Reinforcement Learning Homework 1

tw2906

Problem 1

1.1

For Arm 1, the average reward is:

$\frac{0.3}{1} = 0.3$

For Arm 2, the average reward:

$\frac{0+1+0+0}{4} = 0.25$

The average reward for arm 1 is greater than arm 2's.

According to the greedy policy, only the arm with best current average reward will be chosen.

So the arm 1 will be played at $t = 6, 7$.

1.2

As calculated before, Q for ARM 1 is 0.3 and Q for ARM 2 is 0.25.

At $t=6$, the probability to play arm 2 is $\frac{\varepsilon}{N} = \frac{0.1}{2} = 0.05 = 5\%$

Three conditions may happen when $t=7$:

①. when $t=6$, arm 1 selected (95%), then $Q(1) > Q(2)$, p for arm 2 at $t=7$ still 0.05.

②. when $t=6$, arm 2 selected (5%), and rewards 1 (60%), New Q for arm 2 should be:

$\frac{0+1+0+0+1}{5} = 0.4$

Now $Q_7(2) > Q_7(1)$, probability to select arm 2 when $t=7$ will be $1-\varepsilon+\frac{\varepsilon}{N} = 0.95 = 95\%$.

③. when $t=6$, arm 2 selected (5%), and rewards 0 (40%), New Q for arm 2 should be:

$\frac{0+1+0+0+0}{5} = 0.2$

$Q_7(1) > Q_7(2)$, p to select arm 2 when $t=7$ will be $\frac{\varepsilon}{N} = 0.05 = 5\%$

In general, the probability to play arm 2 when $t=7$:

$0.95 \times 0.05 + 0.05 \times 0.6 \times 0.95 + 0.05 \times 0.4 \times 0.05$

$= 0.0475 + 0.0285 + 0.001$

$= 0.077$

$= 7.7\%$

1.3

Because, the greedy Method will may stuck at the suboptimal option as it is lacking in acquiring the subsequent data from exploration. Exploration may not choose the best option for current, but the data it earned will be beneficial to the algorithm and will finally shows at long-term.

Problem 2

2.1

$$P(a) = \frac{e^{Q_t(a)/\tau}}{\sum_{i=1}^{n} e^{Q_t(i)/\tau}}$$

when $\tau \to 0$, $e^{Q_t(a)/\tau}$, $e^{Q_t(i)/\tau} \to \infty$

$$P(a) = \frac{\infty}{n\infty} \to 1$$

So the softmax will only choose the best current rewards. as greedy action

2.2

when $\tau \to \infty$, $Q_t(a)/\tau \to 0$, $e^{Q_t(a)/\tau} \to 1$

$$P(a) \to \frac{1}{N}$$

So the softmax will give all selection the same probability $\frac{1}{N}$.

2.3

two actions $\to$ $n=2$

$$P(a) = \frac{e^{Q_t(a)/\tau}}{e^{Q_t(a)/\tau} + e^{Q_t(b)/\tau}}$$

$$= \frac{1}{1 + \frac{e^{Q_t(b)/\tau}}{e^{Q_t(a)/\tau}}}$$

$$= \frac{1}{1 + e^{-(Q_t(a) - Q_t(b))/\tau}}$$

when $n=2$, the softmax has same form as sigmoid function,

as $t = (Q_t(a) - Q_t(b))/\tau$

Problem 3.

$$V_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}$$

$$V_{n+1} = \frac{\sum_{k=1}^{N} W_k G_k}{\sum_{k=1}^{n} W_k}$$

$$= \frac{W_n G_n + \sum_{k=1}^{n-1} W_k G_k}{C_n}$$

$$= \frac{W_n G_n + V_n \cdot \sum_{k=1}^{n-1} W_k}{C_n} = \frac{W_n G_n + V_n C_{n-1}}{C_n}$$

$$= \frac{W_n G_n + V_n C_{n-1} + (V_n W_n - V_n W_n)}{C_n}$$

$$= \frac{W_n G_n + V_n C_n - V_n W_n}{C_n}$$

$$V_{n+1} = V_n + \frac{W_n}{C_n}(G_n - V_n)$$

$\because C_n$ sums the weight given to the first n returns.

$$\therefore C_n = \sum_{k=1}^{n} W_k$$

$$C_{n+1} = \sum_{k=1}^{n+1} W_k$$

$$C_{n+1} = \sum_{k=1}^{n} W_k + W_{n+1}$$

$$C_{n+1} = C_n + W_{n+1}$$

Problem 4.

4.1.

$$V = 4 + 0.2 \times 6 + 0.3 \times 8 + 0.5 \times 10$$
$$= 12.6$$

4.2

$$V = 0.5 \times 4 + 0.5 \times (0.6 \times 4 + 0.4 \times (4 + \overset{1.2}{0.2 \times 6} + \overset{2.4}{0.3 \times 8} + \overset{5}{0.5 \times 10}))$$
$$= 2 + \frac{1}{2} \times (2.4 + 0.4 \times 12.6)$$
$$= 5.72$$

4.3

$$V = 0.5 \times 4 + 0.5 \times (0.4 \times 0.5 + 0.6 \times 5)$$
$$= 3.6$$

Problem 5.

5.1.

- For the original MDP:

$$Q_\pi(s,a) = E_\pi[R_{t+1} + \gamma R_{t+2} \mid s, a]$$

- For the modified MDP with new reward function $\alpha + R_s^a$:

$$Q'_\pi(s,a) = E_\pi[\alpha + R_{t+1} + \gamma(\alpha + R_{t+2}) \mid s, a]$$
$$Q'_\pi(s,a) = E_\pi[R_{t+1} + \gamma R_{t+2} \mid s, a] + \alpha + \gamma\alpha$$
$$\underline{Q'_\pi(s,a) = Q_\pi(s,a) + \alpha + \gamma\alpha}$$

Since that, we can say that change that reward function will not take any affect about the original policy. given the optimal policy for the modified MDP:

$$\pi_M^*(s) = \arg\max Q'_\pi(s,a)$$
$$\pi_M^*(s) = \arg\max Q_\pi(s,a) + \alpha + \gamma\alpha$$
$$\pi_M^*(s) = \pi_O^*(s) + \alpha + \gamma\alpha$$
$$\underline{\pi_M^*(s) = \pi_O^*(s)}$$

Problem 5

6.2

- For the original MDP:

$$Q_\pi(S,a) = E_\pi[R_{t+1} + \gamma R_{t+2} \mid S, a]$$

- For the modified MDP:

$$Q'_\pi(S,a) = E_\pi[\beta R_{t+1} + \beta\gamma R_{t+2} \mid S, a]$$

$$Q'_\pi(S,a) = \beta E_\pi[R_{t+1} + \gamma R_{t+2} \mid S, a]$$

$$\underline{Q'_\pi(S,a) = \beta Q_\pi(S,a)}$$

Since that, we can say that the change of the reward function will not take affect about the optimal policy, the optimal policy for the modified MDP will be:

$$\pi^*_M(S,a) = \operatorname{argmax} Q'_\pi(S,a)$$

$$\pi^*_M(S,a) = \beta \operatorname{argmax} Q_\pi(S,a)$$

$$\pi^*_M(S,a) = \beta \pi^*_O(S,a)$$

$$\underline{\pi^*_M(S,a) = \pi^*_O(S,a)}$$

Problem 6.

6.1

State function:

$$T(S, stop, done) = 1 \quad , \quad S \in \{0, 2, 3, 4, 5\}$$

$$T(S, draw, S') = \frac{1}{2}, \quad S = 0, \quad S' \in \{2, 3\}$$

$$T(S, draw, S') = \frac{1}{2}, \quad S = 2, \quad S' \in \{4, 5\}$$

$$T(S, draw, S') = \frac{1}{2}, \quad S = 3, \quad S' \in \{5, done\}$$

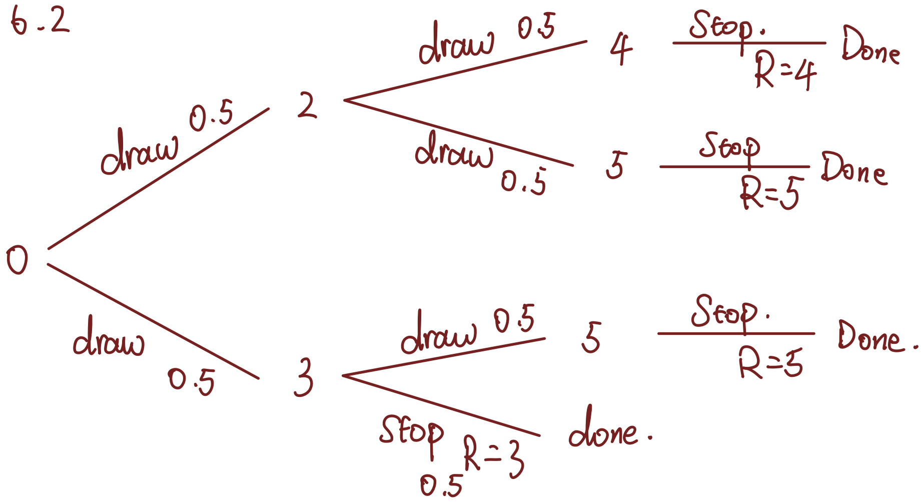
$$T(S, draw, S') = 1, \quad S \in \{4, 5\}, \ S' = done$$

$$T(S, draw, S') = 0 \quad , \text{otherwise.}$$

Reward function:

$$R(s, stop, done) = s, \quad s \leq 5$$
$$R(s, draw, s') = 0 \quad \text{otherwise.}$$

6.2



For the opsimal state-value functions:

$$V_*(0) = 0.5\times(0.5\times4+0.5\times5) + 0.5\times3 = 3.75$$

$$V_*(2) = 0.5\times4 + 0.5\times5 = 4.5$$

$$V_*(3) = 3$$

$$V_*(4) = 4$$

$$V_*(5) = 5$$

For the optimal action-value functions:

$Q_*(0, draw) = 3.75 \qquad Q_*(0, stop) = 0$

$Q_*(2, draw) = 4.5 \qquad Q_*(2, stop) = 2$

$Q_*(3, draw) = 2.5 \qquad Q_*(3, stop) = 3$

$Q_*(4, draw) = 0 \qquad Q_*(4, stop) = 4.$

$Q_*(5, draw) = 0 \qquad Q_*(5, stop) = 5$

6.3 According to the action-value function, when the state is in "0" and "2", action "draw" will result more rewards rather than "stop". And for states "3", "4" and "5", action "stop" can results more rewards.

So, the optimal policies should be:

$$\pi_*(a|s) = \begin{cases} 1, & \begin{cases} a = \text{draw} & \text{if } s \in \{0, 2\} \\ a = \text{stop} & \text{if } s \in \{3, 4, 5\} \end{cases} \\ 0, & \text{otherwise} \end{cases}$$