Markov Decision Processes

Rupam Mahmood

January 20, 2020

Admin

- ✓ This week’s assignment and deadline are a bit different
- ✓ The assignment is divided into two parts: submission and review
- ✓ Submissions (due Thursday) will be graded according to the median of three reviews (due Sunday)
- ✓ However, if someone submits and does not participate in reviewing, they get 0

Bandits review

- ✓ What is the experiment?
- ✓ What are the outcomes?
- ✓ What are the random variables involved?

$$P(A = a, R = r) = \underbrace{P(R = r \,|\, A = a)}_{\text{environment determines}}\underbrace{P(A = a)}_{\text{agent decides}}$$

Contextual bandits

$$P(S = s, A = a, R = r)$$

$$= P(R = r \mid S = s, A = a)P(S = s, A = a)$$

$$= P(R = r \mid S = s, A = a)P(A = a \mid S = s)P(S = s)$$

Markov decision processes

- ✓ What is the experiment?
- ✓ What are the outcomes?
- ✓ What are the random variables involved?

$$P(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1, A_1 = a_1, R_2 = r_2, \cdots)$$

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

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

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

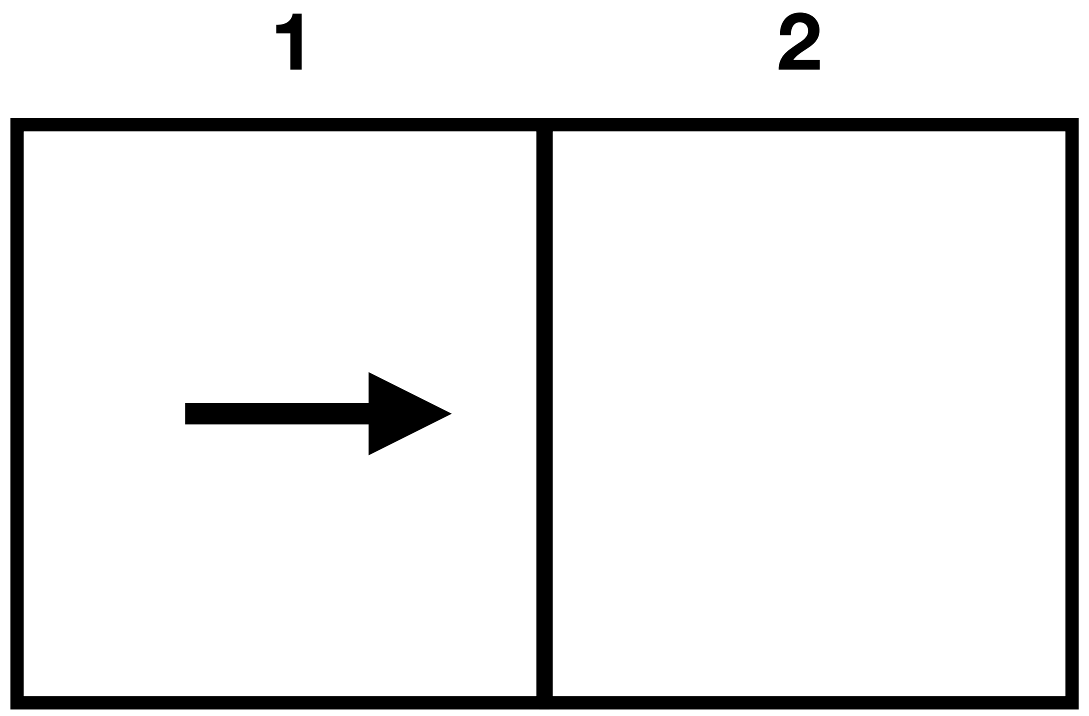
$$= P(R_{t+1} = r, S_{t+1} = s' \mid S_t = s, A_t = a)P(A_t = a \mid S_t = s)P(H_t = h, S_t = s)$$

Markov property

A choice that does not hurt

Same logic applies here recursively

Example 1: An MDP



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

Action is ← or →

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

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

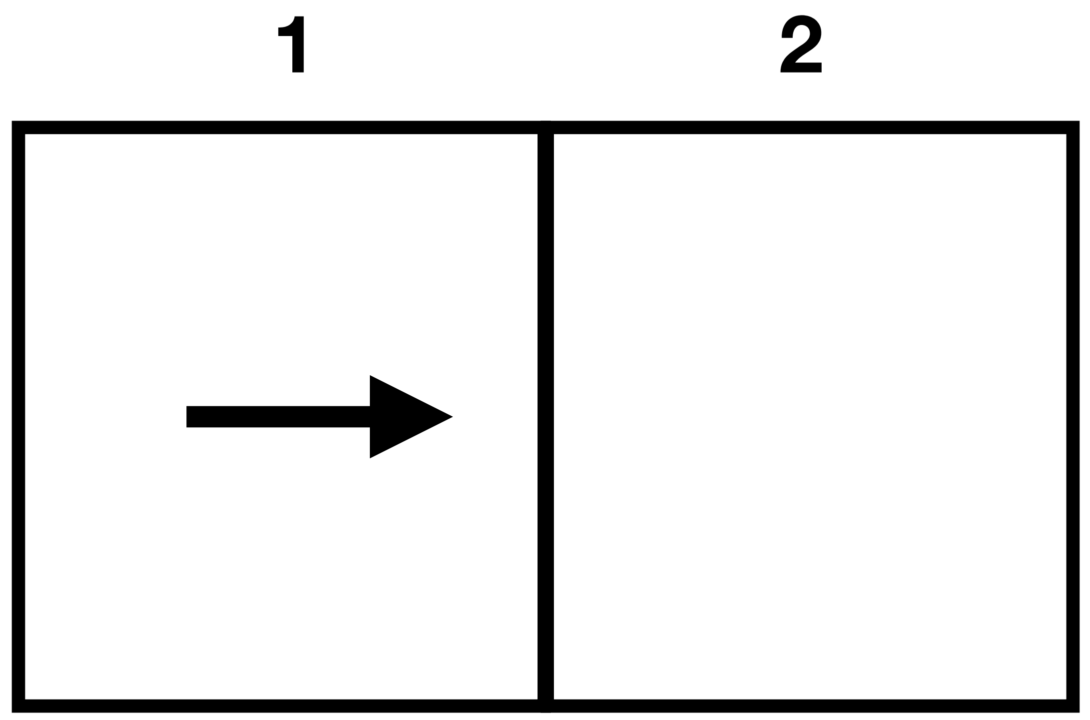
$$P(S_{t+1} = (2, \leftarrow) \,|\, S_t = (2, \rightarrow), A_t = \leftarrow) = 1$$

Example 1 (continued): A sample sequence

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

A_2	R_3	S_3	A_3	R_4	S_5
$\leftarrow$	1	$\boxed{}\boxed{\leftarrow}$	$\leftarrow$	1	$\boxed{\leftarrow}\boxed{}$

Example 2: Not an MDP



State is just the location: 1

Action is ← or →

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

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