Homework 1

ELEN 6885 Reinforcement Learning

Problem 1

1. $t=1$, ①, 0.3

   $t=2$, ②, 0

   $t=3$, ②, 1

   $t=4$, ②, 0

   $t=5$, ②, 0

   $P_1 = \frac{0.3}{1} = 0.3$, $P_2 = \frac{0+1+0+0}{4} = \frac{1}{4} = 0.25$.

   $P_1 > P_2$, so arm 1 will be played if greedy method is used.

2. $t=6$, $P_6(arm2) = \frac{\varepsilon}{2} = 0.05$. $P_6(arm1) = 1 - P(arm2) = 0.95$.

   $t=7$, ~~$P_7(arm2) = P_6(arm1) + P_6(arm2)$~~

   if ① arm=1 is played when $t=6$, $P_1 > P_2$, $P_7'(arm2) = 0.05$.
   ② arm 2 is played when $t=6$ and reward $=0$, $P_1 > P_2$, $P_7''(arm2) = 0.05$.
   ③ arm 2 is played when $t=6$ and reward $=1$, $P_1 < P_2$, $P_7'''(arm2) = 0.95$.

   $$\Rightarrow P_7(arm2) = P① \cdot P_7'(arm2) + P② \cdot P_7''(arm2) + P③ \cdot P_7'''(arm2)$$
   $$= 0.95 \times 0.05 + 0.05 \times 0.4 \times 0.05 + 0.05 \times 0.6 \times 0.95$$
   $$= 0.077$$

   $\therefore$ $t=6$, the probability to play arm 2 is 0.05,
   $t=7$, the probability to play arm2 is 0.077.

3. Because the greedy method gets stuck performing suboptimal actions.

Problem 2

1. $$P_a(\tau \to 0) = \frac{e^{\frac{Q_t(a)}{\tau}}}{\sum_{i=1}^{n} e^{\frac{Q_t(i)}{\tau}}} = \frac{1}{\sum_{i=1}^{n} e^{\frac{Q_t(i) - Q_t(a)}{\tau}}}.$$

   if a is the action with max $Q_t(i)$, $P_a(\tau \to 0) = 1$.

if not, $P_a(\tau \to 0) = 0$.

$\Rightarrow$ Softmax action selection becomes the same as greedy method.

2. if $\tau \to \infty$

$$P_a(\tau \to \infty) = \frac{e^{\frac{Q_t(a)}{\tau}}}{\sum_{i=1}^{n} e^{\frac{Q_t(i)}{\tau}}} = \frac{1}{\sum_{i=1}^{n} e^{\frac{Q_t(i)-Q_t(a)}{\tau}}} = \frac{1}{\sum_{i=1}^{n} e^{0}} = \frac{1}{n}.$$

$\Rightarrow$ Softmax method yields equiprobable selection among all actions.

3. if $n=2$.

$$P_a(n=2) = \frac{e^{\frac{Q_t(a)}{\tau}}}{e^{\frac{Q_t(a)}{\tau}} + e^{\frac{Q_t(2)}{\tau}}} = \frac{1}{1 + \frac{e^{\frac{Q_t(2)}{\tau}}}{e^{\frac{Q_t(a)}{\tau}}}} = \frac{1}{1 + e^{-\left(\frac{Q_t(a)-Q_t(2)}{\tau}\right)}}$$

which has the same form as sigmoid function.

Problem 3

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

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

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

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

Problem 4

1. $v(S_t) = r(a_1) + \gamma \left(\sum_a P_a \cdot r \right)$

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

2. $V(S_t) = \cancel{P(a_1)} \, P(\text{right}) \cdot \left(P(\text{right-down}) \cdot r_1 + P(\text{right-up}) \cdot \left(r_2 + \sum_a P_a r_a \right) \right)$
$+ P(\text{left}) \cdot r_3$
$= 0.5 \times (0.6 \times 4 + 0.4 \times (4 + (0.2 \times 6 + 0.3 \times 8 + 0.5 \times 10))) + 0.5 \times 4$
$= 5.72$

3. $V(S_t) = P(a_2) \cdot r_{21} + P(a_1) \cdot (P(r_{11}) \cdot r_{11} + P(r_{12}) \cdot r_{12})$
$= 0.5 \times 4 + 0.5 \times [0.4 \times 0.5 + 0.6 \times 5)$
$= 3.6$

Problem 5

1. $Q_\pi(\text{original}) = E_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \mid s, a]$.
$Q_\pi(\text{modified}) = E_\pi[R_{t+1} + \alpha + \gamma(R_{t+2} + \alpha) + \gamma^2(R_{t+3} + \alpha) + \cdots \mid s, a]$
$= E_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + (\alpha + \gamma\alpha + \gamma^2\alpha + \cdots) \mid s, a]$.
$= Q_\pi(\text{original}) + \alpha + \gamma\alpha + \gamma^2\alpha + \cdots$

we can tell that the new reward function doesn't change ~~where~~ $a = \arg\max q_*(s,a)$, so the modified MDP has the same policy as the original MDP.

2. $Q_\pi' = E_\pi[\beta R_{t+1} + \gamma\beta R_{t+2} + \gamma^2 \beta R_{t+3} \cdots \mid s, a]$
$= \beta Q_\pi(\text{original})$.

apparently, $\arg\max_\pi Q_\pi' = \arg\max_\pi Q_\pi(\text{original})$, so the modified MDP has the same policy as the original MDP.

Problem 6

1. State transition function:
$P_{s,\text{Done}}^{\text{stop}} = 1, \quad s \in \{1, 2, 3, 4, 5\}$

$P^{Draw}_{0,s'} = 0.5, \; s' \in \{2, 3\}$.

$P^{Draw}_{2,s'} = 0.5, \; s' \in \{4, 5\}$.

$P^{Draw}_{3,s'} = 0.5, \; s' \in \{5, done\}$.

$P^{Draw}_{s,done} = 1, \; s' \in \{4, 5\}$

reward function:

$$\begin{cases} R_s^{stop} = s, & s \in \{0, 2, 3, 4, 5\} \\ R_s^a = 0, & \text{others}. \end{cases}$$

2. Optimal state-Value function:

~~$Q_*(0, stop) = 0, \; Q_*(0, draw) =$~~

~~$Q_*(0,$~~ $Q_*(5, draw) = 0, \; Q_*(5, stop) = 5$.

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

$Q_*(3, draw) = 0.5 \times 5 + 0 \times 0.5 = 2.5, \; Q_*(3, stop) = 3$.

$Q_*(2, draw) = 0.5 \times 4 + 0.5 \times 5 = 4.5, \; Q_*(2, stop) = 2$.

~~$Q_*(1, draw) = 0.5 \times 2.5 + 0.5 \times 4 = 3.75, \; Q_*(1, stop) =$~~

$Q_*(0, draw) = 0.5 \times 4.5 + 0.5 \times 3 = 3.75, \; Q_*(0, stop) = 0$.

Optimal action-Value function:

$$V_*(s) = \begin{cases} 5, & s = 5 \\ 4, & s = 4 \\ 3, & s = 3 \\ 4.5, & s = 2 \\ 3.75, & s = 0. \end{cases}$$

3. From optimal state-value function, we can tell the optimal policy for this MDP is:

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