Lecture 8 - Feb 7, 2023

- Dynamic Programming

  - Policy Iteration
  - Value Iteration

  } Vector-Form

  - Policy Iteration
  - Value Iteration

  } Matrix-Form

Project 1 → Due Feb 8

HW2 is posted → Due Feb 17

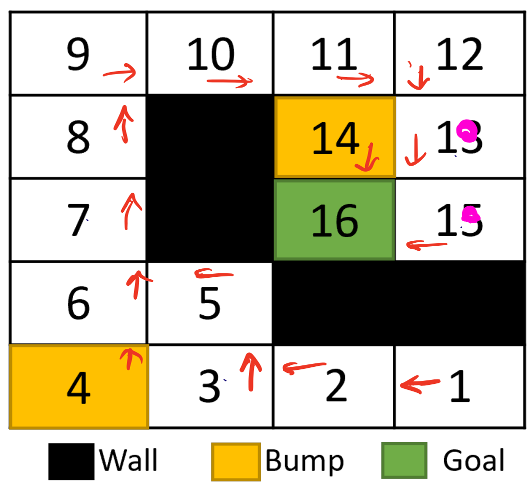
TA's office hour: Wendsdays, 2pm-3pm (in-person)
Fridays, 2pm-3pm (virtual)

Bellman Eq.

$$V_\pi(s) = E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid s_t = s, \pi]$$

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

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

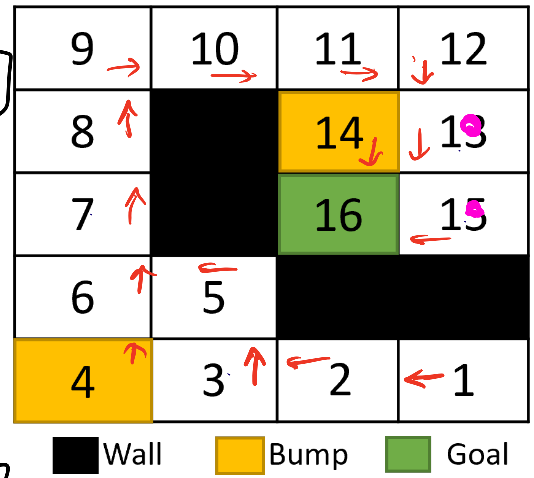


$$Q_\pi(s, a) = E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid s_t = s, a_t = a, \pi]$$

$$= E[R_{t+1} + \gamma V_\pi(s') \mid s_t = s, a_t = a, \pi]$$

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

Deterministic



$$V_\pi(15) = \sum_{s'} P(s'|s,\pi(15))[R+\gamma V_\pi(s')]$$

$$= 1\times[R(15, L, Goal) + \gamma \underbrace{V_\pi(Goal)}_{0}]$$

$$= 99$$

$$V_\pi(13) = \sum_{s'} P(s'|13, D)[R + \gamma V_\pi(s')]$$

$$= 1\times[\underbrace{R(13,D,15)}_{-1} + \gamma \underbrace{V_\pi(15)}_{99}] = 98$$

$$Q_\pi(15,U) = -1 + \gamma \underbrace{V_\pi(13)}_{98} = 97$$

$$Q_\pi(15,D) = -1+\gamma \underbrace{V_\pi(15)}_{99} = 98$$

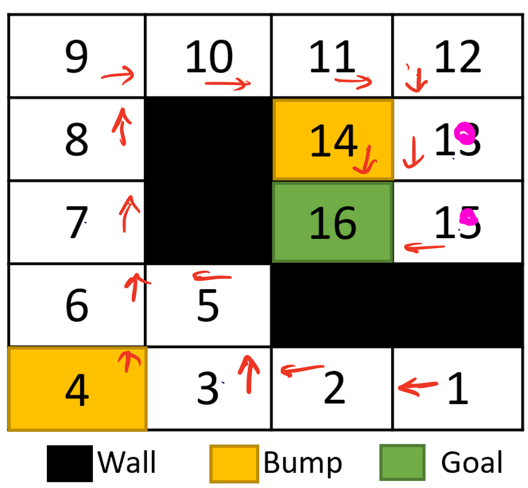
$$Q_\pi(15,R) = -1+\gamma V_\pi(15) = 98$$

$$Q_\pi(15,L) = 99 + \gamma \overbrace{V_\pi(Goal)}^{0} = 99$$

Bellman Eq $\Rightarrow$ allows us to evaluate $V_\pi(s)$

$$\pi^*(s) = \underset{\pi \in \Pi}{\arg\max}\ V_\pi(s) \quad \text{for all } s \in S$$

$$\pi^*(s) = \underset{\pi \in \Pi}{\arg\max}\ Q_\pi(s,a) \quad \text{for all } s, a$$



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

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

Bellman Optimality Eq.

$$\pi^*: \; V_{\pi^*}(s) \geq V_\pi(s), \text{ for all } s, \; \pi \in \Pi$$

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

$$V_{\pi^*}(s) = V^*(s) = \max_{a \in A} E\left[R_{t+1} + \gamma V^*(s') \mid S_t = s, a_t = a\right]$$

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

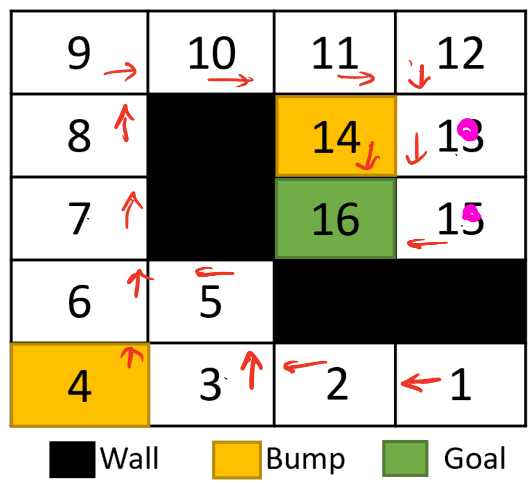
$$V^*(s) = \max_{a \in A} Q^*(s, a)$$

$V_\pi(s) \leq Q_\pi(s,a)$ ✗ Not true

$\underbrace{V_\pi(s)} \leq \max_{a \in A} Q_\pi(s,a)$

$Q_\pi(s, \pi(s))$

$Q_\pi(15, L) \rightarrow V_\pi(15)$
U
D
R



Bellman Opt $V^*(s) = \max_{a \in A} Q^*(s,a)$

$\underbrace{\text{For } \pi^*}$ $V_{\pi^*}(s) < \max_{a \in A} Q_{\pi^*}(s,a) = Q^*(s, a_s)$

$a_s \neq \pi^*(s)$

$\pi'(s) = a_s$

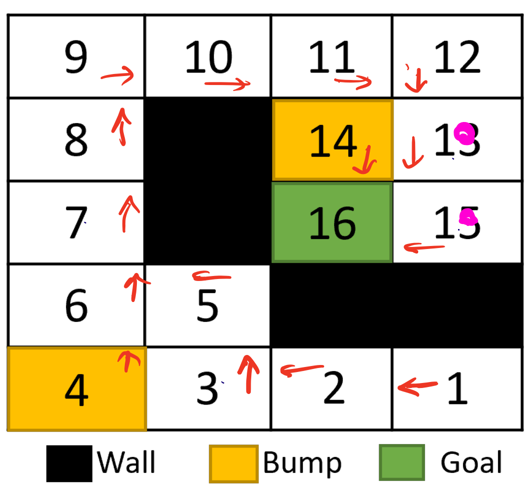
$\pi'(s) = \pi(s)$

$V^{\pi^*}(s) < V^{\pi'}(s) = Q^{\pi^*}(s, a^s)$

$$Q^*(s,a) = E\left[\overset{R(s,a,s')}{R_{t+1}} + \gamma \underbrace{V^*(s')}_{\max_{a' \in A} Q^*(s',a')} \mid S_t = s,\ a_t = a\right]$$

$$= E\left[R_{t+1} + \gamma \max_{a' \in A} Q^*(s',a') \mid S_t = s,\ a_t = a\right]$$

$$Q^*(s,a) = \sum_{s'} P(s' \mid s,a)\left[R(s,a,s') + \gamma \max_{a' \in A} Q^*(s',a')\right]$$



Dynamic Programming (DP)

- Policy Iteration $\rightarrow$ optimal policy π^*
- Value Iteration $\rightarrow$ optimal policy π^*

- Policy Iteration

- Policy Evaluation
- Policy Improvement

$$\pi^0 \xrightarrow{\text{Policy Evaluation}} V^{\pi^0} \xrightarrow{\text{Policy Improvement}} \pi^1 \xrightarrow{\text{PE}} V^{\pi^1} \rightarrow \pi^2 \rightarrow V^{\pi^2} \rightarrow \pi^3 \text{ — } \cdots$$

π^T and π^{T+1} are the same $\Rightarrow \pi^T = \pi^{T+1} = \pi^*$

Policy Evalation Step is policy Iteration Algorithm

① Exact Solution according to Bellman Eq.

A	B

$$M(a^1) = \begin{array}{c|cc} & A & B \\ \hline A & 0.9 & 0.1 \\ B & 0.1 & 0.9 \end{array}$$

$$\text{Reward} \begin{cases} +5 & \text{go } B \\ -1 & \text{if } a^2 \end{cases}$$

$$M(a^2) = \begin{array}{c|cc} & A & B \\ \hline A & 0.1 & 0.9 \\ B & 0.9 & 0.1 \end{array}$$

$$\left\{ \pi^1 = \begin{bmatrix} a^1 \\ a^1 \end{bmatrix}, \pi^2 = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix}, \pi^3 = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix}, \pi^4 = \begin{bmatrix} a^2 \\ a^2 \end{bmatrix} \right\} = \Pi$$

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

Bellman Eq.

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

N states

N Equation

N Variables $V^\pi(s^1) \dots V^\pi(s^N)$

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

$$V_\pi(A) = \underbrace{P(A \mid A, \overset{a^1}{\pi(A)})}_{0.9} \left[\underbrace{R(A, a^1, A)}_{0} + \overset{0.9}{\gamma} V_\pi(A) \right]$$

$$+ \underbrace{P(B \mid A, \overset{a^1}{\pi(A)})}_{0.1} \left[\underbrace{R(A, a^1, B)}_{+5} + \gamma V_\pi(B) \right]$$

$$\boxed{V_\pi(A) = 0.81\, V_\pi(A) + 0.5 + 0.09\, V_\pi(B)}$$

$$V_\pi(B) = \underbrace{P(A \mid B, \overset{a^2}{\pi(B)})}_{0.9} \left[\underbrace{R(B, a^2, A)}_{-1} + \gamma V_\pi(A) \right]$$

$$+ \underbrace{P(B \mid B, \overset{a^2}{\pi(B)})}_{0.1} \left[\underbrace{R(B, a^2, B)}_{4} + \gamma V_\pi(B) \right]$$

$$\boxed{V_\pi(B) = -0.9 + 0.81\, V_\pi(A) + 0.4 + 0.09\, V_\pi(B)}$$

$$0.19\, V_\pi(A) - 0.09\, V_\pi(B) = 0.5$$

$$-0.81\, V_\pi(A) + 0.91\, V_\pi(B) = -0.5$$

$$\underbrace{\begin{bmatrix} 0.19 & -0.09 \\ -0.81 & 0.91 \end{bmatrix}}_{A} \begin{bmatrix} V_\pi(A) \\ V_\pi(B) \end{bmatrix} = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}$$

$$\begin{bmatrix} V_\pi(A) \\ V_\pi(B) \end{bmatrix} = A^{-1} \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}$$

(2) Approximate Policy Evaluation

- Start with arbitrary $V_0(s)$ for all $s \in S$

$$V_{k+1}(s) = \sum_{s'} P(s' \mid s, \pi(s)) \left[R(s, \pi(s), s') + \gamma V_k(s') \right] \text{ for } s \in S \qquad (I)$$

$$\| V_{t+1} - V_t \|_\infty = \max_{s \in S} | V_{t+1}(s) - V_{t+1}(s) | < \theta$$

threshold

$$\hookrightarrow V^\pi = V_{t+1}$$

$$V_0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \xrightarrow{(I)} V_1 = \begin{bmatrix} \; \\ \; \\ \vdots \\ \; \end{bmatrix} \xrightarrow{(I)} V_2 = \begin{bmatrix} \; \\ \; \\ \; \end{bmatrix} \cdots$$

$$\pi^2 = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix} \leftarrow \text{Randomly} \xrightarrow{\text{Policy Evaluation}} V^{\pi^2}$$

$$V_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$V_1(A) = \sum_{s'} P(s' \mid S=A, \pi(A)=a^1)\left[R(S, a^1, s') + \gamma V_0(s')\right]$$

$$= 0.9\left[\underbrace{R(A, a^1, A)}_{0} + \gamma \underbrace{V_0(A)}_{0}\right] + 0.1\left[\underbrace{R(A, a^1, B)}_{5} + \gamma \underbrace{V_0(B)}_{0}\right] = 0.5$$

$$V_1(B) = 0.9\left[\overbrace{R(B, a^2, A)}^{-1} + \gamma \overbrace{V_0(A)}^{0}\right] + 0.1\left[\underbrace{R(B, a^2, B)}_{4} + \gamma \underbrace{V_0(B)}_{0}\right] = -0.5$$

$$V_1 = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}$$

$$\theta = 0.05$$

$$\| V_1 - V_0 \|_\infty = \max_i |V_1(i) - V_0(i)| = 0.5$$

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

$$V_2(A) = 0.9\,[\underbrace{R(A, a^1, A)}_{0} + \gamma \underbrace{V_1(A)}_{0.5}]$$

$$+ 0.1\,[\underbrace{R(A, a^1, B)}_{5} + \gamma \underbrace{V_1(B)}_{-0.5}] = 0.86$$

$$V_2(B) = 0.9\,[\overbrace{R(B, a^2, A)}^{-1} + \gamma \overbrace{V_1(A)}^{0.5}]$$

$$+ 0.1\,[\underbrace{R(B, a^2, B)}_{4} + \gamma \underbrace{V_1(B)}_{-0.5}] = -0.14$$

$$V_2 = \begin{bmatrix} 0.86 \\ -0.14 \end{bmatrix}$$

$$\max_b |V_2 - V_1| = 0.36 < \underline{\theta\ 0.05}\ \times$$

$$\begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} \quad \begin{bmatrix} 0.86 \\ -0.14 \end{bmatrix}$$

V_3

$$|V_{99} - V_{100}| < \theta \quad \checkmark\checkmark \quad V_{100} = V^{\pi} = \begin{bmatrix} 3.73 \\ 2.82 \end{bmatrix}$$

$$\| T(U) - T(V) \|_\infty < \gamma \| U - V \|_\infty$$

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

U

V

Policy Evaluation

Input π, the policy to be evaluated.

Initialization: a small threshold $\theta > 0$, $V(s) = 0$ for $s \in S$.

Loop:

  $\Delta \leftarrow 0$

  Loop for each $s \in \mathcal{S}$:

    $v \leftarrow V(s)$

    $V(s) \leftarrow \sum_{s',r} p(s', r \mid s, \pi(s)) \big[r + \gamma V(s') \big]$

    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$

until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

$V^{\pi}(s) = V(s)$ for $s \in S$

$$\text{Random } \pi^0 \xrightarrow{\text{Policy Eval}} \underline{V^{\pi^0}} \xrightarrow{\text{Policy Impr}} \pi^1$$

Policy Improvement

π' that is better than π

$$\pi'(s) = \underset{a \in A}{\operatorname{argmax}}\ Q^{\pi}(s, a) \quad \text{for all } s$$

$$= \underset{a \in A}{\arg\max} \sum_{s'} P(s'|s,a)\left[R + \gamma V_\pi(s')\right]$$