ENGG6600 Reinforcement Learning Assignment-2

Fall 2022

Yaowen Mei
1177855

Yaowen Mei
1177855

Homework 2

ENGG6600
November 13, 2022

Problem 1- First-Visit vs Every-Visit Monte Carlo

In the Mars Rover example in the lectures, use $\gamma = 1$. Assume the policy is given as below:

- TL in all the states
- S1, and S7 transition to terminal state upon any action.
- $$\eta = \underbrace{S_3, TL}_{t_0}, \underbrace{0, S_3, TL}_{t_1}, \underbrace{0, S_2, TL}_{t_2}, \underbrace{0, S_1, TL}_{t_3}, \underbrace{1, End}_{t_4}$$

	s1	**s2**	**s3**	**s4**	**s5**	**s6**	**s7**
r	1	0	0	0	0	0	10
$\boldsymbol{\pi}$	TL	TL	TL	TL	TL	TL	TL

Questions List:

1. Use **first-visit** Monte Carlo to estimate the value functions of all the states.
2. Use **every-visit** Monte Carlo to estimate the value functions of all the states.

SOLUTION-1:
Derive all the Returns for each time stamp ($i = 1$ for the first iteration):

$$G_{i,0}(S_3) = R_{i,1} + \gamma R_{i,2} + \gamma^2 R_{i,3} + \gamma^3 R_{i,4} = 1$$
$$G_{i,1}(S_3) = R_{i,2} + \gamma R_{i,3} + \gamma^2 R_{i,4} = 1$$
$$G_{i,2}(S_2) = R_{i,3} + \gamma R_{i,4} = 1$$
$$G_{i,3}(S_1) = R_{i,4} = 1$$

Along the trajectory, S3, S2, and S1 are visited.

	s1	**s2**	**s3**	**s4**	**s5**	**s6**	**s7**
$\hat{\boldsymbol{v}}(\boldsymbol{s})$	0	0	0	0	0	0	0
N(s)	0	0	0	0	0	0	0
G(s)	0	0	0	0	0	0	0

SOLUTION-1.1 First-Visited MC:
For S1:

$$N(S_1) = N(S_1) + 1 = 0 + 1 = 1$$
$$G(S_1) = G(S_1) + G_{i,3}(S_1) = 0 + 1 = 1$$
$$\hat{v}_\pi(S_1) = {}^{G(S_1)}/_{N(S_1)} = \tfrac{1}{1} = 1$$

For S2:

$$N(S_2) = N(S_2) + 1 = 0 + 1 = 1$$
$$G(S_2) = G(S_2) + G_{i,2}(S_2) = 0 + 1 = 1$$
$$\hat{v}_\pi(S_2) = {}^{G(S_2)}/_{N(S_2)} = \tfrac{1}{1} = 1$$

For S3 **(S3 were visited twice, but only the first time visit (t_0) evokes a training)**:

$$N(S_3) = N(S_3) + 1 = 0 + 1 = 1$$
$$G(S_3) = G(S_3) + G_{i,0}(S_3) = 0 + 1 = 1$$
$$\hat{v}_\pi(S_3) = {}^{G(S_3)}/_{N(S_3)} = \tfrac{1}{1} = 1$$

For S4 to S7, they were not learned anything from this trajectory.

	s1	**s2**	**s3**	**s4**	**s5**	**s6**	**s7**
$\boldsymbol{v^\pi(s)}$	1	1	1	0	0	0	0
N(s)	1	1	1	0	0	0	0
G(s)	1	1	1	0	0	0	0

Table 1: First-visit MC estimation of value functions for all states.

SOLUTION-1.2 Every-Visit MC:
For S1:

$$N(S_1) = N(S_1) + 1 = 0 + 1 = 1$$
$$G(S_1) = G(S_1) + G_{i,3}(S_1) = 0 + 1 = 1$$
$$\hat{v}_\pi(S_1) = {}^{G(S_1)}/_{N(S_1)} = \tfrac{1}{1} = 1$$

For S2:

$$N(S_2) = N(S_2) + 1 = 0 + 1 = 1$$
$$G(S_2) = G(S_2) + G_{i,2}(S_2) = 0 + 1 = 1$$
$$\hat{v}_\pi(S_2) = {}^{G(S_2)}/_{N(S_2)} = \tfrac{1}{1} = 1$$

For S3 **(S3 were visited twice, so it will be trained twice**:

The first time visit S3:

$$\begin{aligned}
&N\left(S_3\right)=N\left(S_3\right)+1=0+1=1\\
&G\left(S_3\right)=G\left(S_3\right)+G_{i,0}\left(S_3\right)=0+1=1\\
&\hat{v}_\pi\left(S_3\right)={}^{G\left(S_3\right)}\!/_{N\left(S_3\right)}=\tfrac{1}{1}=1
\end{aligned}$$

The second time visit S3:

$$\begin{aligned}
&N\left(S_3\right)=N\left(S_3\right)+1=1+1=2\\
&G\left(S_3\right)=G\left(S_3\right)+G_{i,1}\left(S_3\right)=1+1=2\\
&\hat{v}_\pi\left(S_3\right)={}^{G\left(S_3\right)}\!/_{N\left(S_3\right)}=\tfrac{2}{2}=1
\end{aligned}$$

For S4 to S7, they were not learned anything from this trajectory.

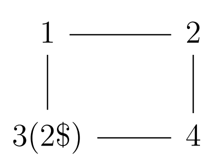
	s1	**s2**	**s3**	**s4**	**s5**	**s6**	**s7**
$\boldsymbol{v^\pi(s)}$	1	1	1	0	0	0	0
N(s)	1	1	2	0	0	0	0
G(s)	1	1	2	0	0	0	0

Table 2: Every-visit MC estimation of value function for all states.

Problem 2–Q Learning vs SARSA

Agent A lives in a 2x2 grid as shown below:

The states correspond to the numbered squares. Her possible actions are **MoveNorth, MoveSouth, MoveEast, MoveWest**.



The agent earns $2 every time she lands in state 3. There are no other rewards or penalties. The reward functions $R(s,a)$ are given. Note that this is a continuing task

Assuming learning rate $\alpha = 1$ and discount rate $\gamma = 0.9$.

Reward $R(s,a)$	**1**	**2**	**3**	**4**
MoveNorth	0	0	0	0
MoveSouth	**2**	0	0	0
MoveEast	0	0	0	0
MoveWest	0	0	0	**2**

State Value Q(s, a)	**1**	**2**	**3**	**4**
MoveNorth	0	0	0	0
MoveSouth	0	0	0	0
MoveEast	0	0	0	0
MoveWest	0	0	0	0

Questions List:

1. Agent A starts in square 1 and performs the following actions: MoveEast, MoveSouth, MoveWest, MoveNorth. After each action, the Q-Table is updated using **Q-Learning**, with the the usual update formula:

$$q\left(s,a\right) \Leftarrow q\left(s,a\right) + \alpha\left(r\left(s,a\right) + \gamma \max_{a'} q\left(s',a'\right) - q\left(s,a\right)\right)$$

$$\begin{tikzcd}
1 \arrow[r, "MoveEast"] & 2 \arrow[d, "MoveSouth"] \\
3 \arrow[u, "MoveNorth"] & 4 \arrow[l, "MoveWest"]
\end{tikzcd}$$

2. Agent A **continues** in square 1 and performs the following actions: MoveSouth, MoveEast,MoveNorth, MoveWest. After each of the first three action, the Q-Table is updated using **SARSA**, with the the usual update formula:

$$q\left(s,a\right) \Leftarrow q\left(s,a\right) + \alpha\left(r\left(s,a\right) + \gamma q\left(s',a'\right) - q\left(s,a\right)\right)$$

$$\begin{tikzcd}
1 \arrow[d, "MoveSouth"] & 2 \arrow[l, "MoveWest"] \\
3 \arrow[r, "MoveEast"] & 4 \arrow[u, "MoveNorth"]
\end{tikzcd}$$

Calculate the Q Table after the last update.

SOLUTION-2-1 Q-Learning:

- Initial state in Episode 0, First action,
- Current State s = 1,

- Current Action a = MoveEast,
- Next State $s' = 2$
- Update $q(1, MoveEast)$

$$\begin{aligned} q\left(1, \mathrm{ME}\right) &\Leftarrow q\left(1, \mathrm{ME}\right) + \alpha \left(r\left(1, \mathrm{ME}\right) + \gamma Max \begin{bmatrix} q\left(2, \mathrm{MN}\right) \\ q\left(2, \mathrm{MS}\right) \\ q\left(2, \mathrm{ME}\right) \\ q\left(2, \mathrm{MW}\right) \end{bmatrix} - q\left(1, \mathrm{ME}\right) \right) \\ &= 0 + 1 \times (0 + 0.9 \times 0 - 0) = 0 \end{aligned}$$

- Episode 0, Second action
- Current State s = 2,
- Current Action a = MoveSouth,
- Next State $s' = 4$
- Update $q(2, MoveSouth)$

$$\begin{aligned} q\left(2, \mathrm{MS}\right) &\Leftarrow q\left(2, \mathrm{MS}\right) + \alpha \left(r\left(2, \mathrm{MS}\right) + \gamma \mathrm{Max} \begin{bmatrix} q\left(4, \mathrm{MN}\right) \\ q\left(4, \mathrm{MS}\right) \\ q\left(4, \mathrm{ME}\right) \\ q\left(4, \mathrm{MW}\right) \end{bmatrix} - q\left(2, \mathrm{MS}\right) \right) \\ &= 0 + 1 \times (0 + 0.9 \times 0 - 0) = 0 \end{aligned}$$

- Episode 0, Third action
- Current State s = 4,
- Current Action a = MoveWest,
- Next State $s' = 3$
- Update $q(4, MoveWest)$

$$\begin{aligned} q\left(4, \mathrm{MW}\right) &\Leftarrow q\left(4, \mathrm{MW}\right) + \alpha \left(r\left(4, \mathrm{MW}\right) + \gamma \mathrm{Max} \begin{bmatrix} q\left(3, \mathrm{MN}\right) \\ q\left(3, \mathrm{MS}\right) \\ q\left(3, \mathrm{ME}\right) \\ q\left(3, \mathrm{MW}\right) \end{bmatrix} - q\left(4, \mathrm{MW}\right) \right) \\ &\Leftarrow 0 + 1 \times (2 + 0.9 \times 0 - 0) = 2 \end{aligned}$$

- Episode 0, Forth action (last action)

- Current State s = 3,
- Current Action a = MoveNorth,
- Next State $s' = 1$
- Update $q(3, MoveNorth)$

$$\begin{aligned} q\left(3, \mathrm{MN}\right) &\Leftarrow q\left(3, \mathrm{MN}\right) + \alpha \left(r\left(3, \mathrm{MN}\right) + \gamma \mathrm{Max} \begin{bmatrix} q\left(1, \mathrm{MN}\right) \\ q\left(1, \mathrm{MS}\right) \\ q\left(1, \mathrm{ME}\right) \\ q\left(1, \mathrm{MW}\right) \end{bmatrix} - q\left(3, \mathrm{MN}\right) \right) \\ &\Leftarrow 0 + 1 \times (0 + 0.9 \times 0 - 0) = 0 \end{aligned}$$

Q-Learning Q Table	**1**	**2**	**3**	**4**
MoveNorth	0	0	0	0
MoveSouth	0	0	0	0
MoveEast	0	0	0	0
MoveWest	0	0	0	**2**

Table 3: Q-Table after episode-0's last update via Q-Learning method.

SOLUTION-2-2 SARSA:

- Episode 1, First action under SARSA,
- Current State s = 1,
- Current Action a = MoveSouth,
- Next State $s' = 3$
- Update $q(1, MoveSouth)$
- Next Action a' = MoveEast

$$q\left(1, MoveSouth\right) \Leftarrow q\left(1, MoveSouth\right) + \alpha\left(r\left(1, MoveSouth\right) + \gamma q\left(3, MoveEast\right) - q\left(1, MoveSouth\right)\right)$$
$$q\left(1, MoveSouth\right) \Leftarrow 0 + 1 \times (2 + 0.9 \times 0 - 0) = 2$$

- Episode 1, Second action under SARSA,
- Current State s = 3,
- Current Action a = MoveEast,

- Next State $s' = 4$
- Update $q(3, MoveEast)$
- Next Action a' = MoveNorth

$$q\left(3, MoveEast\right) \Leftarrow q\left(3, MoveEast\right) + \alpha\left(r\left(3, MoveEast\right) + \gamma q\left(4, MoveNorth\right) - q\left(3, MoveEast\right)\right)$$
$$q\left(3, MoveEast\right) \Leftarrow 0 + 1 \times \left(0 + 0.9 \times 0 - 0\right) = 0$$

- Episode 1, Second action under SARSA,
- Current State s = 4,
- Current Action a = MoveNorth,
- Next State $s' = 2$
- Update $q(4, MoveNorth)$
- Next Action a' = MoveWest

$$q\left(4, MoveNorth\right) \Leftarrow q\left(4, MoveNorth\right) + \alpha\left(r\left(4, MoveNorth\right) + \gamma q\left(2, MoveWest\right) - q\left(4, MoveNorth\right)\right)$$
$$q\left(4, MoveNorth\right) \Leftarrow 0 + 1 \times \left(0 + 0.9 \times 0 - 0\right) = 0$$

SARSA Q Table	**1**	**2**	**3**	**4**
MoveNorth	0	0	0	0
MoveSouth	**2**	0	0	0
MoveEast	0	0	0	0
MoveWest	0	0	0	2

Table 4: Q-Table after episode-1's last update via SARSA.