Lecture 13 - Feb 24, 2023

- Dynamic Programming
    - Policy Iteration
    - Value Iteration
- Approximate Dynamic Programming
    - Asynchronous DP
        - In-Place DP
        - Priorotized Sweeping
        - Real-Time DP
    - Generalized Policy Iteration
- Monte-Carlo Methods
    - First-Visit MC
    - Online MC

Model-Based

Model-Free

HW3 is Posted → Due March 17

Project 2 → Due March 3

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

Approximate Dynamic Programming

Matrix-Form

$$S = \{10{,}000{,}000\} \qquad \underline{M(a)}$$

$$\boxed{A \mid B} \qquad \begin{array}{c|cc} & A & B \\ \hline A & & \\ B & & \end{array}$$

$$V_K(s) = \max_{a \in A} \sum_{s'} P(s' \mid s, a)\left[R(s, a, s') + \gamma V_{k-1}(s')\right]$$

10,000,000

10,000,000

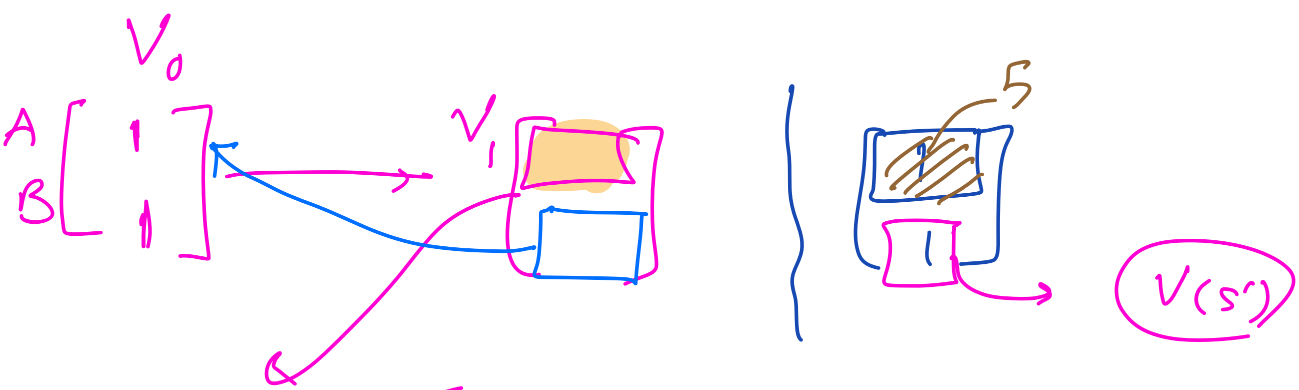
$$V_0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{10{,}000{,}000} \xrightarrow{\text{VIB}} V_1 = \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}_{10{,}000{,}000}$$

In-place Dynamic programming

(Space-Eficient DP)

V_{old} $\uparrow$ $|S|$ $\quad$ V_{New} $\uparrow$ $|S|$ $\longrightarrow$ using $\cdot V$

$$V(s) = \max_{a \in A} \sum_{s'} P(s'|s,a)\left[R + \gamma V(s')\right]$$



$$V_1(s) = \max \sum P(s'|s,a)\left[R + \gamma V_0 \underline{(s')}\right]$$

$$V_1(s) = \max \sum P(s'|s,a)\left[R + \gamma V_0 (s')\right]$$

$$\pi^*(s) = \arg\max_a \sum P(s'|s,a)\left[R + \gamma V^*(s')\right]$$

Approximate DP ← Warren Powell

$\pi \xrightarrow{PE} V^{\pi}$

Prioritized Sweeping

Do I need to Backup for all $s \in S$.

$$V_{k+1}(s) = \max_{a \in A} \sum_{s'} P(s'|s,a)\left[r + \gamma V_k(s')\right]$$

$$|V_{k+1}(s) - V_k(s)|$$

$\Delta \leftarrow$ priority $\quad \Delta = [\;1 \quad\quad\quad 1\;]_{|S|}$

$s_k \sim \Delta$

$$V^{new}(s_k) = \max_{a \in A} \sum P(s'|s_k,a)\left[R + \gamma V^{old}(s')\right]$$

$$\Delta(s_k) = |V^{new}(s_k) - V^{old}(s_k)|$$

$$\Delta(s) = \Delta(s) + \gamma \max_{a \in A} P(s_k \mid s, a)\,\Delta(s_k)$$ $\Leftarrow$ for all s that lead to s_k

$\Delta = [1 \ 1 \ 1 \ 1 \ - - - \ 1]$

$\Delta(S_{13})$

$S_K \sim [\frac{1}{15}, \frac{1}{15} \ - - \ \frac{1}{15}]$

$\sum \Delta$

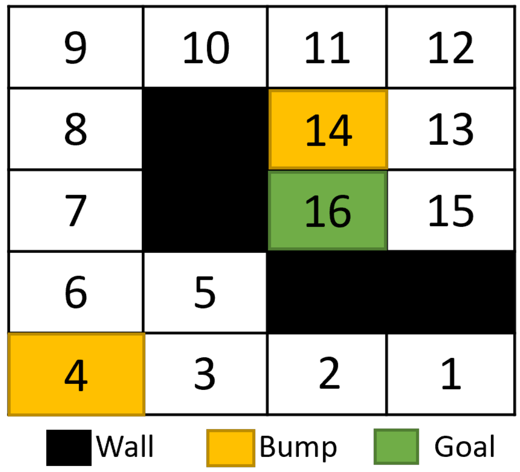
$S_K = 14$

new 14

$V^{new}(S_K) = \max_{a \in A} \sum P(s'|S_K, a)[R + \gamma V^{old}(s')]$

$\max \{ \overset{\text{up}}{-1+\gamma 0}, \overset{\text{Down}}{99+\gamma 0}, \overset{L}{-11}, \overset{R}{-1} \} = 99$

$\Delta(\overset{S_K}{14}) = |\underbrace{V^{new}(14)}_{99} - \underbrace{V^{old}(14)}_{0}| = 99$



1

$\Delta(11) = \Delta(11) + \gamma \underbrace{\max_{a \in A}}_{\text{Down}} \underbrace{P(\overset{S_K}{14} | \overset{S}{11}, a)}_{1} \underbrace{\Delta(14)}_{99} = 100$

$\Delta(13) = \underbrace{\Delta(13)}_{1} + \gamma \max_{\substack{a \in A \\ \text{Left}}} \underbrace{P(14 | 13, a)}_{1} \underbrace{\Delta(14)}_{99} = 100$

$\Delta = \left[1 \quad 1 \quad \overset{11}{100} \quad \overset{12}{1} \quad \overset{13}{100} \quad \overset{14}{99} \quad \overset{15}{1} \right]$

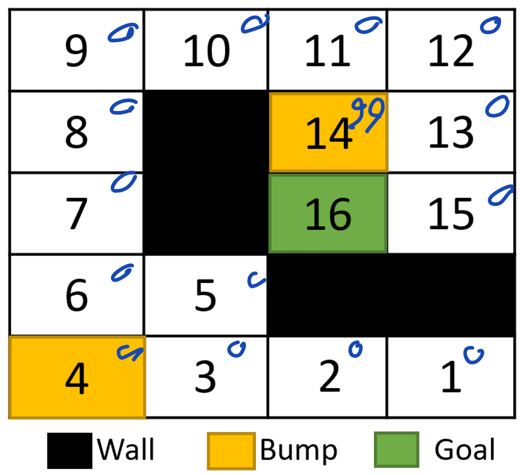
$S_k \leftarrow \Delta$

$S_k = 13$

$$V^{new}(13) = \max \sum_{s'} p(s'|s_k,a)\left[R + \gamma V^{old}(s')\right]$$

$$= \max\{-1, -1, 88, -1\}$$

$$= 88$$



$$\Delta(\overset{13}{S_k}) = \left| V^{old}(13) - V^{new}(13) \right|$$

$$= 88$$

$$\Delta(12) = \Delta(12) + \gamma \max \underbrace{P(13|12,a)}_{1} \underbrace{\Delta(13)}_{88}$$

$$= 1 + 1 \times 88 = 89$$

$$\Delta(15) = \underbrace{\Delta(15)}_{1} + \underbrace{\gamma}_{-} \max \underbrace{P(13|15,a)}_{1} \underbrace{\Delta(13)}_{88} = 89$$

$$\Delta(14) = \underbrace{\Delta(14)}_{99} + \gamma \max P(13|14,a) \underbrace{\Delta(13)}_{88} = 187$$

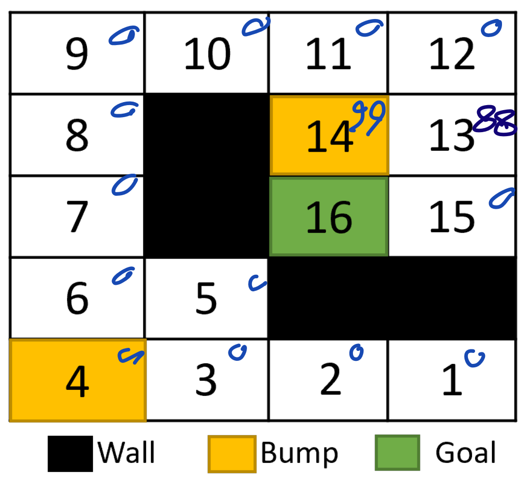
$$\Delta = \begin{array}{cccccccc} 1 & 2 & & 11 & 12 & 13 & 14 & 15 \\ \hline 1 & 1 & \cdot & 100 & 89 & 88 & 187 & 89 \end{array}$$

$s_k \sim \Delta$

$s_k = 11$

$V^{new}(11) = \max\{-1, -1, -1, 88\} = 88$

$\Delta(11) = |0 - 88| = 88$



$\Delta =$

Real-Time DP

$$V(s_k) = \max \sum_{s'} P(s'|s_k,a)\left[R(s_k,a,s') + \gamma V(s')\right]$$

$$a_k = \arg\max \sum_{s'} P(s'|s_k,a)\left[R + \gamma V(s')\right]$$

$$\begin{array}{ccccc} V_0 & & V_1 & & \\ S_0 & \xrightarrow{a \sim VIB} & S_1 & \xrightarrow{a \sim VIB} & S_2 \rightsquigarrow \end{array}$$

Generalized Policy Iteration (GPI)

$$\pi_0 \xrightarrow{PE} V^{\pi_0} \xrightarrow{PI} \pi_1 \xrightarrow{PE} \pi_2$$

$$V_0 \rightsquigarrow V_1 \rightsquigarrow V_2 \cdots V_{T-1} \rightsquigarrow V_T \xrightarrow{\pi'} V_0 \rightsquigarrow V_1 \rightsquigarrow V_2 \cdots$$

$$V_{k+1}(s) = \sum_{s'} P(s'|s,\pi_0(s))\left[R + \gamma V_k(s')\right]$$

$$\pi'(s) = \arg\max_{a \in A} \sum P(s'|s,a)\left[R + \gamma V^{\pi_0}(s')\right]$$

$V_0 \rightsquigarrow V_1 \rightsquigarrow V_2$ — —

$$V_{k+1}(s) = \max \sum_{s'} P(s'|s,a)\left[R + \gamma V_k(s')\right]$$

one-step PE + PI = Value Iteration

GPI →

$V_0 \longrightarrow V_1 \longrightarrow V_2 \xrightarrow{PI} V_3 \longrightarrow V_4 \longrightarrow V_5 \xrightarrow{PI}$

Finger grain switch between PE & PI

Monte-Carlo Methods

$$MDP(S, A, R, \cancel{T})$$

$P(s'|s,a)$

$$s \xrightarrow[P(s'|s,a)]{a} s'$$

$$V_{k+1}(s) = \max \sum_{s'} \cancel{P(s'|s,a)} [R + \gamma V(s')]$$

$\pi(s_0)$ $\pi(s_1)$

$R(s_0, a, s_1)$ $R(s_1, a, s_2)$ $R(s_2, \pi(s_2), s_3)$

$$V^{\pi}(s) = E\Big[\underbrace{R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots}_{G_t} \,\Big|\, s_0 = s, a \sim \pi\Big]$$

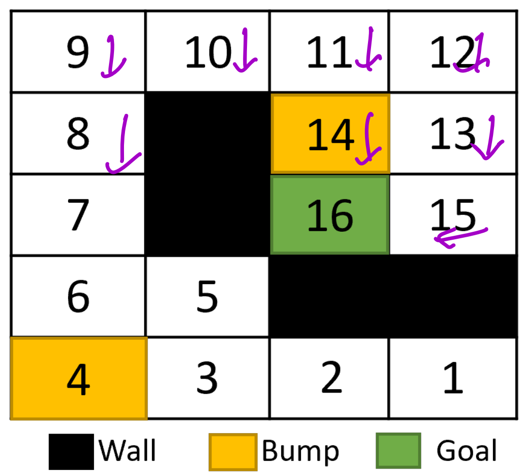
$$= E\Big[R_{t+1} + \gamma \underbrace{(R_{t+2} + \gamma R_{t+3} + \cdots)}_{V^{\pi}(s')} \,\Big|\, s_0 = s, a \sim \pi\Big]$$

$$= E\left[R_{t+1} + \gamma V^{\pi}(s')\right]$$

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

$$V^\pi(s) = E[G_t \mid s_t = s, \pi]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} G_t^i$$



$$13 \xrightarrow[R=-1]{\pi(13)} 15 \xrightarrow[99]{\pi(15)} \qquad G_t^1 = -1 + \gamma 99 = 98$$

$$13 \xrightarrow[-11]{\pi(13)} 14 \xrightarrow[99]{\pi(14)} G \qquad G_t^2 = -11 + \gamma 99 = 88$$