Assn 5 ML 20160463 성해빈

1. (a) $V_\pi(0) = \underbrace{\pi(try|0)}_{''\,1} \sum_{r,s'} \underbrace{p(r,s'|0,try)}_{\text{transition deterministic}} (r+\gamma V_\pi(s')) + \underbrace{\pi(give\text{-}up|0)}_{''\,0} \cancel{\sum_{r,s'} p(r,s'|0,giveup)(\cdots)}$

$= r+\gamma V_\pi(s') = 0+\gamma V_\pi(1) = \dfrac{3\gamma^2}{1-\gamma^3}$

$V_\pi(-1) = 0+\gamma V_\pi(0) = \dfrac{3\gamma^3}{1-\gamma^3}$

(b) No. If discount factor is big enough, reward for trying (+3) is bigger than giving up.

(c) reward for try: $0+0+3\cdot\gamma^2$

reward for give up: $0+\gamma$ if $\gamma = 0.0001$, $3\gamma^2 < \gamma$

give up is the optimal action

(d) if $\gamma = 0.9999$, $3\gamma^2 > \gamma$ $\therefore$ try is the optimal action

(e) No. we could see it by (c) and (d).

2. (a) for $V \in R^{|S|\times 1}$ which is a vector representing every state, $B^\pi V$ can be expressed as

$B^\pi V = R^\pi + \gamma P^\pi V$ where $R^\pi \in R^{|S|\times 1}$, $R^\pi(s) = \sum_a \pi(a|s) r(s,a)$ and

$P^\pi \in R^{|S|\times|S|}$, $P^\pi(s,s') = \sum_a \pi(a|s) p(s'|s,a)$

$\|B^\pi V' - B^\pi V\|_\infty = \|\cancel{R^\pi} + \gamma P^\pi V' - \cancel{R^\pi} - \gamma P^\pi V\|_\infty = \|\gamma P^\pi(V'-V)\|_\infty$

$\leq \gamma \|P^\pi\|_\infty \|(V'-V)\|_\infty$ (norm inequality)

$\leq \gamma \|V'-V\|_\infty$ (max of probability is 1)

(b) For any k, in policy evaluating $V_{k+1} \leftarrow B^\pi V_k$, and true value V_π,

$\|V_k - V_\pi\|_\infty = \|B^\pi V_{k-1} - B^\pi V_\pi\|_\infty \leq \gamma \|V_{k-1} - V_\pi\|_\infty = \|B^\pi V_{k-2} - B^\pi V_\pi\|_\infty \cdots \leq \gamma^k \|V_0 - V_\pi\|_\infty$

So $\|V_k - V_\pi\|_\infty \to 0$ as $k \to \infty$, i.e. $\lim_{k\to\infty} V_k = V_\pi$ $\therefore$ convergence (existence)

if there are two values V, V' verifying Bellman equation for π, $\|V-V'\|_\infty = \|B^\pi V - B^\pi V'\|_\infty \leq \gamma \|V-V'\|_\infty$

this means $1 \leq \gamma$, which is a contradiction since $\gamma < 1$. $\therefore$ unique

(c) $\forall s \in S$, $V_n(s) = \sum_a \pi_n(a|s)\, Q_n(s,a)$

$$\le \max_a Q_n(s,a) = \sum_a \pi_{n+1}(a|s) \sum_{s',r} p(s',r|s,a)(r + \gamma V_n(s'))$$

$$\le B^{\pi_{n+1}} V_n(s) \le (B^{\pi_{n+1}} B^{\pi_{n+1}}) V_n(s) \le \cdots$$

$$\le \lim_{a\to\infty} (B^{\pi_{n+1}})^a V_n(s) = V_{n+1}(s)$$

$\therefore V_n \le V_{n+1}$

(d)

$$\| B^* V'(s) - B^* V(s) \|_\infty \le \gamma \| \max_a \sum_{s'} p(s'|s,a) V'(s') - \max_a \sum_{s'} p(s'|s,a) V(s') \|_\infty$$

$$\le \gamma \max_a \| \sum_{s'} p(s'|s,a) V'(s') - \sum_{s'} p(s'|s,a) V(s') \|_\infty$$

$$= \gamma \max_a \sum_{s'} p(s'|s,a) \| V'(s') - V(s') \|_\infty$$

$$\le \gamma \| V'(s) - V(s) \|_\infty$$

$\therefore \| B^* V' - B^* V \|_\infty \le \gamma \| V' - V \|_\infty$

(e) For any k, in value iteration $V_{k+1} \leftarrow B^* V_k$

$\| V_k - V^* \|_\infty = \| B^* V_{k-1} - B^* V^* \|_\infty \le \gamma \| V_{k-1} - V^* \|_\infty \le \cdots \le \gamma^k \| V_0 - V^* \|_\infty$

So $\| V_k - V^* \|_\infty \to 0$ as $k \to \infty$, i.e. $\lim_{k\to\infty} V_k = V^*$ $\therefore$ convergence (existence)

if there are 2 values V, V' verifying optimal Bellman equation,

$\| V - V' \|_\infty = \| B^* V - B^* V' \|_\infty \le \gamma \| V - V' \|_\infty$, which implies $1 \le \gamma$, but is a contradiction since $\gamma < 1$ $\therefore$ unique

(f) in (e) we showed $\lim_{k\to\infty} V_k = V^*$, thus value iteration converges to optimal value.

V^* is a value that satisfies optimal Bellman equation $V^* = B^* V^* = \boxed{\max_a} \sum_{s'} \sum_r p(s',r|s,a)(r + \gamma V(s'))$

You can see the optimal Bellman equation chose the action that maximizes the value, which means it is following optimal policy: $V^*(s) = \max_\pi V^\pi(s)$ for all $s \in S$.

Thus π_n, the policy extracted from V_n, converges to optimal policy π^* as $V_n \to V^*$.

3. (a) for action + in state 1, reward −1 and next state 2

$$Q(1,+) \leftarrow \underbrace{Q(1,+)}_{0} + 0.5\left[-1 + \underbrace{\max_a Q(2,a)}_{0} - \underbrace{Q(1,+)}_{0}\right] = -0.5$$

	−	+
1	0	−0.5
2	0	0
3	0	0

for action + in state 2, reward 1 and next state 3

$$Q(2,+) \leftarrow \underbrace{Q(2,+)}_{0} + 0.5\left[1 + \underbrace{\max_a Q(3,a)}_{0} - \underbrace{Q(2,+)}_{0}\right] = 0.5$$

	−	+
1	0	−0.5
2	0	0.5
3	0	0

(b)

$$Q(1,-) \leftarrow \underbrace{Q(1,-)}_{0} + 0.5\left[-1 + \underbrace{\max_a Q(1,a)}_{0} - \underbrace{Q(1,-)}_{0}\right] = -0.5$$

	−	+
1	−0.5	−0.5
2	0	0.5
3	0	0

$$Q(1,+) \leftarrow \underbrace{Q(1,+)}_{-0.5} + 0.5\left[-1 + \underbrace{\max_a Q(2,a)}_{0.5} - \underbrace{Q(1,+)}_{-0.5}\right] = -0.5$$

	−	+
1	−0.5	−0.5
2	0	0.5
3	0	0

$$Q(2,+) \leftarrow \underbrace{Q(2,+)}_{0.5} + 0.5\left[1 + \underbrace{\max_a Q(3,a)}_{0} - \underbrace{Q(2,+)}_{0.5}\right] = 0.75$$

	−	+
1	−0.5	−0.5
2	0	0.75
3	0	0

(c) $\pi(s) \leftarrow \arg\max_a Q(s,a)$. $\pi(1) = +$, $\pi(2) = +$

(d) It does not converge will full of transition $(1, -, -1, 1)$ is sampled, because it does not visit other transitions. Q-learning converges to optimal when if we visit various transitions.