Lecture 6 - Jan 31, 2023

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

Project 1 → Due Feb 7

TA's office hour: Wendsdays, 12pm - 1pm (in-person)
Fridays, 12pm - 1pm (virtual)

$$MDP(S, A, R, P)$$

imediate $\longrightarrow R(s, a, s')$ $P(s' | s, a)$

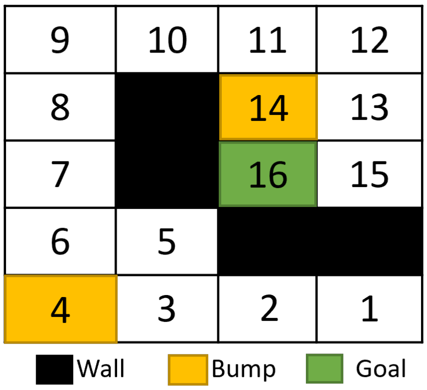
Expected $R(s, a) = \sum_{s'} P(s' | s, a) \, R(s, a, s')$

Task — Episodic

Task — Continuing

Return: Accumulated Reward

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



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

■ Wall 🟨 Bump 🟩 Goal

$\gamma = 1$

$\gamma = 0.9$

$-1 \quad -1 \quad -1 \quad -1 \quad -11 \quad -1 \quad . \; -- \quad 99$

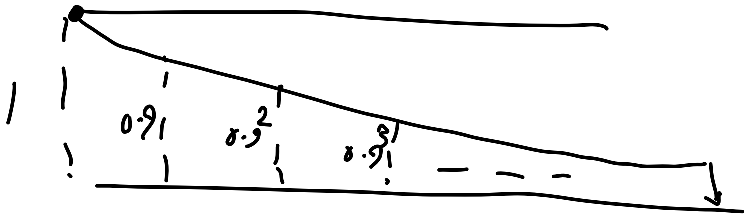
$(0.9)^{20}$

Continuing Task
↳

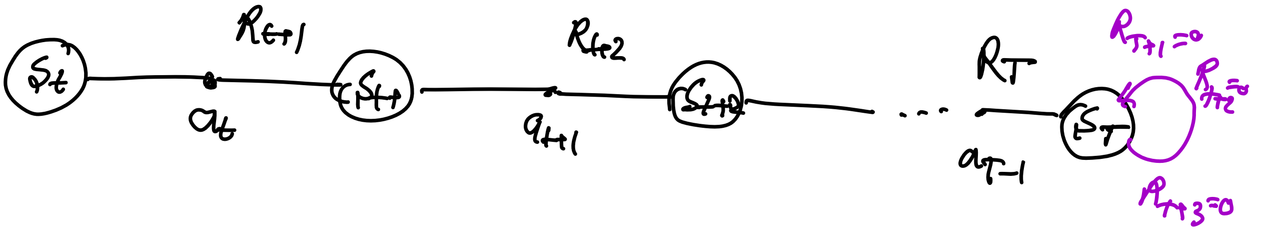
$$G_t = R_{t+1} + \underset{\uparrow}{\gamma} R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

Discount Factor

$\underline{0 < \gamma < 1}$

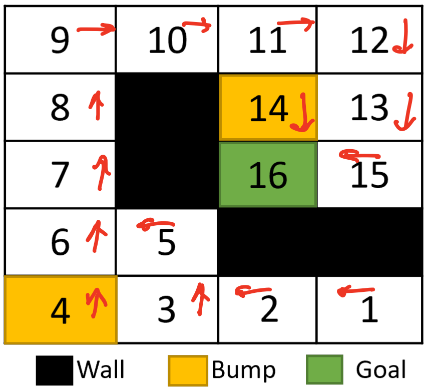


Episodic = Continuing Tasks



$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T} R_T + \gamma^{T+1} R_{T+1} + \gamma^{T+2} R_{T+2} + \cdots$$

$$\pi' = \left[\begin{array}{c} L \\ L \\ L \\ \\ L \end{array}\right] \begin{array}{l} \rightarrow \pi'(s^1) \\ \\ \rightarrow \pi'(s^3) \\ \\ \\ \end{array}$$



9 →	10 →	11 →	12 ↓
8 ↑	■	14 ↓	13 ↓
7 ↑	■	16	15 ←
6 ↑	5 ←	■	■
4 ↑	3 ↑	2 ←	1 ←

■ Wall 🟨 Bump 🟩 Goal

Expected Return = Expected Accomulated Reward

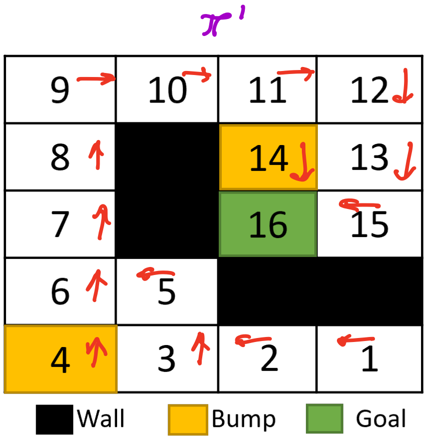
$$V(S_t) = E\left[G_t \mid S_t = 15, \pi' \right]$$

State-Value function

$$V^{\pi'}(15) = E\left[G_t \mid S_t = 15, \pi' \right]$$

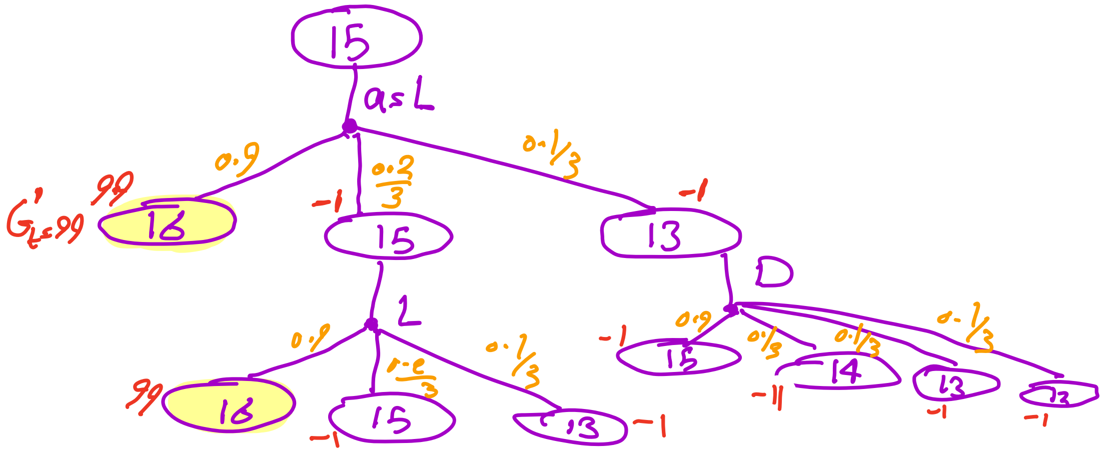
$P(s'|s,a)$

$$P(s'|s=15, a=L) = \begin{cases} 0.9 & s'=16 \\ 2\frac{0.1}{3} & s'=15 \\ \frac{0.1}{3} & s'=13 \end{cases}$$


π'

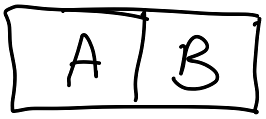
9 →	10 →	11 →	12 ↓
8 ↑	■	14 ↓	13 ↓
7 ↑	■	16	15 ←
6 ↑	5 ←	■	■
4 ↑	3 ↑	2 ←	1 ←

■ Wall 🟨 Bump 🟩 Goal



$$V^{\pi'}(15) = E\left[G_t \mid S_t = 15, \pi'\right]$$

$$= \sum_{\text{trajeltories}} P_t^j \, G_t^j$$



$S = \{A, B\}$

Reward → $\begin{cases} -1 \text{ if } \overset{a^2}{\overline{\text{switch}}} \\ +5 \text{ . be in } B \end{cases}$

$\mathcal{A} = \{a^1, a^2\}$

a^1 → Keep your location

a^2 → switch

$R(S, a, S')$

$R(A, a^1, A) = 0$

$R(A, a^1, B) = 5$

$R(B, a^1, A] = 0$

$\vdots$

$R(B, a^2, B) = 4$

$$P(a^1) = M(a^1) = \begin{array}{cc|cc} & & & k \\ & & A & B \\ \hline k-1 & A & 1 & 0 \\ & B & \boxed{0} & 1 \end{array}$$

$P(S_k = A \mid S_{k-1} = B, \underset{k-1}{a} = a^1)$

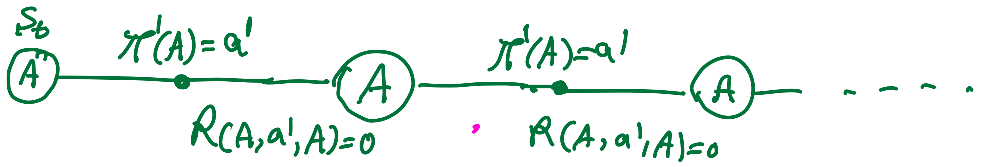
$$P(a^2) = M(a^2) = \begin{array}{cc|cc} & & & k \\ & & A & B \\ \hline & A & 0 & 1 \\ k-1 & B & 1 & 0 \end{array}$$

$\pi^1 = \begin{bmatrix} a^1 \\ a^1 \end{bmatrix}$ $\pi^2 = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix}$ $\pi^3 = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix}$ $\pi^4 = \begin{bmatrix} a^2 \\ a^2 \end{bmatrix}$

$$\pi' = \begin{bmatrix} \pi'(A) \\ \pi'(B) \end{bmatrix} = \begin{bmatrix} a' \\ a' \end{bmatrix} \leftarrow \text{Stay at your state} \qquad \gamma = 0.9$$

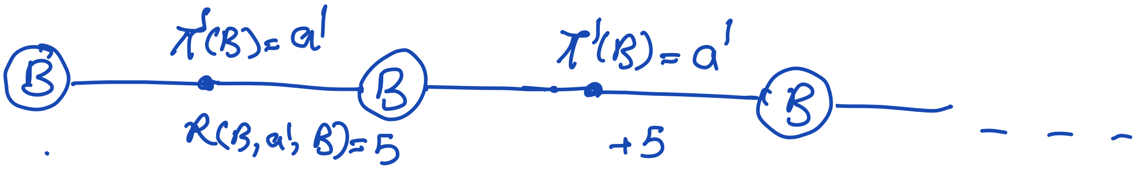
$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$V_{\pi'}(s) = E[G_t \mid S_t = s, \pi']$$



$$\underbrace{V_{\pi'}(A)} = 0 + \gamma 0 + \gamma^2 0 + \cdots = 0$$

Expected Accumulated Rewards starting from A and following policy π'



$$V_{\pi'}(B) = E[R_{t+1} + \gamma R_{t+2} + \cdots \mid S_t = B, \pi']$$

$\gamma = 0.9$

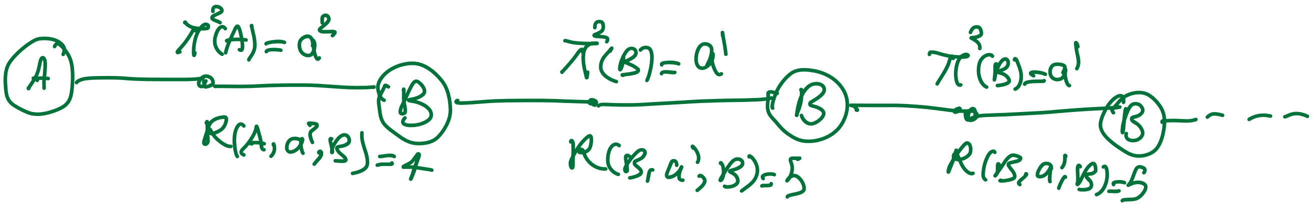
$$= 5 + \gamma 5 + \gamma^2 5 + \gamma^3 5 + \cdots$$

$$= 5(1 + \gamma + \gamma^2 + \gamma^3 + \cdots)$$

$$= 5 \underbrace{\frac{1}{1-\gamma}} = \frac{5}{1-0.9} = 50$$

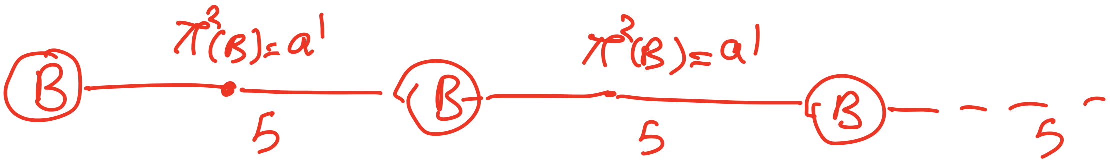
$$1 + x + x^3 \cdots = \frac{1}{1-x}$$
$$0 < x < 1$$

$$\pi^2 = \begin{bmatrix}\pi^2(A) \\ \pi^2(B)\end{bmatrix} = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix}$$



$$V_{\pi^2}(A) = 4 + \gamma 5 + \gamma^2 5 + \gamma^3 5 + \cdots$$

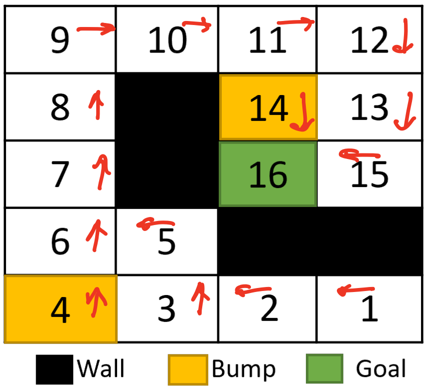
$$= 4 + \gamma 5 \underbrace{(1 + \gamma + \gamma^2 + \cdots)}_{\frac{1}{1-\gamma}} = 49$$



$$V_{\pi^2}(B) = 5 + \gamma 5 + \gamma^2 5 + \cdots = 5(1 + \gamma + \gamma^2 + \cdots) = 50$$

$$\pi' = \begin{bmatrix} \pi'(1) \\ \vdots \\ \pi'(15) \end{bmatrix} = \begin{bmatrix} L \\ L \\ U \\ \vdots \\ L \end{bmatrix}$$

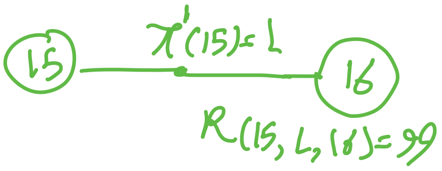
Deterministic



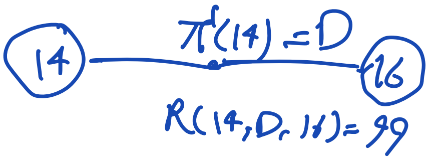
9 →	10 →	11 →	12 ↓
8 ↑	■	14 ↓	13 ↓
7 ↑	■	16	15 ←
6 ↑	5 ←	■	■
4 ↑	3 ↑	2 ←	1 ←

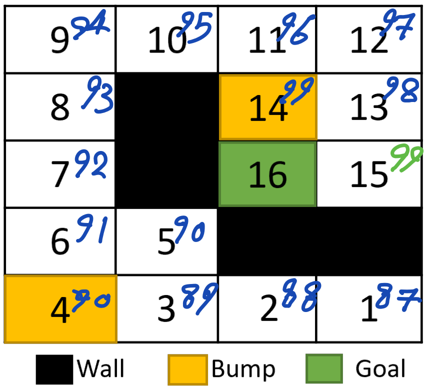
■ Wall ▨ Bump 🟩 Goal

$V_{\pi'}(15) = 99$



$V_{\pi'}(14) = 99$

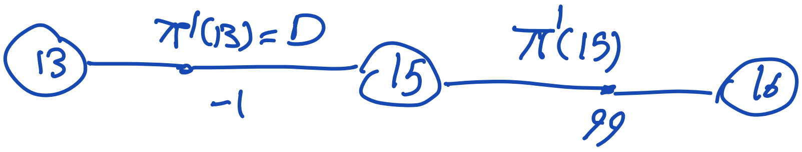




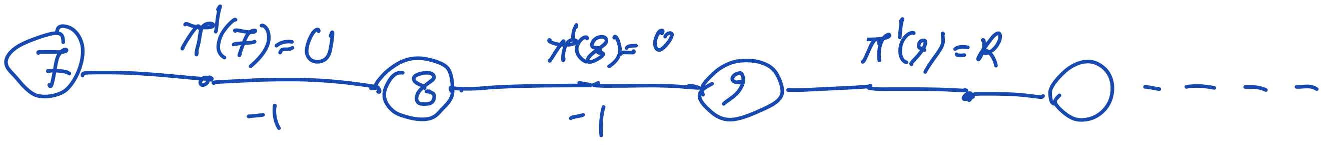
9 94	10 95	11 96	12 97
8 93	■	14 99	13 98
7 92	■	16	15 99
6 91	5 90	■	■
4 90	3 89	2 88	1 87

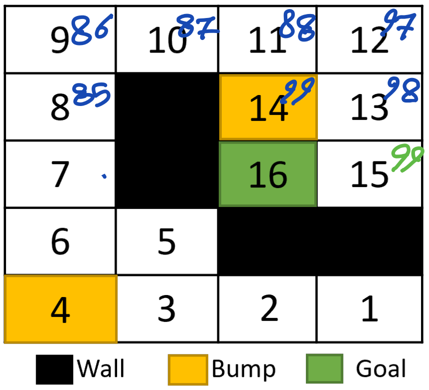
■ Wall ▨ Bump 🟩 Goal

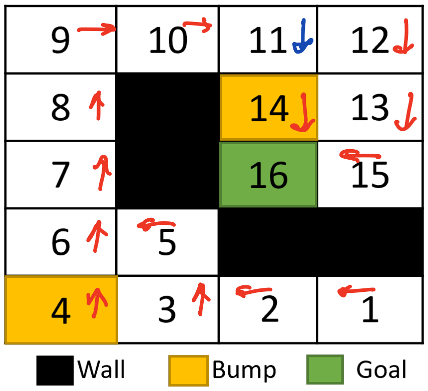
$V_{\pi'}(13) = -1 + 99 = 98$

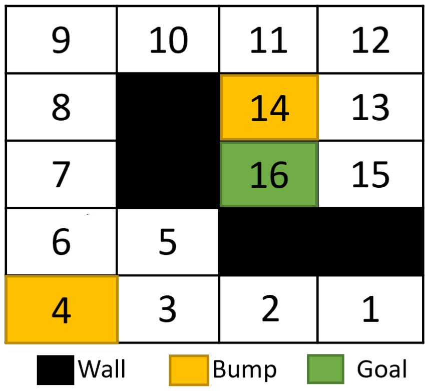


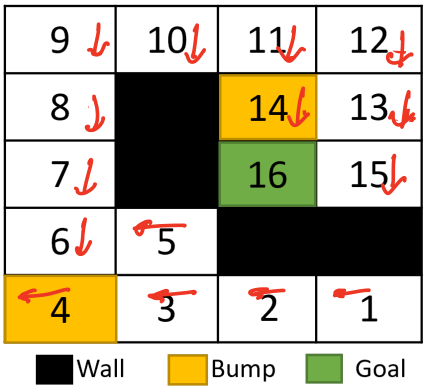
$V_{\pi'}(7) =$




9 86
10 87
11 88
12 97
8 85
14 99
13 98
7
16
15 99
6
5
4
3
2
1
Wall
Bump
Goal


9 →
10 →
11 ↓
12 ↓
8 ↑
14 ↓
13 ↓
7 ↑
16
15 ←
6 ↑
5 ←
4 ↑
3 ↑
2 ←
1 ←
Wall
Bump
Goal

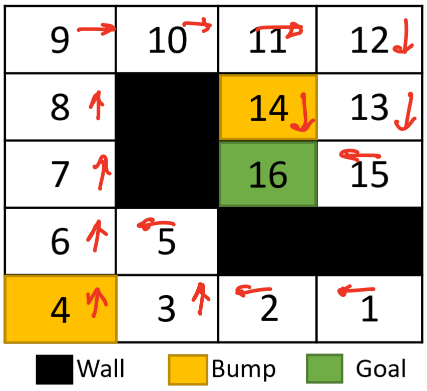

9
10
11
12
8
14
13
7
16
15
6
5
4
3
2
1
Wall
Bump
Goal


9 ↓
10 ↓
11 ↓
12 ↓
8 ↓
14 ↓
13 ↓
7 ↓
16
15 ↓
6 ↓
5 ←
4 ←
3 ←
2 ←
1 ←
Wall
Bump
Goal

State-Action Value Function

$$Q_\pi(s, a) = E[G_t \mid S_t = s, a_t = a, \pi]$$

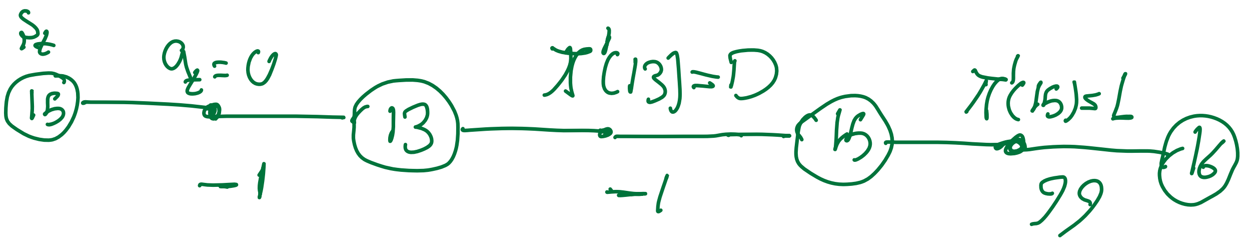
$$\pi' = \begin{bmatrix} L \\ L \\ U \\ \vdots \\ L \end{bmatrix}$$



9 →	10 →	11 →	12 ↓
8 ↑	■	14 ↓	13 ↓
7 ↑	■	16	← 15
6 ↑	← 5	■	■
4 ↑	3 ↑	← 2	← 1

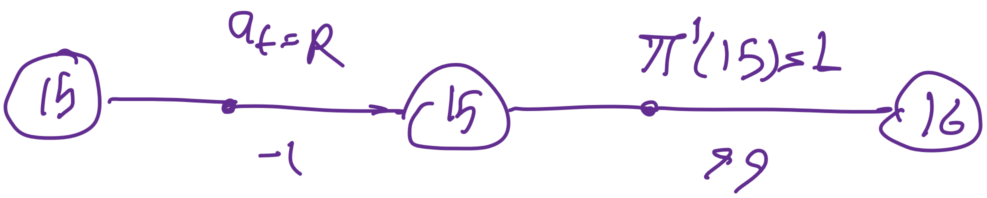
■ Wall 🟨 Bump 🟩 Goal

$Q_{\pi'}(15, U) =$



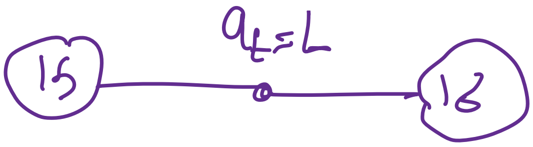
$$Q_{\pi'}(15, U) = -1 \; -1 \; + 99 = 97$$

$$Q_{\pi'}(15, R) = -1 + 99 = 98$$



$$Q_{\pi^1}(15, D) = -1 + 99 = 98$$

$$Q_{\pi^1}(15, L) =$$



$$Q_{\pi}(S, \pi(S)) = V_{\pi}(S)$$

A	B

$\pi^1 = \begin{bmatrix} a1 \\ a1 \end{bmatrix}$ $\pi^2 = \begin{pmatrix} a1 \\ a2 \end{pmatrix}$ π^3 π^4

$\overset{0}{V_{\pi^1}(A)}$ $\overset{50}{V_{\pi^2}(A)}$ $V_{\pi^3}(A)$ $V_{\pi^4}(A)$

$V_{\pi^1}(B)^{49}$ $V_{\pi^2}(B)^{50}$ $V_{\pi^3}(B)$ $V_{\pi^4}(B)$