HW3

This is a preview of the published version of the quiz

Started: Aug 15 at 7:05pm

Quiz Instructions

Question 1 **0 pts**

Honor Pledge: Please type your name as a signature in the form.

I avow that I will not give or receive any unauthorized help on this exam, and that all work will be my own.

Edit View Insert Format Tools Table

12pt Paragraph

p 0 words

Question 2 2 pts

Answer the following question?

1. Explain how the ε-greedy approach balances exploration and exploitation.
2. Explain how the incremental mean method in the utility mean update reduces the memory usage.
3. Explain the difference between the Monte-Carlo (MC) method and the Temporal-Difference (TD) method.

4. Explain how the temporal-difference (TD) method in the utility update reduces the memory usage.
5. Explain in what occasion the TD Q-value update might result different results under SARSA learning and Q-Learning, respectively.

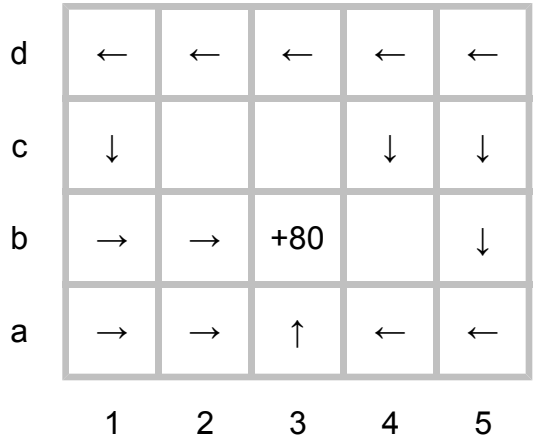
Edit View Insert Format Tools Table

12pt Paragraph

p 0 words

Question 3 2 pts

We extend the windy maze defined in HW1 with probabilistic outcome after an action. The maze map is shown here for your convenience.



The agent can drift to the left or the right with probability 0.2 or go straight with probability 0.6. If the drifting direction is a wall, it will be bounced back to the original position. We consider the policy defined in the map. Calculate which open squares can be reached from 4d after three sequential actions under this policy and with what probabilities.

Edit View Insert Format Tools Table

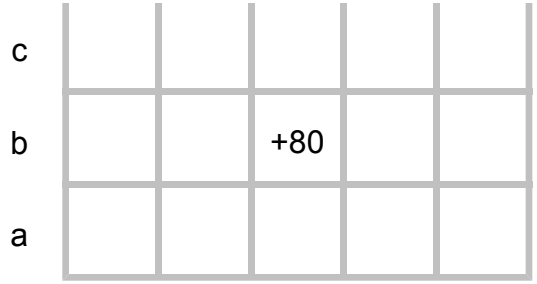
12pt Paragraph

p 0 words

Question 4 4 pts

We consider Value Iteration for the above MDP problem with the same drift probability assumption but without the given policy as shown here.

d



We still assume the wind comes from the north and the cost of one step is defined as follows (the reward will be the negation of the cost): 1 for moving southward; 2 for moving eastward or westward; 3 for moving northward. The reward function R(s,a) will be the negative of the cost. We also assume that the reward reaching at the goal is 80. Since the reward function R(s; a) here depends on both the state and the action taken at this state, all utility equations should be revised as

$U(s) \leftarrow max_a\ (R(s,a) + \gamma\ \Sigma_{s'}\ P(s'|s,a)\ U(s'))$

We choose γ=1. We assume the initial utility at any state is 0 except for 80 at the goal state. We perform one-iteration update of the utility of all states in this order: a1, a2, ..., a5, b1, b2, ..., b5, c1, ..., c5, d1, d2, ..., d5. Please give the updated utility at each state.

Edit View Insert Format Tools Table

12pt Paragraph

p 0 words

Question 5 3 pts

We consider SARSA with ε-Greedy algorithm for the above MDP problem. However, the model including transition probabilities and reward functions is unknown to the agent. The following tables show the latest values of *N(s,a)* (top) and *Q(s,a)* (bottom):

c	6	3	1	

	10 5 11	5 12 3	0 24 1	
b	5 18 9 13		11 55 2 4	
a	6 8 19 6	7 3 15 4	23 1 8 4	4 7 8 74
	1	2	3	4

c	-2.4 -2.4 16.4 -2.4	-1.8 -1.8 60.8 -1.8	41.0 0.0 87.8 3.4	+100
b	-2.9 -2.9 -2.9 -2.9		44.8 12.2 -101.0 -51.5	-100
a	-2.6 -2.5 -2.3 -2.4	-1.8 -2.0 0.4 -2.1	12.0 -2.0 -2.1 -1.9	-101.0 -16.0 -14.9 -3.5
	1	2	3	4

Then we run an additional trial in sequence (drifting might happen or random action might be chosen during the trial): ***a1,E,-2; a2,N,-1; a2,E,-2; a3,N,-1; a4,S,-3; a4,W,-2; a3,N,-1; b3,N,-1; c3,E,-2; c4***. The agent performs TD updates sequentially regarding this trial with the following equations:

$$N(s,a) \leftarrow N(s,a) + 1$$

$$Q(s,a) \leftarrow Q(s,a) + 1/N(s,a)\ (R(s,a) + \gamma Q(s',a') - Q(s,a))$$

We choose $\gamma=1$. Please show the updated Q-values $Q(s,a)$ during this trial (Keep two decimal values for each Q-value):

1. Q(a1,E):
2. Q(a2,N):
3. Q(a2,E):
4. Q(a3,N):
5. Q(a4,S):
6. Q(a4,W):
7. Q(a3,N):
8. Q(b3,N):
9. Q(c3,E):

Edit View Insert Format Tools Table

12pt Paragraph

p 0 words

Question 6 3 pts

We replace SARSA learning with Q-Learning in the above question and redo the question, where you should use the following TD Q-value updates:

$N(s,a) \leftarrow N(s,a) + 1$

$Q(s,a) \leftarrow Q(s,a) + 1/N(s,a)\ (R(s,a) + \gamma\ max_{a'}\ Q(s',a') - Q(s,a))$

Not saved

Submit Quiz