lec-10

* Policy Improvement theorem:

For all deterministic policies, π & π', if $\forall s \in S$

$$q^{\pi}(s, \pi'(s)) \geq v^{\pi}(s)$$

Then $\pi' \geq \pi$

$$\forall s \in S, \quad v^{\pi'}(s) \geq v^{\pi}(s)$$

Proof:

$$v^{\pi}(s) \leq q^{\pi}(s, \pi'(s))$$

$$= \mathbb{E}\left[R_t + \gamma R_{t+1} \cdots \middle| \begin{array}{l} S_t = s, \\ A_t = \pi'(s), \pi \end{array}\right]$$

$$= \mathbb{E}\left[R_t + \gamma \cdot v^{\pi}(S_{t+1}) \middle| S_t = s, \pi'\right]$$

$$\leq \mathbb{E}\left[R_t + \gamma \cdot q^{\pi}(S_{t+1}, \pi'(S_{t+1})) \middle| \begin{array}{l} S_t = s, \\ \pi' \end{array}\right]$$

$$= \mathbb{E}\left[R_t + \gamma \cdot \mathbb{E}\left[R_{t+1} + \gamma \cdot v^{\pi}(S_{t+2}) \middle| S_t = s, \pi'\right] \middle| \begin{array}{l} S_t = s, \\ \pi' \end{array}\right]$$

$$= \mathbb{E}\left[R_t + \gamma R_{t+1} + \gamma^2 v^{\pi}(S_{t+2}) \middle| S_t = s, \pi'\right]$$

Refer page 95, Sutton Barto

$$\leq \mathbb{E}\left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \gamma^3 v^{\pi}(S_{t+3}) \middle| S_t = s, \pi'\right]$$

$$\leq \mathbb{E}\left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots \middle| S_t = s, \pi'\right]$$

$$\leq v^{\pi'}(s)$$

Policy improvement

- For all states select an action 'a' that maximizes $q^{\pi}(s,a)$. This is 'Greedy' policy, greedy wrt q^{π}.

-

$$\pi'(s) \in \arg\max_{a} \sum_{s'} P(s,a,s')\left(R(s,a,s') + \gamma \cdot v^{\pi}(s')\right)$$

- what if $\pi' = \pi$?

so,

$$V^{\pi'}(s) = \max_{a} \sum_{s'} P(s,a,s')\left(R(s,a,s') + \gamma \cdot V^{\pi'}(s')\right)$$

Policy Iteration $\Rightarrow \pi_0 \xrightarrow{\text{evaluate}} V^{\pi_0} \xrightarrow{\text{improve}} \pi_1 \rightarrow \dots$

- This will terminate after finite iteration
- Exists an optimal policy that is deterministic (there could be multiple optimal solutions)

* Value Iteration

- One full backup during policy evaluation in policy iteration

V_0 : arbitrary , π_0 = arbitrary

$$V_1 : \forall s \quad V_1(s) = \sum_{s'} P(s, \pi_0(s), s') \left[R(s, \pi_0(s), s') + \gamma \cdot V_0(s') \right]$$

$$\pi_1 : \forall s \quad \pi_1(s) \in \arg\max_a \sum_{s'} P(s,a,s') \left[R(s, \pi_0, s') + \gamma \cdot V_1(s') \right]$$

$\vdots$

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

• VI converges ~~to optimal policy~~ $V_k \rightarrow V^*$

Proof:

Bellman operator:

View state value function as vectors in $\mathbb{R}^{|S|}$

$$V_K = \begin{bmatrix} V_K(s_1) \\ V_K(s_2) \\ \vdots \\ V_K(s_{|S|}) \end{bmatrix}$$

Operator:

$$T : \mathbb{R}^{|S|} \Rightarrow \mathbb{R}^{|S|}$$

$$T(V_K) \triangleq V_{K+1},$$

$$T_{V_K}(s) = V_{K+1}(s)$$

$$= \max_a \sum_{s'} P(s,a,s')\left(R(s,a,s') + \gamma \cdot V_K(s')\right)$$

Contraction mapping

$f : X \to X$ is a contraction mapping if there exists a $\gamma \in [0,1]$ such that

$\forall\, x \in X$ & $\forall\, y \in X$

$d(f(x), f(y)) \leq \lambda d(x, y)$ where,

d is a distance function

Banach fixed point theorem

if f is a contraction mapping or a non-empty complete normed vector space, then f has a unique fixed point & the sequence defined by $x_{k+1} = f(x_k)$ with

Bellman operator is a Contraction mapping

$$f \leftarrow T, \quad X \leftarrow R^{|S|},$$

$$d(V, V') \triangleq \max_{s} \left| V(s) + V'(s) \right|$$

$$= \|V - V'\|$$

$$\|T_V - T_{V'}\| \leq \lambda \|V - V'\|$$

$$= \max_{s} \left| TV(s) - TV'(s) \right|$$

$$= \max_s \left| \begin{array}{l} \max_a \sum_{s'} P(s,a,s')\left(R(s,a,s') + \gamma \cdot V(s')\right) \\ - \\ \max_a \sum_{s'} P(s,a,s')\left(R(s,a,s') + \gamma \cdot V'(s')\right) \end{array} \right|$$

$$= \max_s \left| \max_a \sum_{s'} P(s,a,s')\left(R(s,a,s') + \gamma \cdot V(s')\right) - \sum_{s'} \text{———} \text{———} + \gamma \cdot V'(s') \right|$$

claim $\Rightarrow \left| \max_x f(x) - \max_x g(x) \right| \leq \max_x \left| f(x) - g(x) \right|$

$\forall x \quad f(x) - g(x) \leq \left| f(x) - g(x) \right|$

$$f(x) \leq \left| f(x) - g(x) \right| + g(x)$$

$$\max_x f(x) \leq \max_x \left| f(x) - g(x) \right| + g(x)$$

$$\leq \max_x \left| f(x) - g(x) \right| + \max_x g(x)$$

$\therefore \max_x f(x) - \max_x g(x) \leq \max_x \left| f(x) - g(x) \right|$