Lecture 7 - Feb 3, 2023

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

Project 1 → Due Feb 7

TA's office hour: Wendsdays, 2pm-3pm (in-person)
Fridays, 2pm-3pm (virtual)

$$MDP(S, A, R, T, \gamma)$$

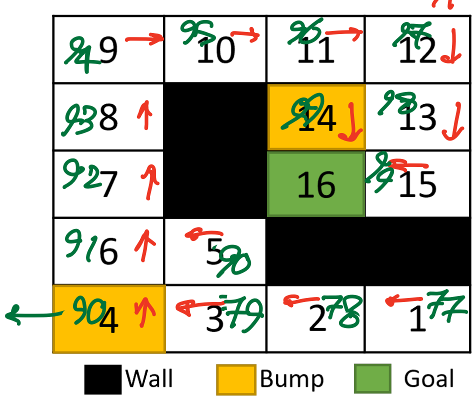
$$T \rightarrow P(s' \mid s, a)$$

$$P(a) = M(a) = \begin{bmatrix} \; & \; \\ & \square \\ \; & \; \end{bmatrix}$$

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

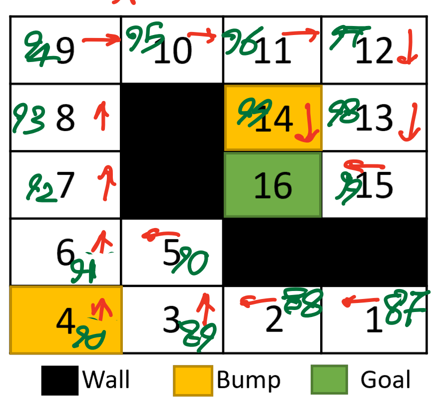
$|A|^{|S|}$ ← diff Policy

$\pi^2$

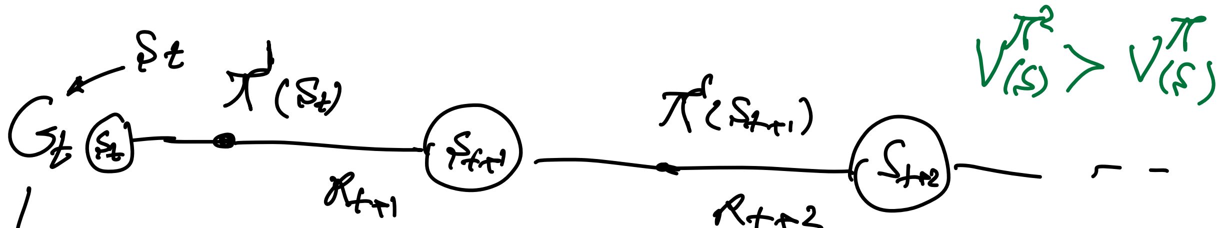



$V^{\pi^2}(4)$

$\pi^1$

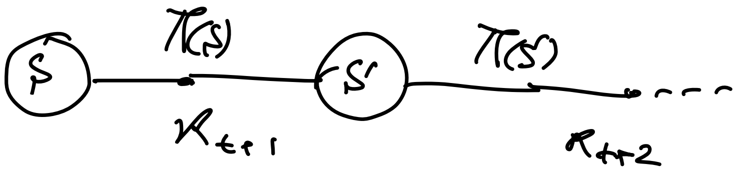


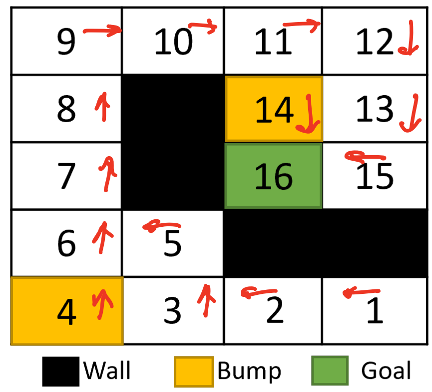

$$V^{\pi^2}(s) > V^{\pi^1}(s)$$

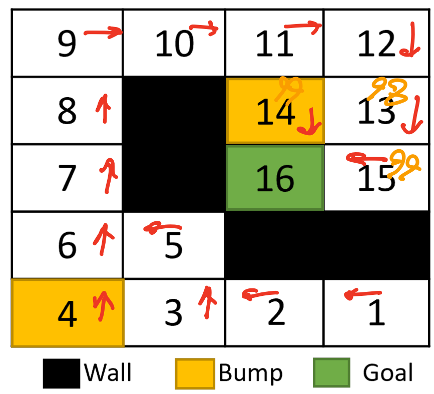



$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$V^{\pi}_{\pi}(s) = E[\,G_t \mid S_t = s, \pi\,]$$



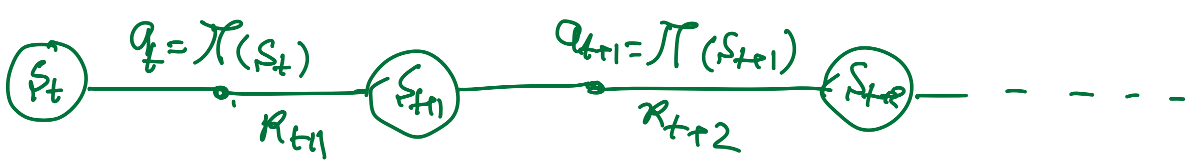




$\pi \geq \pi'$ if and only if

$V_{\pi}(s) \geq V_{\pi'}(s)$ for all $s \in S$

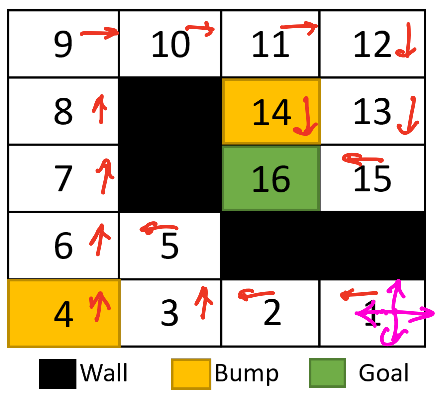
$$V_{\pi}(s) = E[\,G_t \mid S_t = s, \pi\,]$$



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

$$S_t \xrightarrow[R_{t+1}]{a_t = a} S_{t+1} \xrightarrow[R_{t+2}]{a_{t+1} = \pi(S_{t+1})} S_{t+2} \; - - - -$$

$$Q_\pi(s=15, a=U) = 97$$




$$15 \xrightarrow[-1]{U} 13 \xrightarrow[-1]{D} 15 \xrightarrow[99]{R} 16$$

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

$$Q_\pi(15, D) = 98$$

$$Q_\pi(15, L) = 99$$

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

$$V_\pi(s) = E\left[G_t \mid s_t = s,\ a_t = \pi(s),\ a_{t+1:\infty} \sim \pi\right]$$

$$Q_\pi(s,a) = E\left[G_t \mid s_t = s,\ a_t = a,\ a_{t+1:\infty} \sim \pi\right]$$

| $V_{\pi^1}$ | $V_{\pi^2}$ |
|---|---|
| 10 | 10 |
| 9 | 10 |
| 5 | 9 |
| 6 | 4 |
| 7 | 7 |
| 11 | 12 |

$\cancel{V_{\pi^1}(s) \geq V_{\pi^2}(s)}$

→ It is not comparable

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

$\pi^*$ optimal policy

$$\underbrace{V^{\pi^*}(s) \geq V^{\pi}(s) \quad \text{for all } s \quad \pi \in \Pi}$$

$$V_{\pi^*}(s) = V^*(s) = \max_{\pi \in \Pi} V_\pi(s), \text{ for all } s \in S$$

$$Q_{\pi^*}(s,a) = Q^*(s,a) = \max_{\pi \in \Pi} Q_\pi(s,a)$$

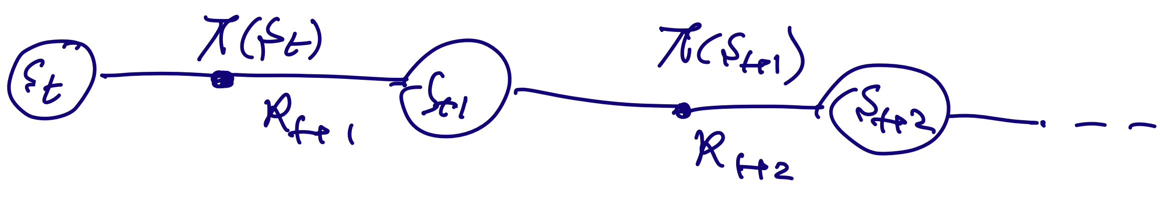
for all $s, a$

---

## Bellman Equation

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

Policy $\pi$



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

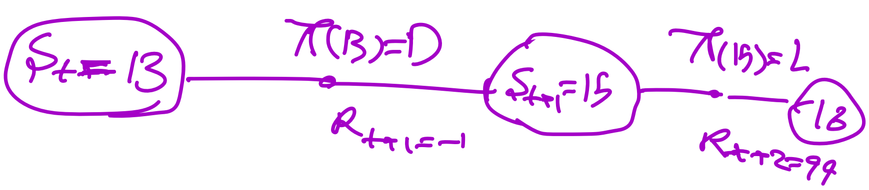
$$= E[R_{t+1} + \gamma \underbrace{R_{t+2} + \gamma^2 R_{t+3} + \dots}_{} \mid S_t = s, \pi]$$

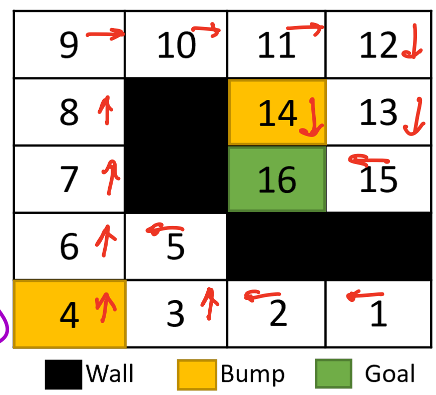
$$= E[R_{t+1} + \gamma \underbrace{(R_{t+2} + \gamma R_{t+3} + \dots)}_{G_{t+1}} \mid S_t = s, \pi]$$

$$V_\pi(s) = E[R_{t+1} + \gamma \underbrace{G_{t+1}}_{V_\pi(S_{t+1})} \mid S_t = s, \pi]$$

Bellman Eq.

$$V_\pi(s) = E[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, \pi]$$



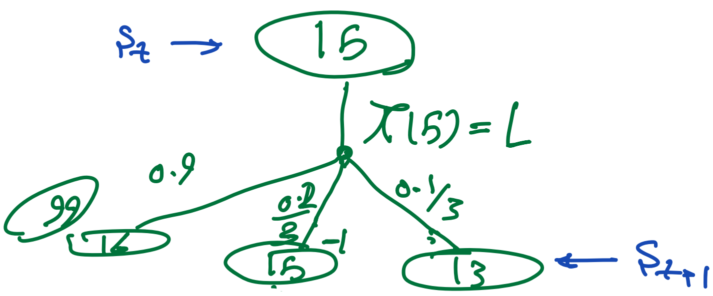



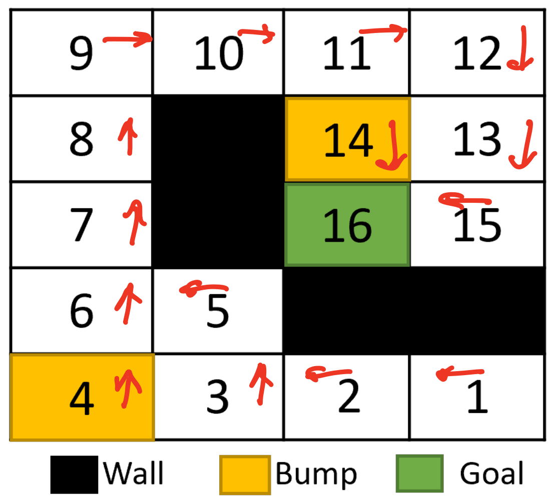
$V_\pi(15) = -1 + 100 = 99$

$V_\pi(13) = E[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = 13, \pi]$

$= -1 + \gamma V_\pi(15)$

## Stochastic Environmt




$V_\pi(s) = E[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, \pi]$

$E[G_t \mid S_t = s, \pi]$

$= \sum p^{traj} G_t^{traj}$

$= P(s' \mid S_t = s, a_t = \pi(s_t))$

| $S_{t+1}$ | | | $R_{t+1}$ |
|---|---|---|---|
| 16 | $\leftarrow$ | $0.9$ | 99 |
| 15 | $\leftarrow$ | $\frac{0.2}{3}$ $\rightarrow$ | $-1$ |
| 13 | $\leftarrow$ | $\frac{0.1}{3}$ | $-1$ |

$$V_\pi(15) = E\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, \pi\right]$$

$$\boxed{V_\pi(s) = \sum_{s'} P(s' \mid s, \pi(s)) \left[R(s, \pi(s), s') + \gamma V_\pi(s')\right]}$$

$$V_\pi(15) = \underbrace{P(s'=16 \mid s=15, \pi(15)=L)}_{0.9} \left[\underbrace{R(15, L, 16)}_{99} + \gamma \underbrace{V_\pi(16)}_{0}\right]$$

$$+ \underbrace{P(s'=15 \mid s=15, \pi(15)=L)}_{\cancel{0.2}\,/3} \left[\underbrace{R(15, L, 15)}_{-1} + \gamma V_\pi(15)\right]$$

$$+ \underbrace{P(s'=13 \mid s=15, \pi(15)=L)}_{0.1/3} \left[\underbrace{R(15, L, 13)}_{-1} + \gamma V_\pi(13)\right]$$

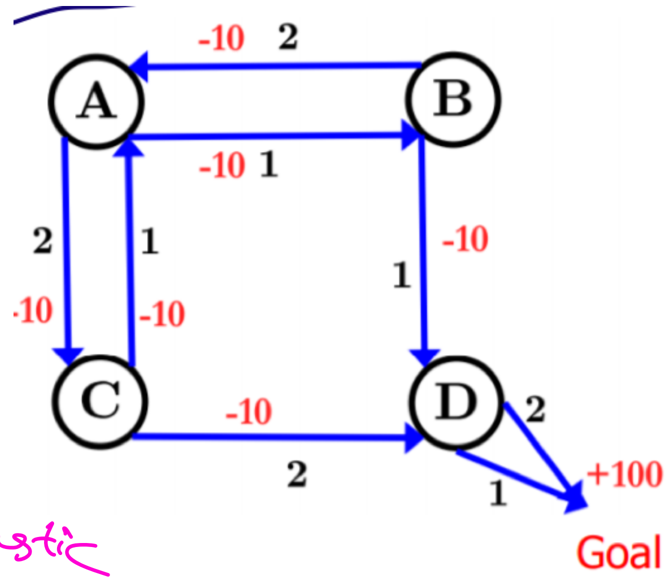
Bellman Eq

$$V_\pi(s) = E\left[R_{t+1} + \gamma V_\pi(s') \mid S_t = s, \pi\right]$$

$$= \sum_{s'} P(s' \mid s, \pi(s)) \left[\underset{\substack{\downarrow \\ R(s,\, \pi(s),\, s')}}{R_{t+1}} + \gamma V_\pi(s')\right]$$

$S = \{A, B, C, D\}$

$A = \{1, 2\}$




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

$P(s' = D \mid s = B, a = 1) = 0.9$

$P(s' = A \mid s = B, a = 1) = 0.1$

$$\pi = \begin{bmatrix} \pi(A) \\ \pi(B) \\ \pi(C) \\ \pi(D) \end{bmatrix} \begin{matrix} 2 \\ 2 \\ 2 \\ 2 \end{matrix}$$

1 1 1 2 $\quad 2^4 = 16$

1 2 1 2

2

1 2 2

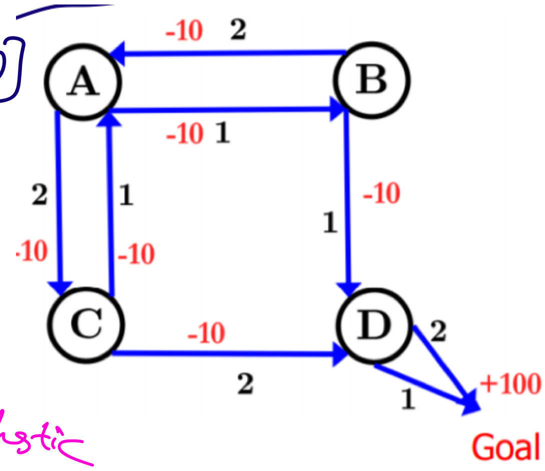
1 1 2 2

---

$$\pi^1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \rightarrow$$

$$V_\pi(s) = \sum_{s'} P(s' \mid s, \pi(s)) \left[ R + \gamma V_\pi(s') \right]$$

$$V_{\pi'}(A) = \sum_{s'} P(s' \mid s=A, \pi'(A)=1) \, [R + \gamma V_{\pi'}(s')]$$




$$A \xrightarrow{a=1} \begin{cases} B & 0.9 & -10 \\ C & 0.1 & -10 \end{cases}$$

chstic

$$V_{\pi'}(A) = \underbrace{P(s'=B \mid s=A, \pi'(A)=1)}_{0.9} \, [-10 + \gamma \cdot V_{\pi'}(B)]$$
$$+ \underbrace{P(s'=C \mid s=A, \pi'(A)=1)}_{0.1} \, [-10 + \gamma V_{\pi'}(C)]$$

$$V_{\pi'}(B) = 0.9 \, [-10 + \gamma V_{\pi'}(D)] + 0.1 \, [-10 + \gamma V_{\pi'}(A)]$$

$$V_{\pi'}(C) = 0.9 \, [-10 + \gamma V_{\pi'}(A)] + 0.1 \, [-10 + \gamma V_{\pi'}(D)]$$

$$V_{\pi'}(D) = 100$$

↳ Four Equations + Four Variables

$V_{\pi'}(A) = x$ $\quad$ $V_{\pi'}(B) = y$ $\quad$ $V_{\pi'}(C) = z$ $\quad$ $V_{\pi'}(D) = d$

$$x = 0.9\left[-10+0.9\,y\right] + 0.1\left[-10+0.9\,z\right]$$

$y = .9$ ___ ___

$z$ ___ ___

$d$ ___ ___

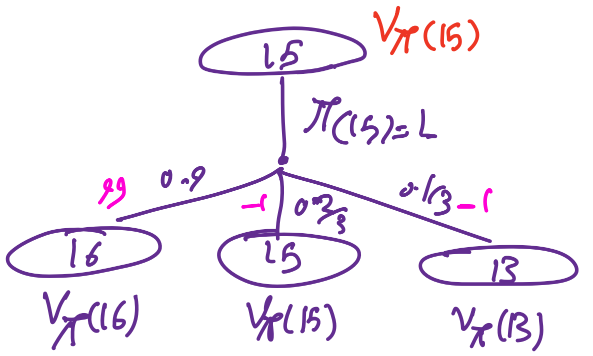
$$\begin{cases} x - 0.81\,y - 0.09\,z + 0\,d = -10 \\ -0.09\,x + y + 0\,z - 0.81\,d = -10 \\ -0.09\,x + 0\,y + z - 0.09\,d = -10 \\ 0\,x + 0\,y + 0\,z + 1\,d = 100 \end{cases}$$

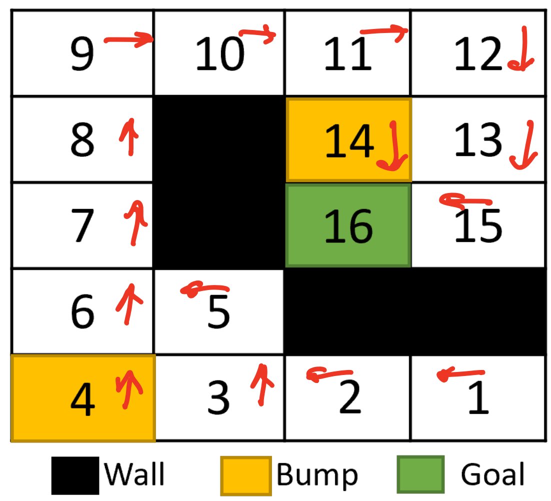
$$A \begin{bmatrix} x \\ y \\ z \\ d \end{bmatrix} = a \rightarrow \begin{bmatrix} -10 \\ -10 \\ -10 \\ 100 \end{bmatrix} \longrightarrow A^{-1}a$$

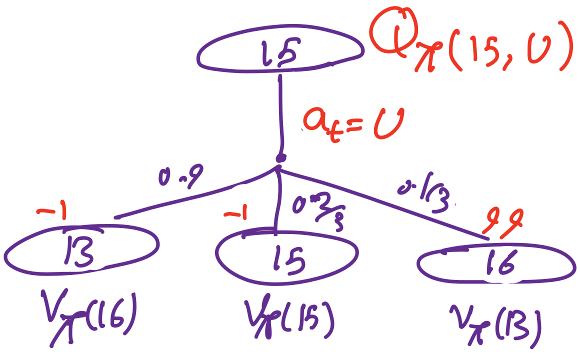
$$\pi' = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \longrightarrow V_{\pi'} = \begin{bmatrix} 75 \\ 87 \\ 68 \\ 100 \end{bmatrix}$$

$$\pi^2 = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \end{bmatrix} \rightarrow V_{\pi^2} = \begin{bmatrix} 75 \\ 68 \\ 87 \\ 100 \end{bmatrix}$$

not comparable




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

$$= E\left[\underbrace{R_{t+1}}_{a} + \gamma \underset{\pi}{R_{t+2}} + \gamma^2 \underset{\pi}{R_{t+3}} + \cdots \mid S_t = s, a_t = a, \pi\right]$$

$$= E\left[R_{t+1} + \gamma \underbrace{\left(R_{t+2} + \gamma R_{t+3} + \cdots\right)}_{\substack{G_{t+1} \\ \downarrow \\ V_\pi(S_{t+1})}} \mid S_t = s, a_t = a, \pi\right]$$

$$= E\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, a_t = a, \pi\right]$$

$$= \sum_{s'} P(s' \mid s, a) \left[R(s, a, s') + \gamma V_\pi(s')\right]$$