ENGG*6600-01 ST: Reinforcement Learning
Written Assignment #2

Due Time: Friday, Oct. 28, 2022

Question 1 In the Mars Rover example in the lectures, use $\gamma = 1$. Assume the policy is given as: TL in all the states. S1 and S7 transition to terminal state upon any action. Given the Trajectory (S3 TL 0 S3 TL 0 S2 TL 0 S1 TL 1 Terminal), use **first-visit** and **every-visit** Monte Carlo algorithms to estimate the value functions of all the states (initial values are zero), respectively.

S1	S2	S3	S4	S5	S6	S7
Okay Field Site +1						Fantastic Field Site +10

Figure 1 Mars Rover Policy Evaluation

(1) First-visit MC:

s	$S1$	$S2$	$S3$	$S4$	$S5$	$S6$	$S7$
$V^{\pi}(s)$							

Show your working here:

(2) Every-visit MC

s	$S1$	$S2$	$S3$	$S4$	$S5$	$S6$	$S7$
$V^{\pi}(s)$							

Show your working here:

Question 2: Agent A lives in a 2×2 grid as shown in Figure 2

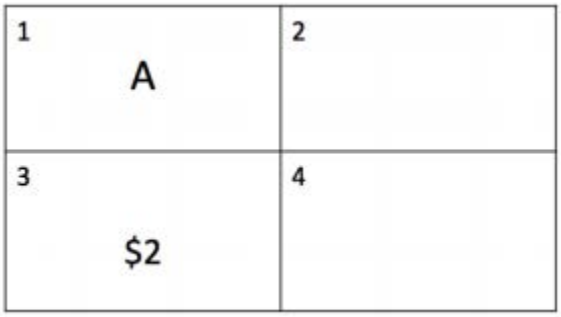

Figure 2 2×2 grid

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 in Table 1. Note that this is a continuing task.

Table 1 Reward functions $R(s, a)$

	1	2	3	4
MoveNorth	0	0	0	0
MoveSouth	$2	0	0	0
MoveEast	0	0	0	0
MoveWest	0	0	0	$2

The Q matrix is initialized to all zeros as shown in Table 2.

Table 2 Initial Q values $Q(s, a)$

	1	2	3	4
MoveNorth	0	0	0	0
MoveSouth	0	0	0	0
MoveEast	0	0	0	0
MoveWest	0	0	0	0

(a) 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 usual update formula:

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

Assuming a learning rate $\alpha = 1$ and discount rate $\gamma = 0.9$, give the nonzero entries of the Q-table after the last update.

Updated Q table:

	1	2	3	4
MoveNorth				
MoveSouth				
MoveEast				
MoveWest				

Show your working here:

(b) Agent A continues from square 1 and performs the following actions: MoveSouth, MoveEast, MoveNorth, MoveWest. After each of the first three actions, the Q-table is updated using SARSA, with the usual update formula:

$$q(s,a) \leftarrow q(s,a) + \alpha(r(s,a) + \gamma q(s',a') - q(s,a))$$

Assuming a learning rate $\alpha = 1$ and discount rate $\gamma = 0.9$, give the nonzero entries of the Q-table after the last update.

Updated Q table:

	1	2	3	4
MoveNorth				
MoveSouth				
MoveEast				
MoveWest				

Show your working here: