CS 188 Fall 2021 — Introduction to Artificial Intelligence — Challenge Q10 HW5

- **Due:** Tuesday 10/5 at 10:59pm.
- **Policy:** Can be solved in groups (acknowledge collaborators) but must be submitted individually.
- **Make sure to show all your work and justify your answers**.
- **Note:** This is a typical exam-level question. On the exam, you would be under time pressure, and have to complete this question on your own. We strongly encourage you to first try this on your own to help you understand where you currently stand. Then feel free to have some discussion about the question with other students and/or staff, before independently writing up your solution.
- Your submission on Gradescope should be a PDF that matches this template. Each page of the PDF should align with the corresponding page of the template (page 1 has name/collaborators, question begins on page 2.). **Do not reorder, split, combine, or add extra pages**. The intention is that you print out the template, write on the page in pen/pencil, and then scan or take pictures of the pages to make your submission. You may also fill out this template digitally (e.g. using a tablet.)

First name	NAMEERA
Last name	FAISAL AKHTAR
SID	3034244256
Collaborators	SARA IMAM

For staff use only:

Q10. Challenge Question (RL)	/16

Q10. [16 pts] Challenge Question (RL)

For this problem assume that the discount factor $\gamma = 1$. The environment in which the agent moves can be seen in Figure 1, which we will refer to as **MDP1**. The agent starts from the start state S. Double squares denote exit states from which the only action the agent can take is *exit*. By taking the *exit* action, the agent collects the reward listed in the double box and then moves to a terminal state where no further rewards can be collected. In all other states (the single boxes), the agent can move to any neighboring state, obtaining a zero reward. For example from state S the agent can go right by taking action →.

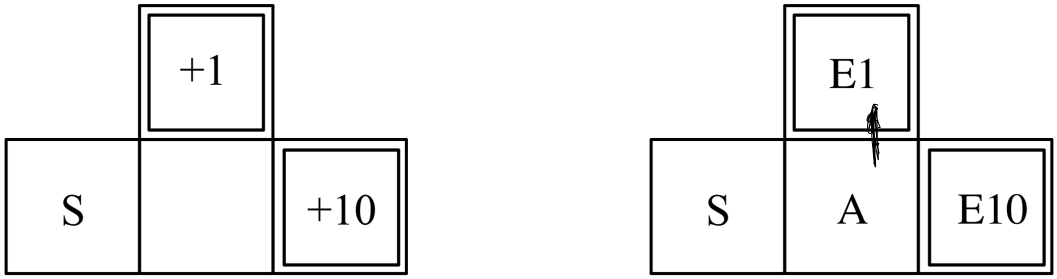


Figure 1: **MDP1**: (Left) Start state and rewards for exit actions. (Right) State names.

10.1) (2 pts) What are the optimal V-values for states A and S?

$V^*(A) =$ + 10

$V^*(S) =$ + 10

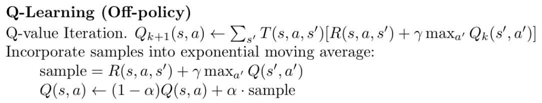
If we start at state A, the only way to earn rewards is by exiting from E_1 or E_{10}. However, the optimal policy is to exit at E_{10} (gaining a reward of 10).

Same logic for state S.

Computing optimal policies when we know the rewards and transitions in a MDP is straightforward. Now we assume that we do not have that information, and thus we would like to implement Q-learning to derive an optimal policy. When we run Q-learning, we will initialize the Q-values to zero. Assume the following sequence of transitions and associated rewards, where X denotes the terminal state:

s	a	s'	r
S	→	A	0
A	↑	E1	0
E1	exit	X	1
S	→	A	0
A	→	E10	0
E10	exit	X	10

- sample = $R(S, \rightarrow, A) + 1(0) = 0 + 0 = 0$
  $Q(S, \rightarrow) = (1-\alpha) . 0 + \alpha . 0 = 0$
- sample = $R(A, \uparrow, E_1) + 1(0) = 0 + 0 = 0$
  $Q(A, \uparrow) = (1-\alpha) . 0 + \alpha . 0 = 0$
- sample = $R(E_1, exit, X) + 1(0) = 1 + 0 = 1$
  $Q(E_1, exit) = (1-\alpha) . 0 + \alpha . 1 = \alpha$
- sample = $R(S, \rightarrow, A) + 1(0) = 0 + 0 = 0$
  $Q(S, \rightarrow) = (1-\alpha) 0 + \alpha . 0 = 0$



Q-Learning (Off-policy)

Q-value Iteration. $Q_{k+1}(s,a) \leftarrow \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma \max_{a'} Q_k(s',a')]$

Incorporate samples into exponential moving average:

sample $= R(s,a,s') + \gamma \max_{a'} Q(s',a')$

$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha \cdot$ sample

- sample = $R(A, \rightarrow, E_{10}) + 1(0) = 0 + 0 = 0$
  $Q(A, \rightarrow) = (1-\alpha) . 0 + \alpha . 0 = 0$
- sample = $R(E_{10}, exit, X) + 1(0) = 10 + 0 = 10$
  $Q(E_{10}, exit) = (1-\alpha) . 0 + \alpha . 10 = 10\alpha$

10.2) (2 pts) Which of the following Q-values are non-zero after running Q-learning on the transition-reward pairs above, assuming that we go through the sequence above only one time? Select all that apply.

A. $Q(S, \rightarrow)$

B. $Q(A, \uparrow)$

C. $Q(A, \rightarrow)$

[✓ D.] $Q(E1, exit)$

[✓ E.] $Q(E10, exit)$

10.3) (2 pts) Assume we use a learning rate α of 0.5. If we run Q-learning on the dataset above for an infinite number of iterations, then what are the Q-values upon convergence? If a Q-value does not converge, write *none* for that value.

$Q(S, \rightarrow) =$ 10

$Q(A, \leftarrow) =$ 0

$Q(A, \uparrow) =$ 1

We never go $(A, \leftarrow)$, so this never changes from its initial value of zero.

The other two actions converge to optimal exit values.

Now let's consider a modified MDP, called **MDP2** in which now state A (denoted with a spiral) is a special state in which the only action is to *escape*. The *escape* action will take the agent to a neighboring state, each with equal probability.

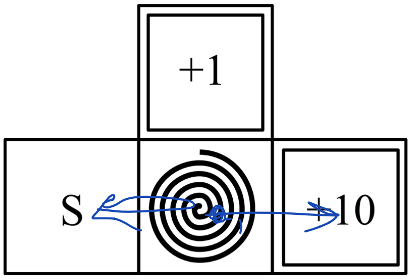


Figure 2: **MDP2**: States and rewards.

10.4) (2 pts) What are the optimal V-values in this new MDP for states S and A?

$V^*(S) =$ 11/2

$V^*(S) = V^*(A) + \text{cost of getting from S to A} = V^*(A) + 0 = V^*(A)$

$V^*(A) =$ 11/2

$$V^*(A) = \underbrace{\frac{1}{3} \cdot V^*(E_1)}_{\text{going up}} + \underbrace{\frac{1}{3} \cdot V^*(E_{10})}_{\text{going right}} + \underbrace{\frac{1}{3} V^*(S)}_{\text{going left}}$$

since $V^*(S) = V^*(A)$

$$V^*(A) = \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 10 + \frac{1}{3} V^*(A)$$

$$\frac{2}{3} V^*(A) = \frac{11}{3} \Rightarrow \boxed{V^*(A) = \frac{11}{2} = V^*(S)}$$

Now consider the following two datasets **S1** and **S2** accumulated from the new MDP. Remember that $E1$ denotes the square corresponding to an *exit* reward of +1 and $E10$ denotes the square corresponding to an *exit* reward of +10:

S1

s	a	s'	r
S	→	A	0
A	escape	E1	0
E1	exit	X	1
S	→	A	0
A	escape	E10	0
E10	exit	X	10

S2

s	a	s'	r
S	→	A	0
A	escape	E1	0
E1	exit	X	1
S	→	A	0
A	escape	E10	0
E10	exit	X	10
S	→	A	0
A	escape	E10	0
E10	exit	X	10

10.5) (2 pts) If we run Q-learning by iterating infinitely over the data sequence **S1** with an appropriately decreasing learning rate, what will the converged values of the following Q-values be? If a Q-value does not converge, write *none* for that value.

$Q^{S1}(S, \rightarrow) =$ 10

$Q^{S1}(A, escape) =$ 1

Taking sequence S1 from S going right, we exit from E_{10}.

Escaping from A takes you to E_1 in S1, so we exit from E_1

10.6) (2 pts) Under the same setup as in 10.5) but for **S2**, what are the values for the following two Q-values? If a Q-value does not converge, write *none* for that value.

$Q^{S2}(S, \rightarrow) =$ 10

$Q^{S2}(A, escape) =$ 10

Taking sequence S2 from S2 going right, we exit from E_{10}.

Escaping from A takes you to E_{10} in S2, so we exit from E_{10}.

10.7) (2 pts) Which of the following options is the true optimal Q-value $Q^*(S, \rightarrow)$ for **MDP2**?

[✓ A.] $Q^{S1}(S, \rightarrow)$

B. $Q^{S2}(S, \rightarrow)$

C. Neither

The first MDP has no randomness b/c A has no escape action. On the other hand, the second MDP has randomness since A has an escape action. As a result, MDP 1 is guaranteed to converge and MDP 2 is not guaranteed to converge b/c of randomness.

10.8) (2 pts) If we run Q-learning with a constant learning rate $\alpha = 1$ and we visit all state-actions pairs infinitely often, then for which of the two MDPs, if any, does Q-learning converge? Select exactly one answer.

A. **MDP1** only

B. **MDP2** only

[✓ C.] **MDP1** and **MDP2**

D. Neither of them

Both will converge eventually since we do enough trials.