CSPB 3202 - Truong - Artificial Intelligence

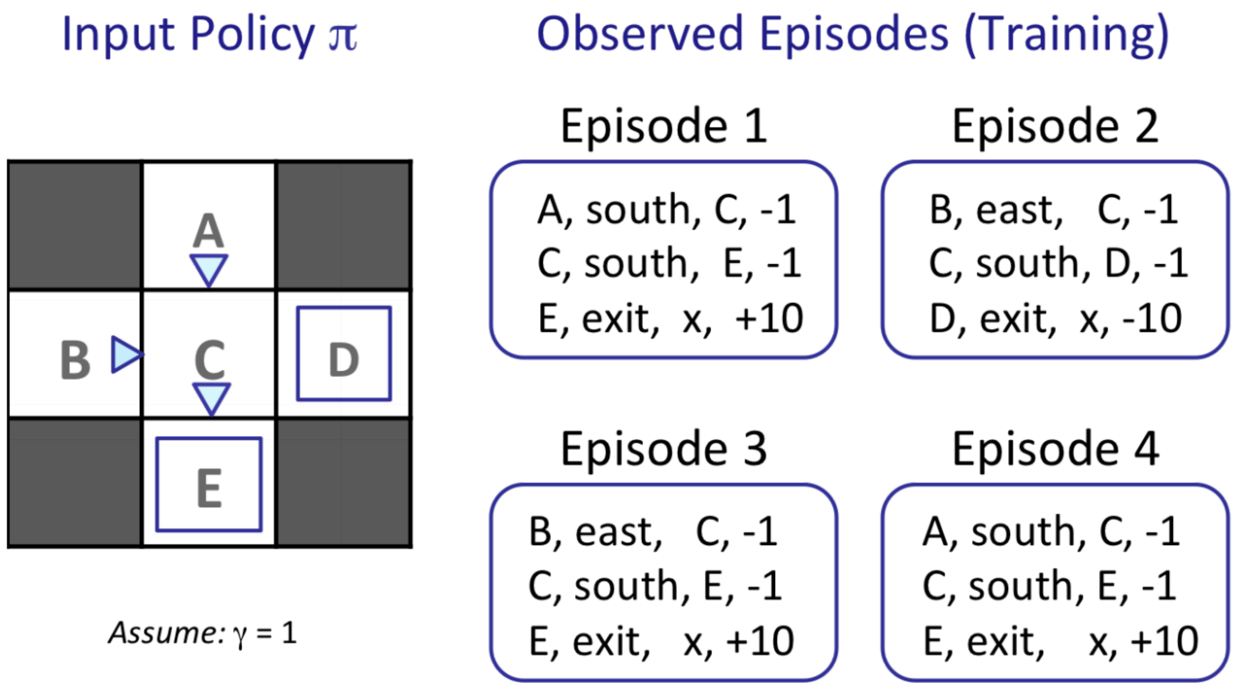
Dashboard / My courses / 2244:CSPB 3202 / 1 July - 7 July / Reading Quiz 6-RL

Started on	Wednesday, 3 July 2024, 6:04 PM
State	Finished
Completed on	Wednesday, 3 July 2024, 6:15 PM
Time taken	10 mins 51 secs

Question **1**

Correct

Marked out of 4.00



Input Policy π

Assume: γ = 1

Observed Episodes (Training)

Episode 1

A, south, C, -1
C, south, E, -1
E, exit, x, +10

Episode 2

B, east, C, -1
C, south, D, -1
D, exit, x, -10

Episode 3

B, east, C, -1
C, south, E, -1
E, exit, x, +10

Episode 4

A, south, C, -1
C, south, E, -1
E, exit, x, +10

What model would be learned from the above observed episodes?

T(A, south, C) = 1 ✓

T(B, east, C) = 1 ✓

T(C, south, E) = 0.75 ✓

T(C, south, D) = 0.25 ✓

Question **2**

Correct

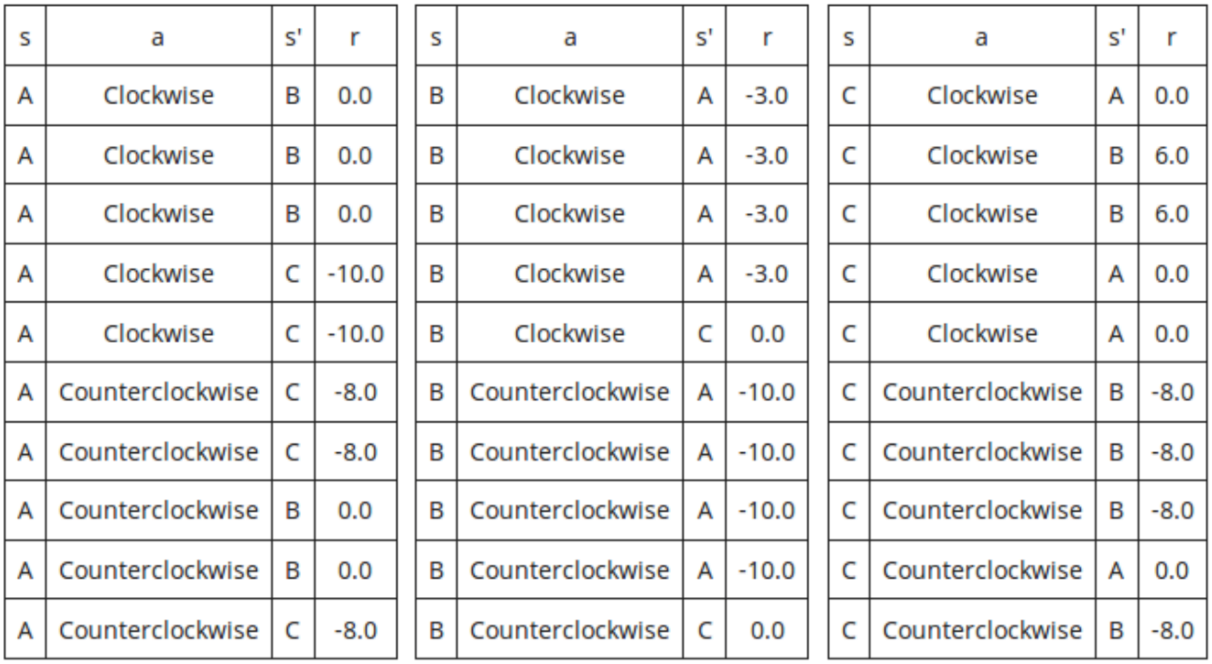
Marked out of 13.00

We recommend you work out the solutions to the following questions on a sheet of scratch paper, and then enter your results into the answer boxes.

Consider an MDP with 3 states, A, B and C; and 2 actions Clockwise and Counterclockwise. We do not know the transition function or the reward function for the MDP, but instead, we are given samples of what an agent experiences when it interacts with the environment (although, we do know that we do not remain in the same state after taking an action). In this problem, we will first estimate the model (the transition function and the reward function), and then use the estimated model to find the optimal actions.

To find the optimal actions, model-based RL proceeds by computing the optimal V or Q value function with respect to the estimated T and R. This could be done with any of value iteration, policy iteration, or Q-value iteration. Last week you already solved some exercises that involved value iteration and policy iteration, so we will go with Q value iteration in this exercise.

Consider the following samples that the agent encountered.



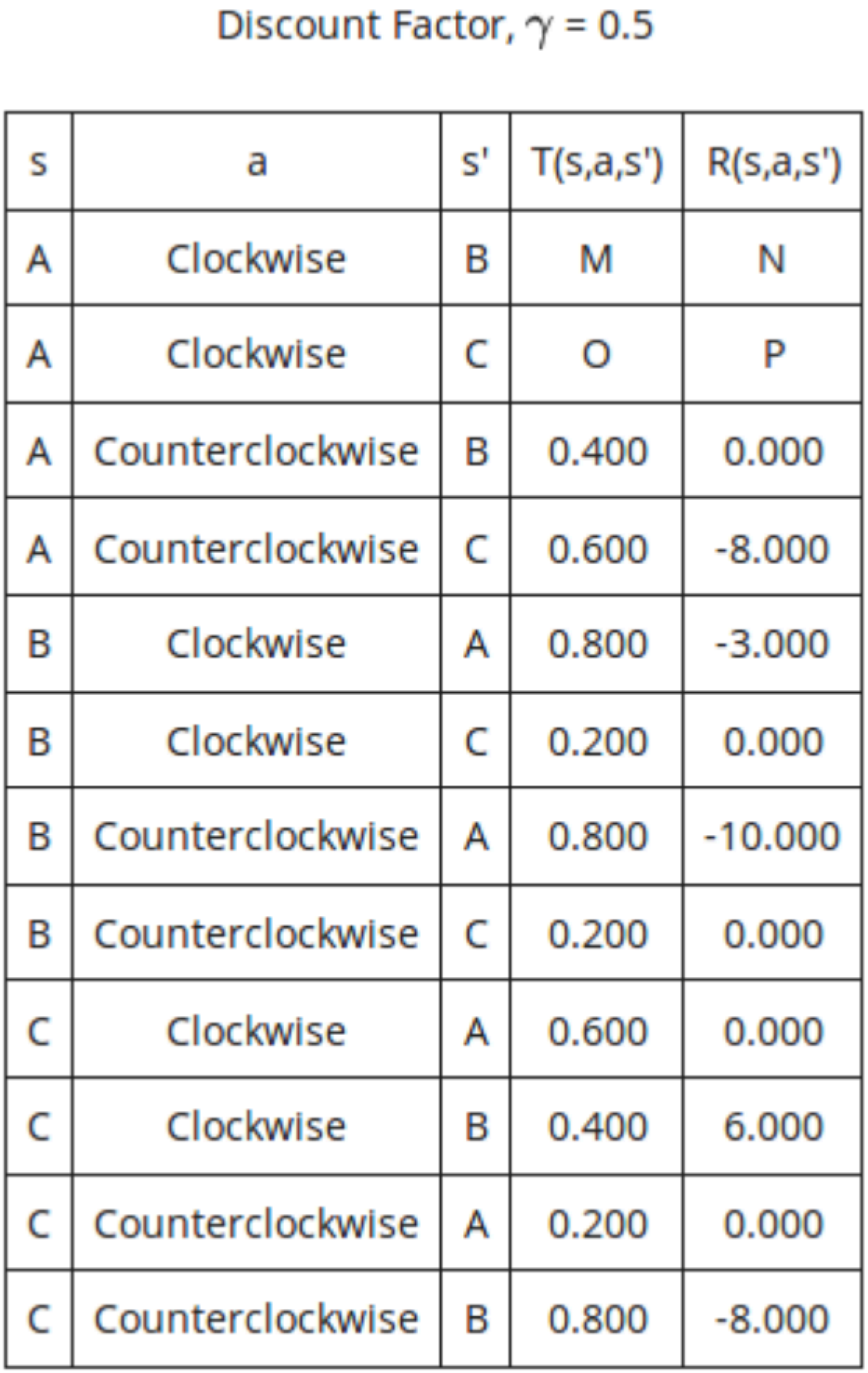
s	a	s'	r
A	Clockwise	B	0.0
A	Clockwise	B	0.0
A	Clockwise	B	0.0
A	Clockwise	C	-10.0
A	Clockwise	C	-10.0
A	Counterclockwise	C	-8.0
A	Counterclockwise	C	-8.0
A	Counterclockwise	B	0.0
A	Counterclockwise	B	0.0
A	Counterclockwise	C	-8.0

s	a	s'	r
B	Clockwise	A	-3.0
B	Clockwise	A	-3.0
B	Clockwise	A	-3.0
B	Clockwise	A	-3.0
B	Clockwise	C	0.0
B	Counterclockwise	A	-10.0
B	Counterclockwise	A	-10.0
B	Counterclockwise	A	-10.0
B	Counterclockwise	A	-10.0
B	Counterclockwise	C	0.0

s	a	s'	r
C	Clockwise	A	0.0
C	Clockwise	B	6.0
C	Clockwise	B	6.0
C	Clockwise	A	0.0
C	Clockwise	A	0.0
C	Counterclockwise	B	-8.0
C	Counterclockwise	B	-8.0
C	Counterclockwise	B	-8.0
C	Counterclockwise	A	0.0
C	Counterclockwise	B	-8.0

Q1:

We start by estimating the transition function, T(s,a,s') and reward function R(s,a,s') for this MDP. Fill in the missing values in the following table for T(s,a,s') and R(s,a,s').



Discount Factor, $\gamma = 0.5$

s	a	s'	T(s,a,s')	R(s,a,s')
A	Clockwise	B	M	N
A	Clockwise	C	O	P
A	Counterclockwise	B	0.400	0.000
A	Counterclockwise	C	0.600	-8.000
B	Clockwise	A	0.800	-3.000
B	Clockwise	C	0.200	0.000
B	Counterclockwise	A	0.800	-10.000
B	Counterclockwise	C	0.200	0.000
C	Clockwise	A	0.600	0.000
C	Clockwise	B	0.400	6.000
C	Counterclockwise	A	0.200	0.000
C	Counterclockwise	B	0.800	-8.000

M = 0.6 ✓

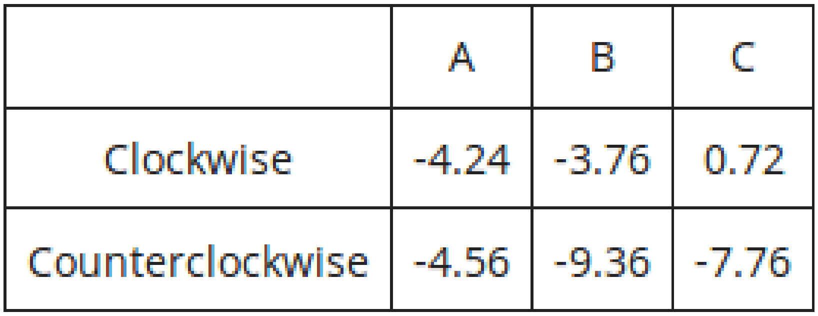
N = 0 ✓

O = 0.4 ✓

P = -10 ✓

Q2:

Now we will run Q-iteration using the estimated T and R functions. The values of $Q_k(s, a)$, are given in the table below.



	A	B	C
Clockwise	-4.24	-3.76	0.72
Counterclockwise	-4.56	-9.36	-7.76

Fill in the values for $Q_{k+1}(s, a)$.

Q(A, clockwise) = -4.984 ✓

Q(A, counterclockwise) = -5.336 ✓

Q(B, clockwise) = -4.024 ✓

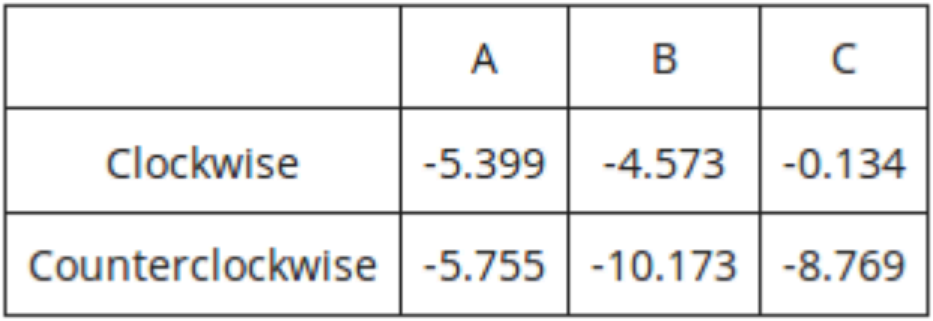
Q(B, counterclockwise) = -9.624 ✓

Q(C, clockwise) = 0.376 ✓

Q(C, counterclockwise) = -8.328 ✓

Q3:

Suppose Q-iteration converges to the following Q^* function, $Q^*(s, a)$.



	A	B	C
Clockwise	-5.399	-4.573	-0.134
Counterclockwise	-5.755	-10.173	-8.769

What is the optimal action, either Clockwise or Counterclockwise, for each of the states?

A? Clockwise ✓

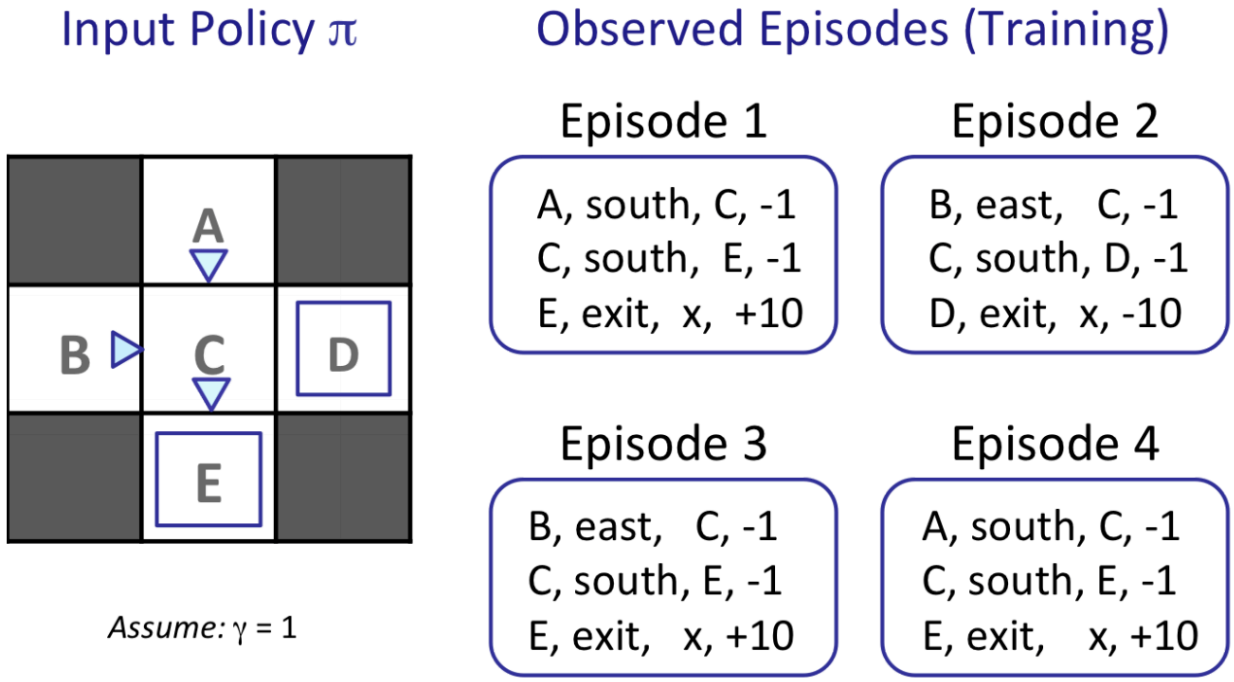
B? Clockwise ✓

C? Clockwise ✓

Question **3**

Correct

Marked out of 5.00



Input Policy π

Assume: γ = 1

Observed Episodes (Training)

Episode 1

A, south, C, -1
C, south, E, -1
E, exit, x, +10

Episode 2

B, east, C, -1
C, south, D, -1
D, exit, x, -10

Episode 3

B, east, C, -1
C, south, E, -1
E, exit, x, +10

Episode 4

A, south, C, -1
C, south, E, -1
E, exit, x, +10

What are the estimates for the following quantities as obtained by direct evaluation:

$V^{\pi}(A)$ = 8 ✓

$V^{\pi}(B)$ = -2 ✓

$V^{\pi}(C)$ = 4 ✓

$V^{\pi}(D)$ = -10 ✓

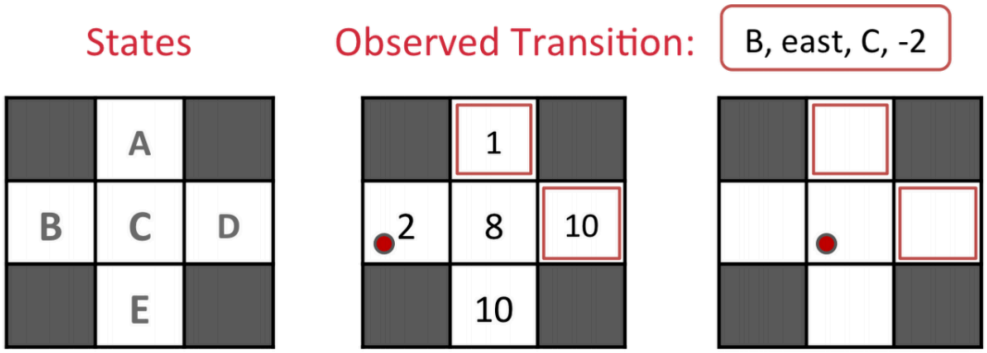
$V^{\pi}(E)$ = 10 ✓

Question **4**

Correct

Marked out of 5.00

Consider the gridworld shown below. The left panel shows the name of each state A through E. The middle panel shows the current estimate of the value function V^{π} for each state. A transition is observed, that takes the agent from state B through taking action east into state C, and the agent receives a reward of -2. Assuming $\gamma = 1, \alpha = 0.5$, what are the value estimates after the TD learning update? (note: the value will change for one of the states only).



Assume: γ = 1, α = 1/2

$$V^{\pi}(s) \leftarrow (1-\alpha)V^{\pi}(s) + \alpha\left[R(s,\pi(s),s') + \gamma V^{\pi}(s')\right]$$

$V^{\pi}(A)$ = 1 ✓

$V^{\pi}(B)$ = 4 ✓

$V^{\pi}(C)$ = 8 ✓

$V^{\pi}(D)$ = 10 ✓

$V^{\pi}(E)$ = 10 ✓

Question **5**

Correct

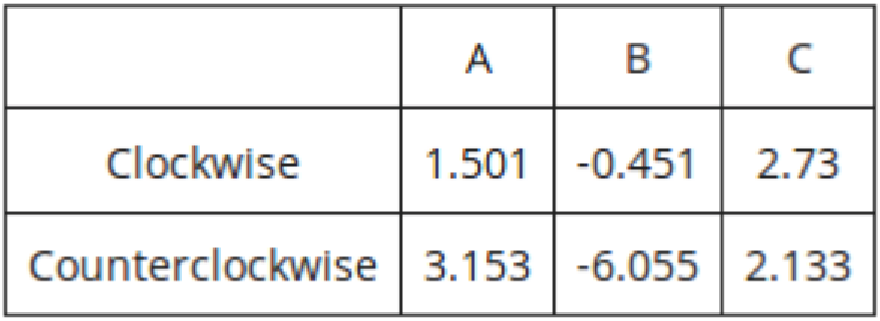
Marked out of 6.00

We recommend you work out the solutions to the following questions on a sheet of scratch paper, and then enter your results into the answer boxes.

Consider an MDP with 3 states, A, B and C; and 2 actions Clockwise and Counterclockwise. We do not know the transition function or the reward function for the MDP, but instead, we are given with samples of what an agent actually experiences when it interacts with the environment (although, we do know that we do not remain in the same state after taking an action). In this problem, instead of first estimating the transition and reward functions, we will directly estimate the Q function using Q-learning.

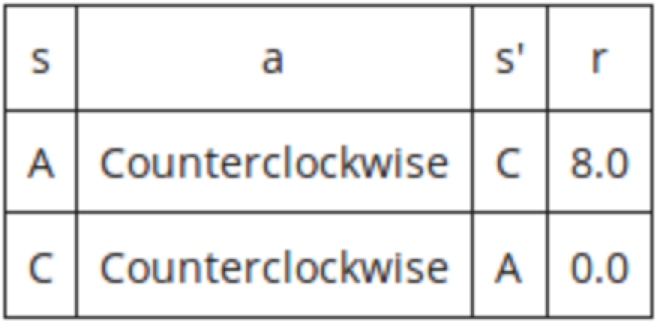
Assume, the discount factor, γ is 0.5 and the step size for Q-learning, α is 0.5.

Our current Q function, , is as follows.



	A	B	C
Clockwise	1.501	-0.451	2.73
Counterclockwise	3.153	-6.055	2.133

The agent encounters the following samples.



s	a	s'	r
A	Counterclockwise	C	8.0
C	Counterclockwise	A	0.0

Process the samples given above. Below fill in the Q-values after both samples have been accounted for.

Q(A, clockwise) = 1.501 ✓

Q(A, counterclockwise) = 6.259 ✓

Q(B, clockwise) = -0.451 ✓

Q(B, counterclockwise) = -6.055 ✓

Q(C, clockwise) = 2.73 ✓

Q(C, counterclockwise) = 2.63125 ✓

Question **6**

Correct

Marked out of 2.00

In general, for Q-Learning to converge to the optimal Q-values...

☐

a. It is necessary that the discount γ is less than 0.5

☑

b. It is necessary that every state-action pair is visited infinitely often.

✓

☐

c. It is necessary that actions get chosen according to $\arg\max_a Q(s,a)$.

☑

d. It is necessary that the learning rate α (weight given to new samples) is decreased to 0 over time.

✓

Question **7**

Correct

Marked out of 4.00

For each of the following action-selection methods, indicate which option describes it best.

A: With probability p , select $argmax_a Q(s,a)$. With probability $1-p$, select a random action. $p = 0.99$

Mostly exploitation ✓

B: Select action a with probability P(a|s) = $\frac{e^{Q(s,a)/\tau}}{\sum_{a'} e^{Q(s,a')/\tau}}$, where τ is a temperature parameter that is decreased over time .

Mix of Both ✓

C: Always select a random action.

Mostly exploration ✓

D: Keep track of a count, $K_{s,a}$, for each state-action tuple, (s,a), of the number of times that tuple has been seen and select $argmax_a[Q(s,a) - K_{s,a}]$

Mix of Both ✓

Question **8**

Correct

Marked out of 2.00

Which of the following method(s) would be advisable to use when doing Q-Learning?

□

a. With probability p , select $argmax_a Q(s,a)$. With probability $1 - p$, select a random action. $p = 0.99$

☑

b. Select action a with probability P(a|s) = $\frac{e^{Q(s,a)/\tau}}{\sum_{a'} e^{Q(s,a')/\tau}}$, where τ is a temperature parameter that is decreased over time .

✓

□

c. Always select a random action.

☑

d. Keep track of a count, $K_{s,a}$, for each state-action tuple, (s,a), of the number of times that tuple has been seen and select $argmax_a[Q(s,a) - K_{s,a}]$

✓