Lecture 5 - Jan 27, 2023

- Reinforcement Learning Preliminaries

I

- State, Action, Reward, Policy
- Returns and Expected Returns
- State Value Function
- State-Action Value Function
- Bellman Equation and optimality

HW1 → Due Jan 28

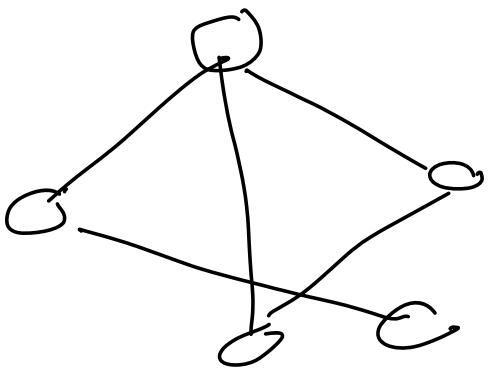
Project 1 → Due Feb 7

TA's office hour: Wednesdays, 12pm - 1pm (in-person)

Fridays, 12pm - 1Am (virtual)

Markov Decision Process (MDP)

$$\langle S, A, T, R \rangle$$



$S = (1\ 1\ 0\ 1\ 1)$

9	10	11	12
8		14	13
7		16	15
6	5		
4	3	2	1

Wall Bump Goal

$A = \{U, D, L, R\}$

→ Transition probability

$T: S \times A \times S$

$P(s' \mid s, a)$

next current

$s = 3,\ a = L \longrightarrow$

$s' = 4$	w.p. 0.9
$s' = 5$	w.p. $\frac{0.1}{3}$
$s' = 2$	w.p. $\frac{0.1}{3}$
$s' = 3$	w.p. $\frac{0.1}{3}$

$$P(s' \mid s=3, a=L) = \begin{cases} 0.9 & s'=4 \\ 0.1/3 & s'=5 \text{ or } 2 \text{ or } 3 \\ 0 & s'=15 \end{cases}$$

$$R: S \times A \times S$$

$$R(s, a, s')$$

9	10	11	12
8		14	13
7		16	15
6	5		
4	3	2	1

Wall Bump Goal

$$R(s=3, a=L, s'=5) = -1$$

$$R(s=3, a=L, s'=4) = -10-1 = -11$$

$$R(s=3, a=U, s'=15) = -1$$

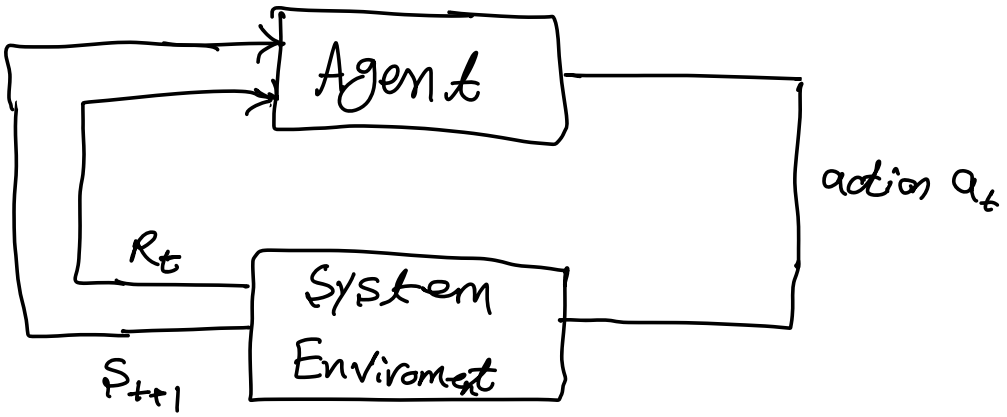
Goal: 100
Bump: -10
movement: -1

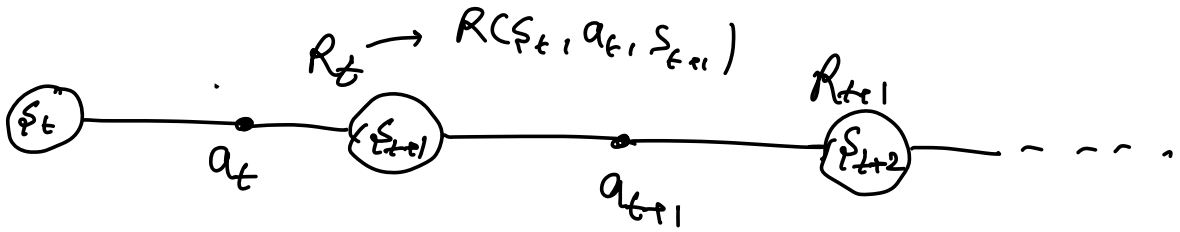
$$P(s_{t+1}=4 \mid s_t=3, a=L, s_{t-1}=2, a_{t-1}=L)$$

$$= P(s_{t+1}=4 \mid s_t=3, a=L) = 0.9$$

9	10	11	12
8		14	13
7		16	15
6	5		
4	3	2	1

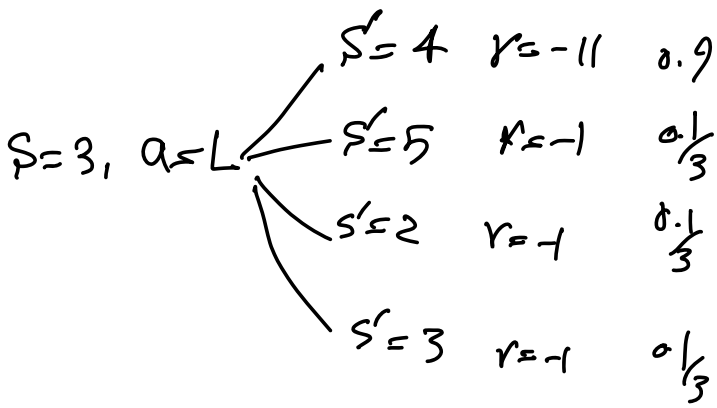
Wall Bump Goal

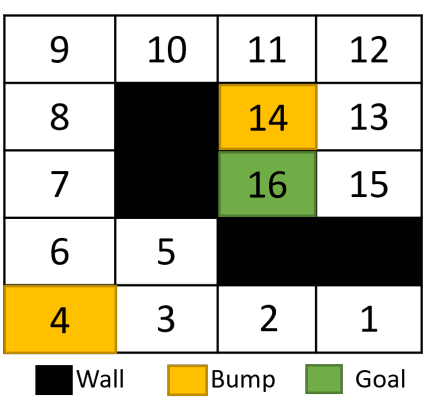




MDP(S, A, T, R)

$P(s', r \mid s, a)$





$P(s'=4, r=-1 \mid s=3, a=L) = 0$

Types of Rewards

Type I : $R(s, a, s')$ Immediate Reward

Type II : $R(s, a)$ Expected Immediate Reward

$$R(s=3, a=L, s') = \begin{cases} -11 & s'=4 & 0.9 \\ -1 & s'=5 & 0.1/3 \\ -1 & s'=2 & 0.1/3 \\ -1 & s'=3 & 0.1/3 \end{cases}$$

$$R(s=3, a=L) = -11 \times 0.9 + (-1) \times \frac{0.1}{3} + (-1) \times \frac{0.1}{3} + (-1) \times \frac{0.1}{3}$$

$$= -10$$

9	10	11	12
8		14	13
7		16	15
6	5		
4	3	2	1

Wall Bump Goal

$P(s', r \mid s, a)$

$P(A, B)$

$P(A) = \sum_{b \in B} P(A, B=b)$

$P(s' \mid s, a) = \sum_{r} P(s', r \mid s, a)$

$\Rightarrow$ Marginalization

$P(r \mid s, a) = \sum_{s'} P(s', r \mid s, a)$

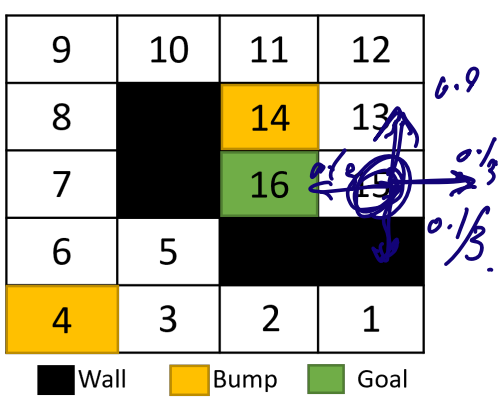
$R(s, a) = E[R(s, a, s') \mid S_t = s, a_t = a]$

$= \sum_{s'} P(s' \mid s, a) \; R(s, a, s')$

$R(s=15, a=U) =$

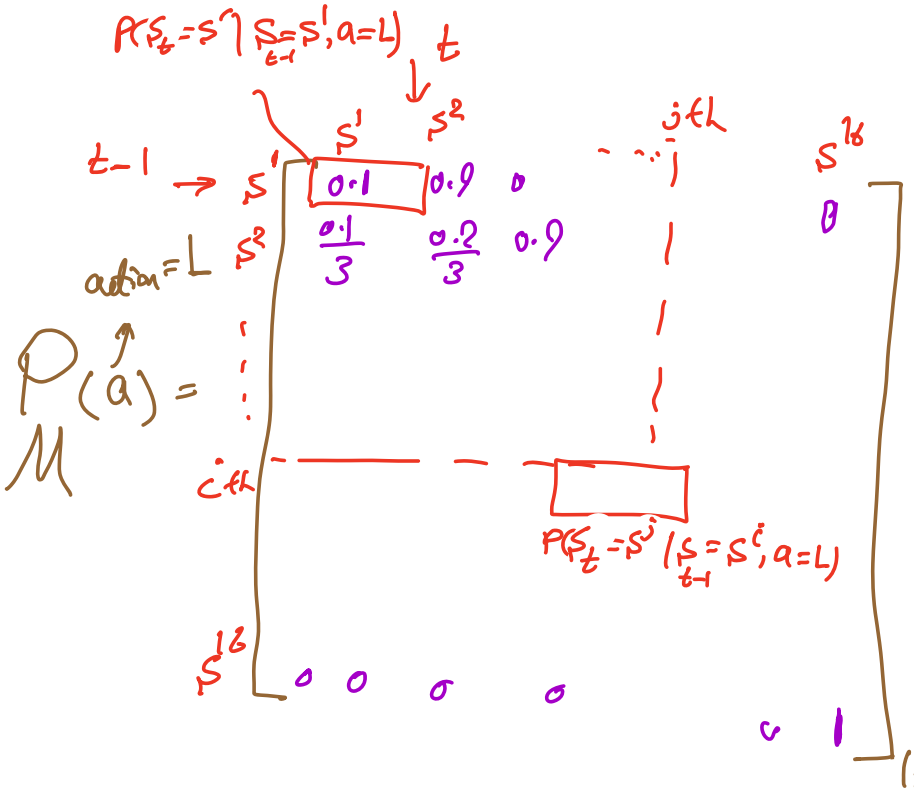
$0.9 \times \frac{s'=13}{-1} + \frac{0.1}{3} \times \frac{s'=15}{-1} + \frac{0.1}{3} \frac{s'=15}{-1}$

$+ \frac{0.1}{3} \times \frac{s'=16}{99}$



Transition probability

- element-wise $P(s'|s,a)$
- Matrix form
  - ↳ Transition Matrix



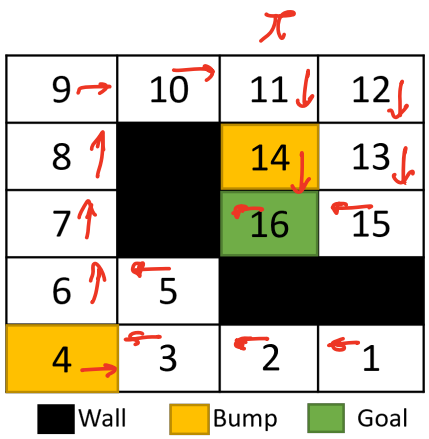
9	10	11	12
8		14	13
7		16	15
6	5		
4	3	2	1

Wall Bump Goal

Policy: Deterministic ⟸ Most common

Stochastic

$\pi : S \rightarrow A$



$\Pi = \{\pi^1, \pi^2, \cdots, \pi^{4^{15}}\}$

$\pi = \begin{bmatrix} \pi(s^1) \\ \pi(s^2) \\ \vdots \\ \pi(s^{16}) \end{bmatrix} = \begin{bmatrix} L \\ L \\ \\ \\ L \end{bmatrix}$

$\pi(s) = U$

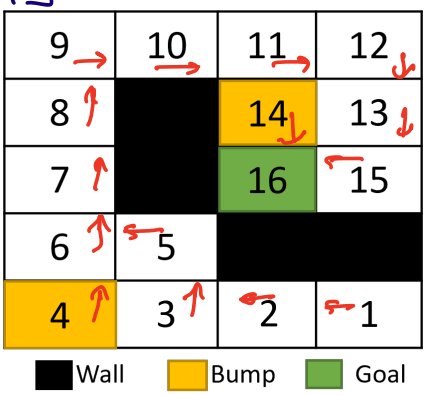
$\pi(a|s) = \begin{cases} a=L & 0.4 \\ a=U & 0.2 \\ a=D & 0.4 \\ a=R & 0 \end{cases}$

↑ stochastic

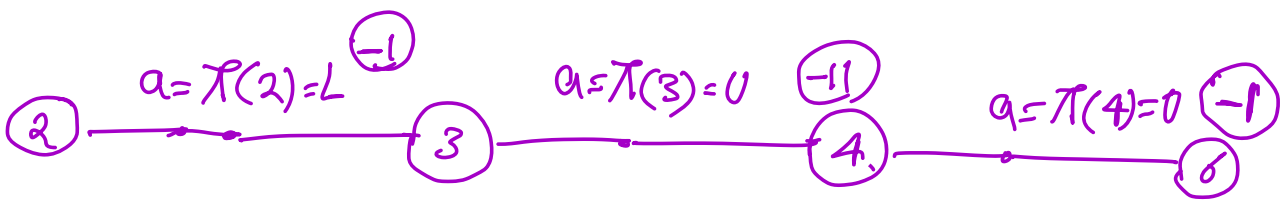
$\pi(s) = a$ ← action that is chosen in state s under policy π

$$\underset{a_{0:T-1}}{\operatorname{argmax}} \; E\left[\sum_{t=0}^{T} R(s_t, a_t, s_{t+1}) \mid s_0 = 2, a_{0:T-1}\right]$$

⇓ Policy



$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmax}} \; E\left[\sum_{t=0}^{T} R(s_t, a_t, s_{t+1}) \mid s_0 = 2, a_{0:T} \sim \pi\right]$$



Return

$\pi \rightarrow R_t \rightarrow R_{t+1} \rightarrow R_{t+2} + \cdots$

(S_t)

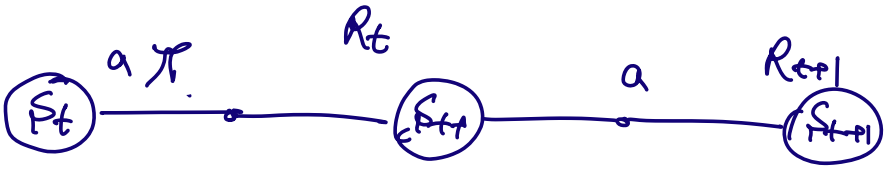
$G_t =$

(S_t)

9	10	11	12
8		14	13
7		16	15
6	5		
4	3	2	1

Wall Bump Goal

- Total Reward G_t: sum of all future rewards in the episode
- Discounted Reward G_t: sum of all future discounted rewards
- Average Reward G_t: Avg Reward per time step



RL Tasks

- Episodic Tasks
- Continuing Tasks

Episodic Tasks:

"Break into episodes"

Total Reward

9	10	11	12
8		14	13
7		16	15
6	5		
4	3	2	1

Wall Bump Goal

$G_t = R_{t+1} + R_{t+2} + \cdots + R_T$ → Terminal point

$E[G_t]$

Episodic