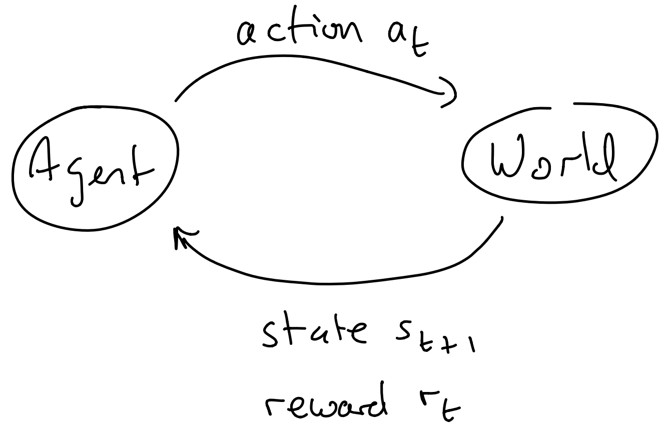
Lecture 20

Markov Decision Processes

CS 181
Apr 2021



time t

Data $D = (s_0, a_0, r_0, s_1, a_1, r_1, \dots)$

Markov Decision Process (S, A, r, p)

states $S = \{1, \dots, |S|\}$ reward function $r(s,a)$

actions $A = \{1, \dots, |A|\}$

transition model $p(s'|s,a)$

next state, current state, action

policy $\pi(a|s)$
prob. of taking action a, given in state s

$\boxed{\text{Markov assumption}}$

$$p_t(s_{t+1}|s_1, s_2, \dots, s_t, a_1, a_2, \dots, a_t)$$
$$= p_t(s_{t+1}|s_t, a_t)$$

$\boxed{\text{Stationary assumption}}$

$$p_t(s_{t+1}|s_t, a_t) = p(s_{t+1}|s_t, a_t)$$

MDP (S, A, r, p)

$\boxed{\text{Planning}}$

Input: MDP

Output: Optimal policy

$\boxed{\text{Reinforcement Learning}}$

Input: Access to the world

Output: actions ($\to$ policy)

Objective

Infinite Horizon

$$\max_{\pi} \mathbb{E}_{\substack{a \sim \pi \\ s \sim p}} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right]$$

$\gamma \in [0, 1)$

$a_0 \sim \pi(a|s_0)$

$r_0 = r(s_0, a_0)$

$s_1 \sim p(s'|s_0, a_0)$

$a_1 \sim \pi(a|s_1)$

$r_1 = r(s_1, a_1)$

$s_2 \sim \quad \ldots$

$\gamma^0 3 + \gamma 3 + \gamma^2 3 + \ldots$

$3 + \gamma 3 + \gamma^2 3 + \ldots$

$= 3/1-\gamma$

$\gamma \to 1$ more patient

$\gamma \to 0$ less patient

$\leftarrow$ example reward always 3

Finite Horizon

$$\max_{\pi} \mathbb{E}_{\substack{a \sim \pi \\ s \sim p}} \left[\sum_{t=0}^{T} r_t \right]$$

Time horizon T

Finite Horizon Planning : Value Iteration

Define $V^*_{(t)}(s)$ = total value from state s under optimal policy with t steps to go

Principle of optimality

An optimal policy consists of:

1. An optimal first action
2. Followed by an optimal policy from the successor state

Value Iteration

Base case $\quad V^*_{(1)}(s) = \max_a r(s,a)$

Inductive case $\quad V^*_{(t+1)}(s) = \max_a \left[r(s,a) + \sum_{s' \in S} p(s'|s,a) V^*_{(t)}(s') \right]$

$\boxed{\text{Computational complexity}}$

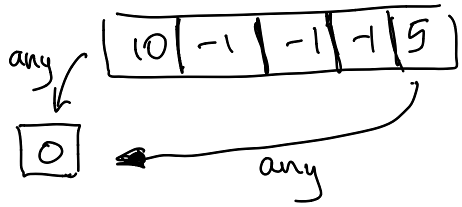
$O(T\,|S|\,|A|\,L)$

L : max # states reachable from any state under any action

$$V^*_{(t+1)}(s) = \max_a \left[r(s,a) + \sum_{s' \in S} p(s'|s,a)\, V^*_{(t)}(s') \right]$$

Example

$$V^*_{(1)}(s) = \max_a r(s,a)$$



rewards for taking any action in a state

Actions: {L, R}

Horizon T=3

$V^*_{(1)}$

0	10	-1	-1	-1	5

$V^*_{(2)}$

0	10	9	-2	4	5

action L: $-1 + 10 = 9$

action R: $-1 - 1 = -2$

$V^*_{(3)}$

0	10	9	8	4	5

action L: $-1 + 9 = 8$

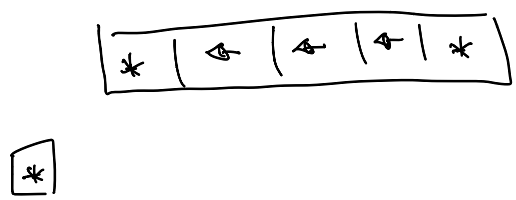
R: $-1 + 4 = 3$

$V^*_{(4)}$

0	10	9	8	7	5

action L: $-1 + 8 = 7$

R: $-1 + 5 = 4$



Policy extraction.

Optimal policy $\pi^*_{(t+1)}(s)$: action to take in state s, t steps to go

$$\arg\max_a \left[r(s,a) + \cdots \right]$$

L = 2 in this gridworld

Infinite Time Horizon Value Iteration

Assume deterministic $\pi(s) \in A$
(w.l.o.g.)

Define MDP value function

$$V^{\pi}(s) = \mathbb{E}_{s \sim p}\left[\sum_{t=0}^{\infty} \gamma^{t} r(s_t, \pi(s_t)) \,\middle|\, s_0 = s \right]$$

expected discounted value from policy
π in state s

Can decompose:

$$V^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s' \mid s, \pi(s)) \, V^{\pi}(s')$$

Define Optimal policy

$$\pi^{*} \in \arg\max_{\pi} V^{\pi}(s) \quad , \text{ for all states}$$

Optimal value function

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

$\hookrightarrow$ $\boxed{\text{Bellman equation + VI. Next lecture!}}$