Lecture 10 - Feb 14, 2023

- Dynamic Programming
  - Policy Iteration
  - Value Iteration

  } Vector-Form

  - Policy Iteration
  - Value Iteration

  } Matrix-Form

- Approximate Dynamic Programming
  - Asynchronous DP
  - Generalized Policy Iteration

Exam 1 → Tuesday, Feb 21

HW2 → Due Feb 17

Project 2 → Due March 3

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

Dynamic Programming

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

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

Approach 1: Policy Iteration

Policy Evaluation (PE)

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

Policy Improvement

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

$$\pi_0 \xrightarrow{PE} V_{\pi_0} \xrightarrow{PI} \pi^1 \xrightarrow{PE} V_{\pi_1} \xrightarrow{PI} \pi_2 \cdots$$

$$\pi_T = \pi_{T-1} = \pi^*$$

Approach 2: Value Iteration (VI)

Value Iteration Backup (VIB)

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

for all $s \in S$

$$V_0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \end{bmatrix} \longrightarrow V_1 = \begin{bmatrix} \\ \\ \end{bmatrix} \longrightarrow V_2 \longrightarrow \cdots V_T$$

$$\max \|V_T - V_{T-1}\| < \theta \longrightarrow V_T = V^*$$

$\boxed{\pi^*(s) = \arg\max Q^*(s,a)}$

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

$$V_0 \rightarrow V_1 \rightarrow V_2 - - - \quad V_T = V^* \longrightarrow \pi^*$$

Proof.

$$V^*(s^1) = \max_{a\in A} \sum_{s'} P(s'|s^1, a)\left[R(s^1, a, s') + \gamma V^*(s')\right]$$

$\vdots$

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

$$V^* = TV^*$$

$$U_0 \xrightarrow{TU_0} U_1 \xrightarrow{TU_1} U_2 - - - - \quad \|U_T - U_{T-1}\|_\infty < \theta$$

V and U are two random vectors

$$|TV(s) - TU(s)| \leq \gamma \|V - U\|_\infty$$

$$|TV(s) - TU(s)| = \Big| \max_{a\in A} \underbrace{\sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V(s')\right]}_{f(a)}$$

$$- \max \underbrace{\sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma U(s')\right]}_{g(a)}$$

Lemma

$$\left|\max_a f(a) - \max_a g(a)\right| \leq \max_a |f(a) - g(a)|$$

$$\leq \max_{a \in A} \left| \gamma \sum_{s'} P(s'|s,a) \left[V(s') - U(s') \right] \right|$$

max happens at a^s

$$= \gamma \left| \sum_{s'} P(s'|s,a^s) \left[V(s') - U(s') \right] \right|$$

$$\leq \gamma \, \| V - U \|_\infty$$

Example:

A | B

$$M(a^1) = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{cc} A & B \\ \left[\begin{matrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{matrix}\right. & \left.\right] \end{array}$$

$$M(a^2) = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{matrix} A & \quad B \end{matrix} \\ \left[\begin{matrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{matrix}\right] \end{array}$$

Reward $\begin{cases} +5 & \text{ending at } B \\ -1 & \text{for } a^2 \end{cases}$

$\gamma = 0.9$

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

$$V_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} V_0(A) \\ V_0(B) \end{bmatrix} \xrightarrow{\text{VIB}} V_1 = \begin{bmatrix} V_1(A) \\ V_1(B) \end{bmatrix}$$

$$V_1(A) = \max_{a \in \{a^1, a^2\}} \sum_{s'} P(s'|A, a)\left[R(A,a,s') + \gamma V_0(s')\right]$$

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

$$\underbrace{\underbrace{P(B|A,a^2)}_{0.9}\left[\underbrace{R(A,a^2,B)}_{4} + \gamma \overset{0}{V_0(B)}\right] + \underbrace{P(A|A,a^2)}_{0.1}\left[\underbrace{R(A,a^2,A)}_{-1} + \gamma \overset{0}{V_0(A)}\right]}_{3.6} \quad a^2$$

$= 3.5$

$$V_1 = \begin{bmatrix} V_1(A) \\ V_1(B) \end{bmatrix} = \begin{bmatrix} 3.5 \\ \quad \end{bmatrix}$$

$$V_1(B) = \max_{a \in A} \sum_{s'} P(s' \mid B, a)\left[R(B, a, s') + \gamma V_0(s')\right]$$

$$= \max \left\{ \overbrace{\underbrace{P(B \mid B, a^1)}_{0.9}\left[\underbrace{R(B, a^1, B)}_{5} + \gamma \overbrace{V_0(B)}^{0}\right] + \underbrace{P(A \mid B, a^1)}_{0.1}\left[\overbrace{R(B, a^1, A)}^{0} + \gamma \underbrace{V_0(A)}_{0}\right]}^{4.5} \right. \quad a^1$$

$$\left. \underbrace{\underbrace{P(A \mid B, a^2)}_{0.9}\left[\underbrace{R(B, a^2, A)}_{-1} + \gamma \overbrace{V_0(A)}^{0}\right] + \underbrace{P(B \mid B, a^2)}_{0.1}\left[\underbrace{R(B, a^2, B)}_{4} + \gamma \overbrace{V_0(B)}^{0}\right]}_{-.5} \right\} \quad a_2$$

$$= 4.5$$

$$V_1 = \begin{bmatrix} 3.5 \\ 4.5 \end{bmatrix} \longrightarrow V_2 \longrightarrow V_3 \longrightarrow V_T$$

$$V^* = V_{100} = \begin{bmatrix} 43.1 \\ 44.1 \end{bmatrix}$$

$$\max |V_{100} - V_{99}| < \theta \quad 0.01 \quad \checkmark \quad V_{100} = V^*$$

$$\pi^* = \begin{bmatrix} \pi^*(A) \\ \pi^*(B) \end{bmatrix} \longleftarrow V^*$$

$$\boxed{\pi^*(s) = \underset{a \in A}{\arg\max} \sum_{s'} P(s' \mid s, a)\left[R(s, a, s') + \gamma V^*(s')\right]}$$

$$V^* = \begin{bmatrix} 43.1 \\ 44.1 \end{bmatrix}$$

$$\pi^*(A) = \underset{a \in \{a^1, a^2\}}{\arg\max} \left\{ \overbrace{\overbrace{P(A \mid A, a^1)}^{0.9}\left[\overbrace{R(A, a^1, A)}^{0} + \gamma \overbrace{V^*(A)}^{43.1}\right] + \overbrace{P(B \mid A, a^1)}^{0.1}\left[\overbrace{R(A, a^1, B)}^{5} + \gamma \overbrace{V^*(B)}^{44.1}\right]}^{39.38} \right.$$

a^1

$$\left. \underbrace{\overbrace{P(B \mid A, a^2)}^{0.9}\left[\overbrace{R(A, a^2, B)}^{4} + \gamma \overbrace{V^*(B)}^{44.1}\right] + \overbrace{P(A \mid A, a^2)}^{0.1}\left[\overbrace{R(A, a^2, A)}^{-1} + \gamma \overbrace{V^*(A)}^{43.1}\right]}_{43.1} \right\}$$

a^2

$= a^2$

$$\pi^*(B) = \underset{a \in \{a^1, a^2\}}{\arg\max} \left\{ \overbrace{\underbrace{P(B \mid B, a^1)}_{0.9}\left[\underbrace{R(B, a^1, B)}_{5} + \gamma \underbrace{V^*(B)}_{44.1}\right] + \underbrace{P(A \mid B, a^1)}_{0.1}\left[\overbrace{R(B, a^1, A)}^{0} + \gamma \overbrace{V^*(A)}^{43.1}\right]}^{44.1} \right.$$

a^1

$$\left. \underbrace{\underbrace{P(A \mid B, a^2)}_{0.9}\left[\underbrace{R(B, a^2, A)}_{-1} + \gamma \underbrace{V^*(A)}_{43.1}\right] + \underbrace{P(B \mid B, a^2)}_{0.1}\left[\underbrace{R(B, a^2, B)}_{4} + \gamma \underbrace{V^*(B)}_{44.1}\right]}_{38.38} \right\}$$

$= a^1$

Value Iteration

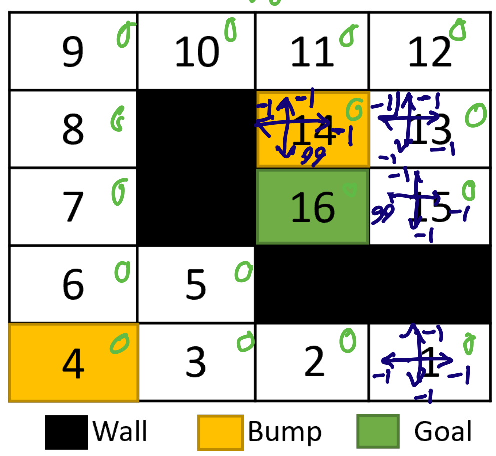
Initialize array V arbitrarily (e.g., $V(s) = 0$ for all $s \in \mathcal{S}^+$)

Repeat
  $\Delta \leftarrow 0$
  For each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \max_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$ (a small positive number)

Output a deterministic policy, π, such that
  $\pi(s) = \arg\max_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]$

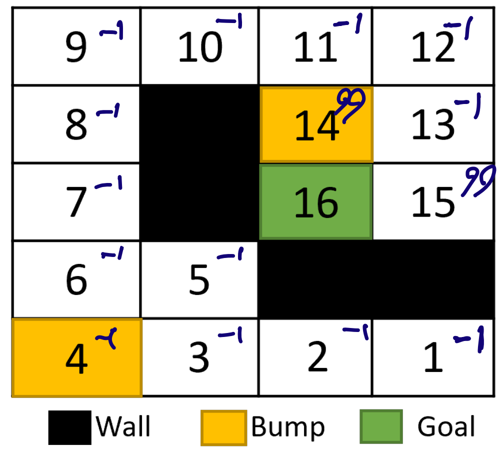
Deterministic $\gamma = 1$

V_0



VIB

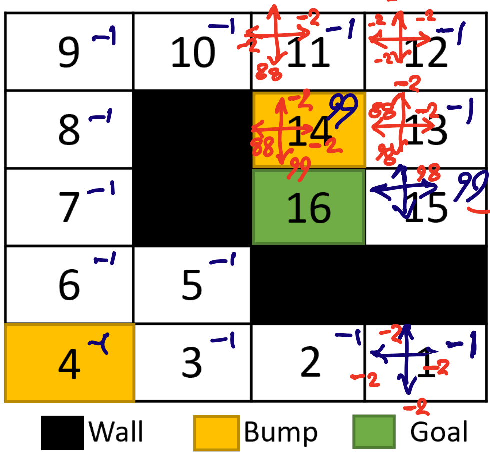
V_1



$$V_1(s) = \max_{a \in A} \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V_0(s')\right]$$

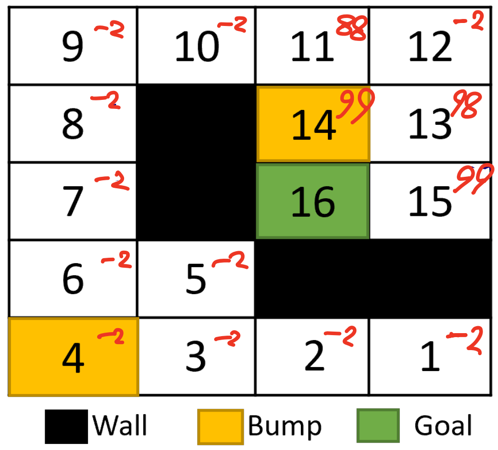
$$s=15 \longrightarrow V_1(15) = \max_{a \in \{U\,L\,D\,R\}} \left\{ \begin{array}{l} \underbrace{\overbrace{P(13|15,U)}^{1}\left[\overbrace{R(15,U,13)}^{-1} + \gamma \overbrace{V_0(13)}^{0}\right] = -1}_{U} \\ \underbrace{\overbrace{P(Goal|15,L)}^{1}\left[\overbrace{R(15,L,Goal)}^{99} + \gamma \overbrace{V_0(Goal)}^{0}\right] = 99}_{L} \\ \underbrace{\overbrace{P(15|15,D)}^{1}\left[\overbrace{R(15,D,15)}^{-1} + \gamma \overbrace{V_0(15)}^{0}\right] = -1}_{D} \\ \underbrace{\qquad\qquad\qquad\qquad = -1}_{R} \end{array} \right\} = 99$$

V_1

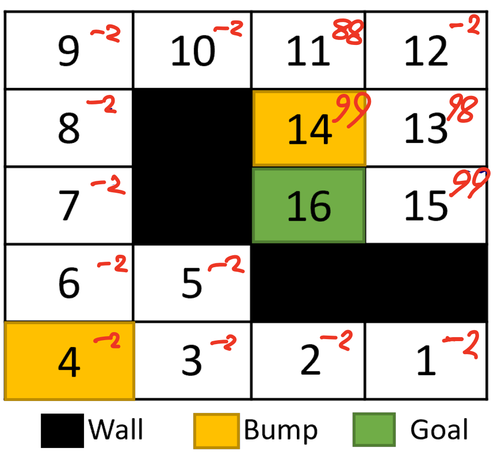


U → -1 + (-1) = -2
D → -1 + 99 = 98
R → -1 + 99 = 98
L → 99

V_2

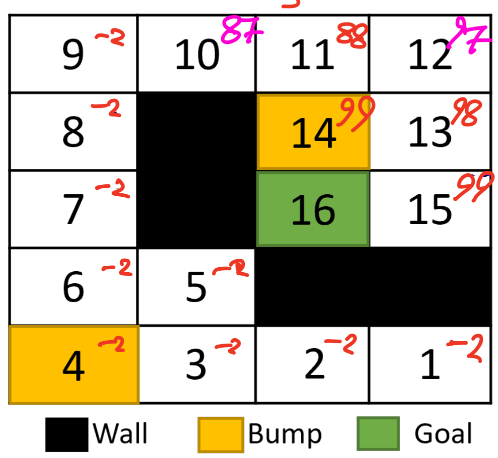


V_2



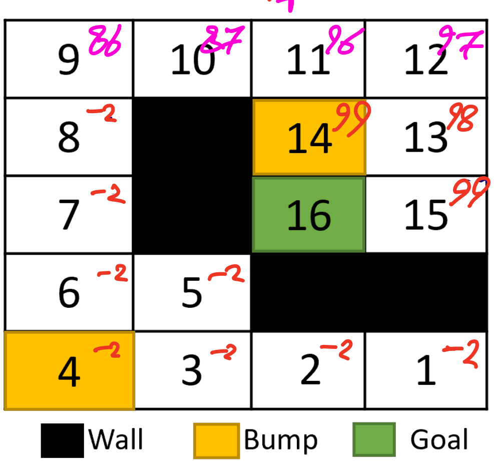
V_3

8



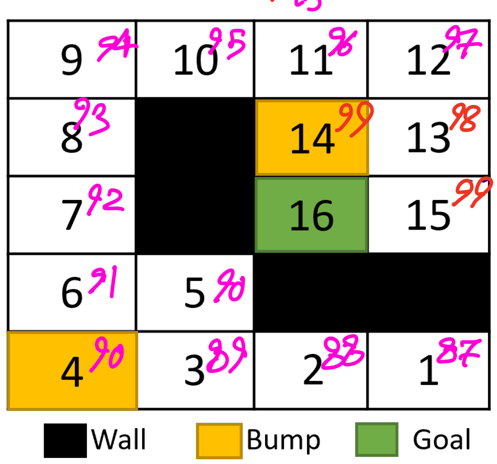
V_4

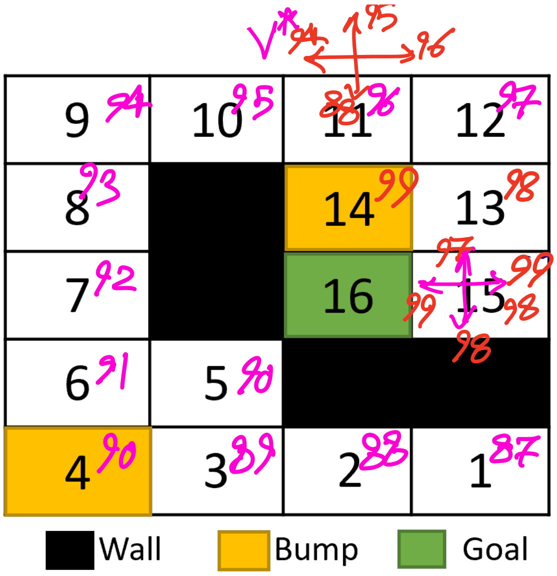
8

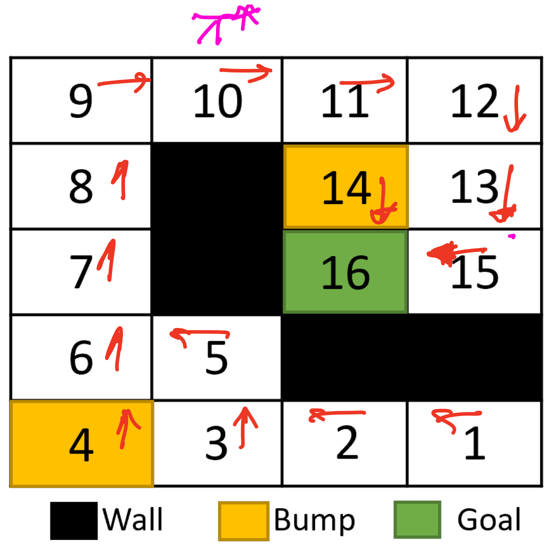


...

$V_{13} = V^*$







$$\pi^*(s) = \arg\max \sum_{s'} P(s'|s,a)\,[R + \gamma V^*(s')]$$