CS234 Assignment1 2019

pcwu1023

January 2020

1 Gridworld

(a)
$r_s = -1$

0	1	2	3
-5	2	3	4
2	3	4	5
1	0	-1	-2

(b)
all the rewards have +2 added to them

12	11	10	9
-3	10	9	8
10	9	8	7
11	12	13	14

(c)

$$V^{\pi}_{new}(s_i) = \sum_{t=0}^{\infty} \gamma^t (r_t + c) \qquad \text{(deterministic)}$$

$$= \sum_{t=0}^{\infty} \gamma^t r_t + \sum_{t=0}^{\infty} \gamma^t c$$

$$= V^{\pi}_{old}(s_i) + \frac{c}{1-\gamma}$$

(d)
The optimal policy will become that the agent keeps taking actions or moving around in order to maximize the reward.
Values of unshaded squares will become infinite.

(e)
No. The optimal policy would not change (same as (d)) after γ changes to $0 < \gamma < 1$ and it will not depend on the choice of γ. However, the value function would change, it will not be infinite, and will converge in the end.

(f)
We can choose very small and less than zero r_s, such as $r_s = -10$, and start from number 16, 15, 14, 13, 9, etc. square, in order to make a shorter path to the red square than the path to the green square. Therefore, the optimal policy can result in termination in the red square.

2 Value of Different Policies

The following equations are used for the proof:

$$V(s) = E[G_t | S_t = s] = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots \mid S_t = s]$$

$$Q^\pi(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^\pi(s')$$

$$Q_t^{\pi_1}(x_t, \pi_1(x_t,t)) - Q_t^{\pi_1}(x_t, \pi_2(x_t,t)) = R(x_t, \pi_1(x_t,t)) + \gamma \sum_{x_t' \in S} P(x_t' | x_t, \pi_1(x_t,t)) V^{\pi_1}(x_t')$$
$$- R(x_t, \pi_2(x_t,t)) + \gamma \sum_{x_t' \in S} P(x_t' | x_t, \pi_1(x_t,t)) V^{\pi_1}(x_t')$$

Follow π_1 in the future

$$\sum_{t=1}^{H} E_{x_t \sim \pi_2} \left[\left(Q_t^{\pi_1}(x_t, \pi_1(x_t,t)) - Q_t^{\pi_1}(x_t, \pi_2(x_t,t)) \right) \right]$$

$$= \sum_{t=1}^{H} E_{x_t \sim \pi_2} \left[R(x_t, \pi_1(x_t,t)) - R(x_t, \pi_2(x_t,t)) \right]$$

↳ linear

$$= E_{x_t \sim \pi_2} \sum_{t=1}^{H} \left[R(x_t, \pi_1(x_t,t)) - R(x_t, \pi_2(x_t,t)) \right]$$

$$= E_{x_t \sim \pi_2} \left[G_1 | S_1 = s \right]^{\pi_1} - E_{x_t \sim \pi_2} \left[G_1 | S_1 = s \right]^{\pi_2}$$

$$= V_1^{\pi_1}(x_1) - V_1^{\pi_2}(x_1)$$

3 Fixed Point

(a) $n=1$, $\|V_2 - V_1\|_\infty = \|BV_1 - BV_0\|_\infty \le \gamma \|V_1 - V_0\|_\infty$

assume

$n=k$, $\|V_{k+1} - V_k\|_\infty \le \gamma^k \|V_1 - V_0\|_\infty$

prove

$n=k+1$

$\|V_{k+2} - V_{k+1}\|_\infty \le \gamma^{k+1} \|V_1 - V_0\|_\infty$

$\gamma \|V_{k+1} - V_k\|_\infty \le \gamma^{k+1} \|V_1 - V_0\|_\infty$

$\|V_{k+2} - V_{k+1}\|_\infty = \|BV_{k+1} - BV_k\|_\infty \le \gamma \|V_{k+1} - V_k\|_\infty$

$(V_{k+1} = BV_k)$

$\therefore \|V_{k+2} - V_{k+1}\|_\infty \le \gamma^{k+1} \|V_1 - V_0\|_\infty$

$\|V_{n+1} - V_n\|_\infty \le \gamma^n \|V_1 - V_0\|_\infty$ for $k=n$

(b)

$$\begin{aligned}
\|V_{n+c} - V_n\|_\infty &\le \|V_{n+c} - V_{n+c-1}\|_\infty + \cdots + \|V_{n+1} - V_n\|_\infty \\
&\le \gamma^{n+c-1} \|V_1 - V_0\|_\infty + \cdots + \gamma^n \|V_1 - V_0\|_\infty \\
&\le \gamma^n \left(\gamma^{c-1} \|V_1 - V_0\|_\infty + \gamma^{c-2} \|V_1 - V_0\|_\infty + \cdots + \|V_1 - V_0\|_\infty \right) \\
&\le \gamma^n \left(\sum_{\infty} \gamma^t \|V_1 - V_0\|_\infty \right) = \gamma^n \left(\frac{1}{1-\gamma} \|V_1 - V_0\|_\infty \right)
\end{aligned}$$

$\therefore \|V_{n+c} - V_n\|_\infty \le \frac{\gamma^n}{1-\gamma} \|V_1 - V_0\|_\infty$

(c)

From (b)

$$\|V_{n+c} - V_n\|_\infty \le \frac{\gamma^n}{1-\gamma}\|V_1 - V_0\|_\infty$$

$\because$ $m, n > K$, Let $n+c = m$

$$\therefore \|V_m - V_n\|_\infty \le \frac{\gamma^n}{1-\gamma}\|V_1 - V_0\|_\infty \le \frac{\gamma^K}{1-\gamma}\|V_1 - V_0\|_\infty < \varepsilon$$

(d)

Since V is a fixed point, $BV = V$

$$\|BV' - BV''\|_\infty = \|V' - V''\|_\infty \le \gamma\|V' - V''\|_\infty$$

$\because \gamma < 1$

Only $\|V' - V''\|_\infty = 0$, can make $\|V' - V''\|_\infty \le \gamma\|V' - V''\|_\infty$ be true.

4 Frozen Lake MDP

(c)
Stochasticity increases the number of iterations for value iteration required to converge. For policy iteration, the number of iterations might remain the same or policy iteration might diverge. Stochasticity would change the optimal policy as well. The optimal policy for stochastic frozen lake is different from the optimal policy for deterministic frozen lake.