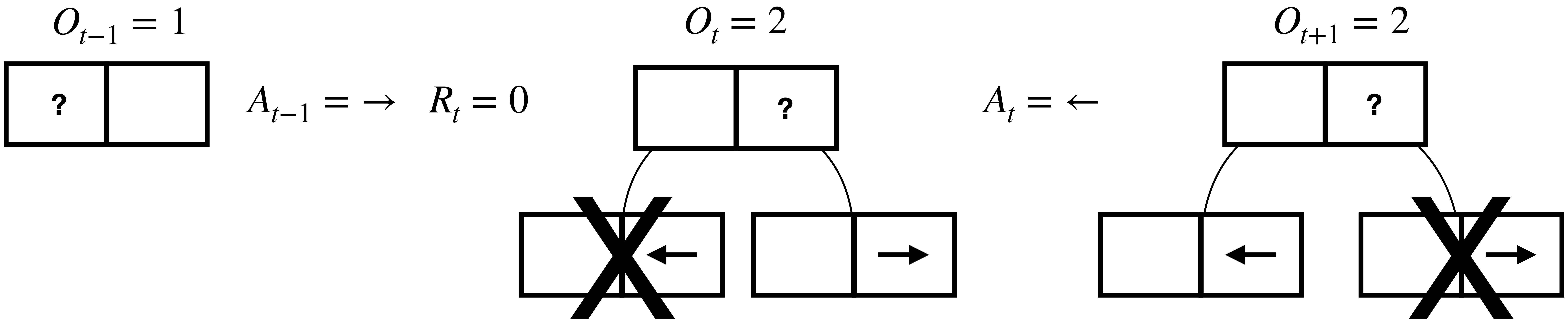


$$P(O_{t+1} = 2 \mid O_t = 2, A_t = \leftarrow) \neq 0 \textbf{ or } 1$$

Example 2: Not an MDP (continued)

Show that: $P(O_{t+1} = 2 \mid O_t = 2, A_t = \leftarrow) \neq P(O_{t+1} = 2 \mid O_t = 2, A_t = \leftarrow, R_t = 0)$

$$P(O_{t+1} = 2 \mid O_t = 2, A_t = \leftarrow, R_t = 0)$$



$$P(O_{t+1} = 2 \mid O_t = 2, A_t = \leftarrow, R_t = 0) = 1$$

The goal of a bandit agent

Maximize expected reward R

$$\pi(a) = P(A = a)$$

$$v_\pi = E_\pi[R] = E_\pi[E[R \mid A]] = E_\pi[q_*(A)]$$

Choose policy π **that maximizes** v_π

The goal of a contextual bandit agent

Maximize expected reward R for all state S

$$\pi(a \mid s) = P(A = a \mid S = s)$$

$$v_\pi(s) = E_\pi[R \mid S = s] = E_\pi[E[R \mid S = s, A]] = E_\pi[q_*(s, A)]$$

Choose policy π that maximizes v_π for all state S

The goal of an agent in an MDP

Maximize expected sum of discounted future rewards R **from all states** S

Maximize expected return G **from all states** S

return: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

$$= R_{t+1} + \gamma G_{t+1}$$

value function: $v_\pi(s) = E_\pi[G_t \mid S_t = s]$

Choose policy π **that maximizes** v_π **for all state** S

Worksheet question 1

1. Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2, R_2 = -2, R_3 = 0$ followed by an infinite sequence of 7s. What are G_1 and G_0?

Worksheet question 2

2. Assume you have a bandit problem with 4 actions, where the agent can see rewards from the set $\mathcal{R} = \{-3.0, -0.1, 0, 4.2\}$. Assume you have the probabilities for rewards for each action: $p(r|a)$ for $a \in \{1, 2, 3, 4\}$ and $r \in \{-3.0, -0.1, 0, 4.2\}$. How can you write this problem as an MDP? Remember that an MDP consists of $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P, \gamma)$.

   More abstractly, recall that a Bandit problem consists of a given action space $\mathcal{A} = \{1, ..., k\}$ (the k arms) and the distribution over rewards $p(r|a)$ for each action $a \in \mathcal{A}$. Specify an MDP that corresponds to this Bandit problem.

Worksheet question 3

3. Prove that the discounted sum of rewards is always finite, if the rewards are bounded: $|R_{t+1}| \leq R_{\max}$ for all t for some finite $R_{\max} > 0$.

$$\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty \qquad\qquad \text{for } \gamma \in [0, 1)$$

Hint: Recall that $|a + b| < |a| + |b|$.

The reward hypothesis

That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward).