Q9: Reinforcement Learning

Q9: Reinforcement Learning

Imagine an unknown game which has only two states $\{A, B\}$ and in each state the agent has two actions to choose from: $\{Up, Down\}$. Suppose a game agent chooses actions according to some policy π and generates the following sequence of actions and rewards in the unknown game:

t	s_t	a_t	s_{t+1}	r_t
0	A	Down	B	2
1	B	Down	B	-4
2	B	Up	B	0
3	B	Up	A	3
4	A	Up	A	-1

Unless specified otherwise, assume a discount factor $\gamma = 0.5$ and a learning rate $\alpha = 0.5$

Part 1

0.0/2.0 points (graded)

Recall the update function of Q-learning is:

$$Q\left(s_t, a_t\right) \leftarrow (1-\alpha)\, Q\left(s_t, a_t\right) + \alpha\left(r_t + \gamma \max_{a'} Q\left(s_{t+1}, a'\right)\right)$$

Assume that all Q-values initialized as 0. What are the following Q-values learned by running

Generating Speech Output ove experience sequence?

$Q(A, Down)$

1

Answer: 1

$Q(B, Up)$

7/4

Answer: 7/4

Explanation

Perform Q-learning update 4 times, once for each of the first 4 observations.

Submit

You have used 0 of 2 attempts

Part 2

0.0/2.0 points (graded)

In model-based reinforcement learning, we first estimate the transition function $T(s, a, s')$ and the reward function $R(s, a, s')$. Fill in the following estimates of T and R, estimated from the experience above. Write "x" if not applicable or undefined.

$\hat{T}(A, Up, A)$

1

Answer: 1

$\hat{T}(A, Up, B)$

0

Answer: 0

$\hat{T}(B, Up, A)$

0.5

Answer: 0.5

$\hat{T}(B, Up, B)$

Generating Speech Output

0.5

Answer: 0.5

$\hat{R}(A, Up, A)$

-1

Answer: -1

$\hat{R}(A, Up, B)$

x

Answer: x

$\hat{R}(B, Up, A)$

3

Answer: 3

$\hat{R}(B, Up, B)$

0

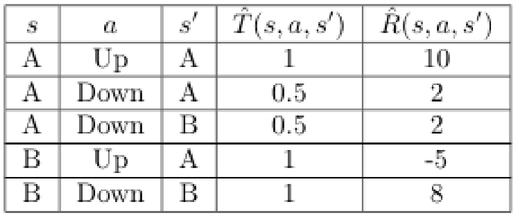
Answer: 0

Explanation
Count transitions above and calculate frequencies. Rewards are observed rewards.

Submit

You have used 0 of 2 attempts

To decouple this question from the previous one, assume we had **a different experience** and ended up with the following estimates of the transition and reward functions:



s	a	s'	$\hat{T}(s,a,s')$	$\hat{R}(s,a,s')$
A	Up	A	1	10
A	Down	A	0.5	2
A	Down	B	0.5	2
B	Up	A	1	-5
B	Down	B	1	8

Generating Speech Output *ed on the above experience.*

Part 3

0.0/1.0 point (graded)

Give the optimal policy $\hat{\pi}^*(s)$ for the MDP with transition function $\hat{T}$ and reward function $\hat{R}$
.

$\hat{\pi}^*(A)$

Up ▾ **Answer:** Up

$\hat{\pi}^*(B)$

Down ▾ **Answer:** Down

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

Part 4

0.0/1.0 point (graded)

Give $\hat{V}^*(s)$ for the MDP with transition function $\hat{T}$ and reward function $\hat{R}$.
Hint: for any $x \in \mathbb{R}$, $|x| < 1$, *we have* $1 + x + x^2 + x^3 + x^4 + \cdots = 1/(1 - x)$.

$\hat{V}^*(A)$

20 **Answer:** 20

$\hat{V}^*(B)$

16 **Answer:** 16

Explanation

Based on the optimal policy in Part 3, calculate the value function using a Bellman equation.

Generating Speech Output

Submit You have used 0 of 2 attempts

Part 5

0.0/2.0 points (graded)

If we repeatedly feed this new experience sequence through our Q-learning algorithm, what values will it converge to? Assume the learning rateα_t is properly chosen so that convergence is guaranteed.

● the values found above, $\hat{V}^*$ ✓

○ the optimal values, V^*

○ neither $\hat{V}^*$ nor V^*

○ not enough information to determine

Explanation

The Q-learning algorithm will not converge to the optimal values V^* for the MDP because the experience sequence and transition frequencies replayed are not necessarily representative of the underlying MDP. (For example, the true $T\left(A, Down, A\right)$ might be equal to 0.75, in which case, repeatedly feeding in the above experience would not provide an accurate sampling of the MDP.) However, for the MDP with transition function $\hat{T}$ and reward function $\hat{R}$, replaying this experience repeatedly will result in Q-learning converging to its optimal values $\hat{V}^*$.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

Generating Speech Output