Lecture 9 - Feb 10, 2023

- Dynamic Programming

$\left.\begin{array}{l}\bullet \text{ Policy Iteration} \\ \bullet \text{ Value Iteration}\end{array}\right\}$ Vector-Form

$\left.\begin{array}{l}\bullet \text{ Policy Iteration} \\ \bullet \text{ Value Iteration}\end{array}\right\}$ Matrix-Form

HW2 → Due Feb 17

Project 2 will be assigned → Due March 3

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

Policy Iteration

$$\pi^*_{(s)} = \arg\max_{\pi \in \Pi} V_\pi(s) \quad \text{for all } s$$

① Policy Evaluation (PE)

② Policy Improvement (PI)

Random Policy

$$\pi^0 \xrightarrow{PE} V_{\pi^0} \xrightarrow{PI} \pi^1 \xrightarrow{PE} V_{\pi^1} \xrightarrow{PI} \pi^2 \longrightarrow V_{\pi^2} \ - - - \ \pi_{T-1} \longrightarrow V_{\pi_{T-1}} \longrightarrow \pi_T$$

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

PE: Computes V_π for any given π

Bellman Eq

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

① Solving Bellman Eqs $\rightarrow \left\{ \begin{array}{l} N \text{ Equations} \\ N \text{ Variables} \end{array} \right.$

(Small State Space)

②

$$V_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}_{N \times 1} \leftarrow \text{Random}$$

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

$$\underset{V_0}{\begin{bmatrix} 0 \\ 0 \\ \cdot \\ \vdots \end{bmatrix}} \rightarrow \underset{V_1}{\begin{bmatrix} 0.1 \\ 0.2 \\ 1 \\ 1 \end{bmatrix}} \text{———} \cdots \quad \underset{V_t}{\begin{bmatrix} \\ \\ \\ \end{bmatrix}} \quad \underset{V_{t+1}}{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}$$

$$\max \left| V_{t+1} - V_t \right| = \left\| V_{t+1} - V_t \right\|_\infty < \theta$$

Bellman Eq ← vector form

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

$$\vdots$$

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

$$\hookrightarrow T^{\pi}(V)$$

$$T^{\pi}(V)(s^1) = \sum_{s'} P(s'|s^1, \pi(s^1)) \left[R(s^1, \pi(s^1), s') + \gamma V(s')\right]$$

$$\vdots$$

$$T^{\pi}(V)(s^N) = \sum_{s'} P(s'|s^N, \pi(s^N)) \left[R(s^N, \pi(s^N), s') + \gamma V(s')\right]$$

$$V = T^{\pi} V$$

$\hookrightarrow$ Contraction-mapping Theorem

$$\|T^{\pi}(V) - T^{\pi}(U)\|_{\infty} \leq \gamma \|V - U\|_{\infty}$$

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

$$\vdots$$

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

$$\vdots$$

$$\|T^{\pi}(V) - T^{\pi}(U)\|_{\infty} = \|\gamma \sum_{s'} P(s'|s, \pi(s)) \; V(s') - U(s')\|_{\infty}$$

$$\vdots$$

$$\le \gamma \|V-U\|_\infty$$

PI :

Bellman Optimality Eq.

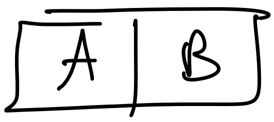
$$\pi_0 \xrightarrow{PE} V_{\pi_0} \xrightarrow{PI} \pi'$$

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

Policy Improvement

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

Example:



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

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

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

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

Policy Iteration:

Random policy $\pi^0 = \begin{pmatrix} a^1 \\ a^2 \end{pmatrix}$, $\theta = 0.1$, $\gamma = 0.9$

$$\pi^0 \xrightarrow{PE} V^{\pi^0} \longrightarrow \pi^1 \longrightarrow V^{\pi^1} - \;\; — \;\; —$$

$$V_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \rightsquigarrow V_{k+1}(s) = \sum_{s'} P(s' \mid s, \pi^0(s)) \left[R + \gamma V_k(s') \right]$$

$$\downarrow$$

$$V_1 = \begin{bmatrix} \; \\ \; \end{bmatrix}$$

$$V_1(A) = \sum_{s'} P(s' \mid A, \pi^0(A) = a^1) \left[R(A, a^1, s') + \gamma V_0(s') \right]$$

$$= P(A \mid A, \pi^0(A) = a^1) \left[R(A, a^1, A) + \gamma V_0(A) \right] = 0$$

$\qquad 1 \qquad\qquad 0 \qquad\qquad 0$

$$V_1(B) = \sum_{s'} P(s' \mid B, \pi^0(B) = a^2)\left[R(B, a^2, s') + \gamma V_0(s')\right]$$

$$= \underbrace{P(A \mid B, \pi^0(B) = a^2)}_{1}\left[\underbrace{R(B, a^2, A)}_{-1} + \gamma \underbrace{V_0(A)}_{0}\right] = -1$$

$$V_1 = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \qquad \max \| \underbrace{V_1}_{\begin{bmatrix} 0 \\ -1 \end{bmatrix}} - \underbrace{V_0}_{\begin{bmatrix} 0 \\ 0 \end{bmatrix}} \| = 1 < \theta \xrightarrow{} 0.1 \quad \times$$

Continue

$$V_2(A) = \sum_{s'} P(s' \mid A, \pi^0(A) = a^1)\left[R(A, a^1, s') + \gamma V_1(s')\right]$$

$$= \underbrace{P(A \mid A, a^1)}_{1}\left[\underbrace{R(A, a^1, A)}_{0} + \gamma \underbrace{V_1(A)}_{0}\right] = 0$$

$$V_2(B) = \sum_{s'} P(s' \mid B, \pi^0(B) = a^2)\left[R(B, a^2, s') + \gamma V_1(s')\right]$$

$$= \underbrace{P(A \mid B, a^2)}_{1}\left[\underbrace{R(B, a^2, A)}_{-1} + \gamma \underbrace{V_1(A)}_{0}\right] = -1$$

$$V_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \qquad \max | \underbrace{V_2}_{\begin{bmatrix} 0 \\ -1 \end{bmatrix}} - \underbrace{V_1}_{\begin{bmatrix} 0 \\ -1 \end{bmatrix}} | = 0 < \underset{\theta}{0.1} \quad \checkmark \text{ stop}$$

$$\underbrace{V_2 = V_{\pi^0}} \qquad\qquad V_{\pi^0} = \begin{bmatrix} V_{\pi^0}(A) \\ V_{\pi^0}(B) \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

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

for all s

$$\pi^0 = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix} \qquad \underbrace{V_{\pi^0} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}}$$

$$\pi'(A) = \underset{a \in \{a^1, a^2\}}{\text{argmax}} \left\{ \sum_{s'} P(s' \mid A, a) \left[R(A, a, s') + \gamma V_{\pi^0}(s') \right] \right\}$$

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

$$\pi^1(A) = \text{argmax} \left\{ \underbrace{\sum_{s'} P(s' \mid A, a^1) \left[R(A, a^1, s') + \gamma V_{\pi^0}(s') \right] = 0}_{a^1} \right.$$

$$\overbrace{P(B \mid A, a^2)}^{1} \left[\overbrace{R(A, a^2, B)}^{4} + \overbrace{\gamma}^{0.9} \overbrace{V_{\pi^0}(B)}^{-1} \right]$$

$$\underbrace{\sum_{s'} P(s' \mid A, a^2) \left[R(A, a^2, s') + \gamma V_{\pi^0}(s') \right] = 3.1}_{a^2} = a^2$$

$$\pi^1(B) = \underset{a \in A}{\arg\max} \left\{ \sum_{s'} P(s' | B, a) \left[R + \gamma V_{\pi^0}(s') \right] \right\}$$

$$\pi^1(B) = \arg\max \left\{ \underbrace{\overbrace{P(B | B, a^1)}^{1} \left[\overbrace{R(B, a^1, B)}^{5} + \gamma \overbrace{V_{\pi^0}(B)}^{-1} \right] = 4.1}_{a^1} \right.$$

$$\left. \underbrace{\overbrace{P(A | B, a^2)}^{1} \left[\overbrace{R(B, a^2, A)}^{-1} + \gamma \overbrace{V_{\pi^0}(A)}^{0} \right] = -1}_{a^2} \right\} = a^1$$

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

PE: V^{π^1}

$$V_0 \to V_1 \to V_2 \text{ ———}$$

Bellman Eqs

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

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

for all s

$\hookrightarrow$ N equation
N variables $\Big\}$ linear

$$V_{\pi'}(A) = \sum_{s'} P(s' \mid A, \pi'(A) = a^2) \left[R(A, a^2, s') + \gamma V_{\pi'}(s') \right]$$

$$= \underbrace{P(B \mid A, a^2)}_{1} \left[\underbrace{R(A, a^2, B)}_{4} + \gamma \underline{V_{\pi'}(B)} \right]$$

$$V_{\pi'}(A) - 0.9\, V_{\pi'}(B) = 4 \qquad (\mathrm{I})$$

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

$$= \underbrace{P(B \mid B, a^1)}_{1} \left[\underbrace{R(B, a^1, B)}_{5} + \underbrace{\gamma}_{0.9} \underline{V_{\pi'}(B)} \right]$$

$$0.1\, V_{\pi'}(B) = 5 \rightarrow V_{\pi'}(B) = 50$$

$$\hookrightarrow (\mathrm{I}) \quad V_{\pi'}(A) = 49$$

$$\pi^0 = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix} \longrightarrow V_{\pi^0} = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \longrightarrow \pi^1 = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix} \to V_{\pi^1} = \begin{bmatrix} 49 \\ 50 \end{bmatrix}$$

$$\pi^2 \quad PI$$

$$\pi^2(A) = \operatorname*{argmax}_{a \in \{a^1, a^2\}} \left\{ \sum_{s'} P(s' \mid A, a) \left[R(A, a, s') + \gamma V_{\pi^1}(s') \right] \right\}$$

$$= \overbrace{P(A \mid A, a^1)}^{1} \left[\overbrace{R(A, a^1, A)}^{0} + \gamma \overbrace{V_{\pi^1}(A)}^{49} \right] = 44$$

$$\pi^2(A) = \operatorname{argmax} \underbrace{\left\{ \sum_{s'} P(s' \mid A, a^1) \left[R(A, a^1, s') + \gamma V_{\pi^1}(s') \right] = 0 \right.}_{a^1}$$

$$\overbrace{P(B \mid A, a^2)}^{1} \left[\overbrace{R(A, a^2, B)}^{4} + \overbrace{\gamma}^{0.9} \overbrace{V_{\pi^1}(B)}^{50} \right]$$

$$\underbrace{\sum_{s'} P(s' \mid A, a^2) \left[R(A, a^2, s') + \gamma V_{\pi^1}(s') \right] = 49}_{a^2} = a^2$$

$$\pi^2(B) = \underset{a \in A}{\operatorname{argmax}} \left\{ \sum_{s'} P(s' \mid B, a) \left[R + \gamma V_{\pi^1}(s') \right] \right\}$$

$$\pi^2(B) = \operatorname{argmax} \left\{ \underbrace{\overbrace{P(B \mid B, a^1)}^{1} \left[\overbrace{R(B, a^1, B)}^{5} + \gamma \overbrace{V_{\pi^1}(B)}^{50} \right] = 50}_{a^1} \right.$$

$$\left. \underbrace{\overbrace{P(A \mid B, a^2)}^{1} \left[\overbrace{R(B, a^2, A)}^{-1} + \gamma \overbrace{V_{\pi^1}(A)}^{49} \right] = 44}_{a^2} \right\} = a^1$$

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

$$\underbrace{\pi^1 = \pi^2 \overset{\checkmark\checkmark\checkmark}{=} \pi^*}$$

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
   $\quad \Delta \leftarrow 0$
   $\quad$ Loop for each $s \in \mathcal{S}$:
   $\quad\quad v \leftarrow V(s)$
   $\quad\quad V(s) \leftarrow \sum_{s',r} p(s', r \mid s, \pi(s))\big[r + \gamma V(s')\big]$
   $\quad\quad \Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   $\quad$ *old-action* $\leftarrow \pi(s)$
   $\quad \pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$
   $\quad$ If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

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

π' is better than π

$$V_{\pi'} \geq V_\pi \Rightarrow V_{\pi'}(s) \geq V_\pi(s) \text{ for all } s$$

$$V_\pi(s) \leq Q_\pi(s, \pi'(s))$$

$$= E[\overline{R_{t+1}}^{\pi'(s)} + \gamma \underbrace{\overline{V_\pi(S_{t+1})}^{\pi}}_{\leq Q_\pi(S_{t+1}, \pi'(S_{t+1}))} \mid S_t = s, a_t = \pi'(s), a_{t+1:\infty} \sim \pi]$$

$$\leq E[R_{t+1} + \gamma Q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s, a_t = \pi'(s)]$$

$$= E[R_{t+1} + \gamma (\overline{R_{t+2}}^{\pi'(S_{t+1})} + \gamma V_\pi(S_{t+2})) \mid S_t = s, a_t = \pi'(s), a_{t+1} = \pi'(S_{t+1})]$$

$$<$$

$$= E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s, a_t = \pi'(s), a_{t+1} = \pi'(S_{t+1}) \dots \pi']$$

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