CS 188
Fall 2022

Introduction to
Artificial Intelligence

Challenge Q10 HW5

- **Due:** Friday 10/7 at 11:59pm.
- **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	
Last name	
SID	
Collaborators	

For staff use only:

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 →.

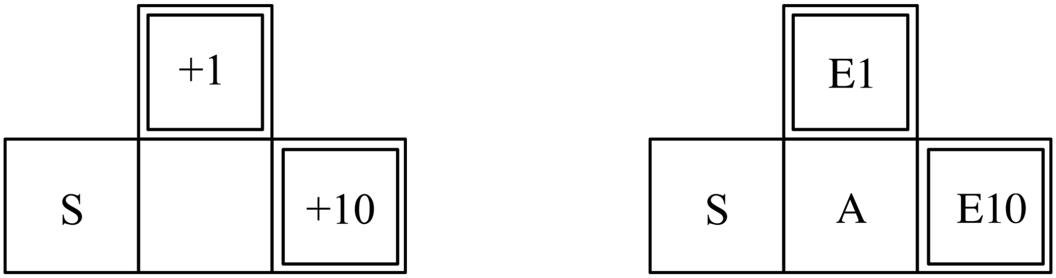

Figure 1: **MDP1**: (Left) Start state and rewards for exit actions. (Right) State names.

The Q-learning update equation is $Q'(s,a) = (1-\alpha)Q(s,a) + \alpha[R(s,a,s') + max_{a'}Q(s',a')]$. However, this problem can be solved without manually computing any Q-value updates.

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

$V^*(A) =$ 10

$V^*(S) =$ 10

In a deterministic undiscounted ($\gamma = 1$) MDP the optimal value is the maximum return from the state.

Computing 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

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. $Q(E10, exit)$

Q-values are only updated when a transition is experienced. $Q(E1, exit)$, $Q(E10, exit)$ are updated to the reward earned, but the other states were updated when all the Qs were still zero.

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.

The MDP is undiscounted and deterministic, so Q-learning converges even though the learning rate is constant. With infinite visits the Q-values will converge to the true values.

$Q(S, \rightarrow)$ =10, $Q(S, \rightarrow)$ is the only state-action for S, so it converges to the optimal value $V^*(S)$.

$Q(A, \leftarrow)$ =0, The episode A, ← is never experienced so it is unchanged after initialization.

$Q(A, \uparrow)$ =1, The only return possible after A, ↑ is 1.

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.

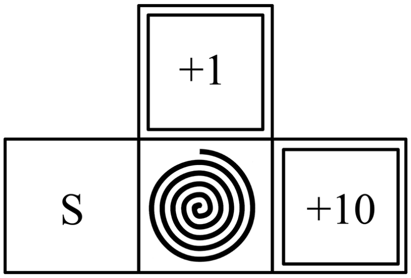

Figure 2: **MDP2**: States and rewards.

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

The optimal value is the mean of the end returns 1 and 10 because the exit states have equal probability. The value of S is the same as A since the discount $\gamma = 1$ and the transition S, →, A is deterministic. The transition A, escape, S has no impact on the value because the MDP is undiscounted / infinite horizon.

$V^*(S) =$ 5.5

$V^*(A) =$ 5.5

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.

$Q^{S1}(S, \rightarrow)$ =5.5

$Q^{S1}(A, escape)$ =5.5

The conditions for convergence are satisfied and the Q-values converge to the expected return. The expectation of returns is $1/2 \times 1 + 1/2 \times 10$.

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.

$Q^{S2}(S, \to) = 7$

$Q^{S2}(A, escape) = 7$

The expectation of returns is $1/3 \times 1 + 2/3 \times 10$ because two out of three exits in the sequence have reward 10.

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

A. [✓ $Q^{S1}(S, \to)$]

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

C. Neither

The sequence $S1$ has the same distribution of returns as the true distribution, even though all of the possible transitions are not experienced.

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

For learning rate $\alpha = 1$ the Q-learning update sets $Q(s, a)$ to the sample $[R(s, a, s') + \max_{a'} Q(s', a')]$ with no regard for the previous value of $Q(s, a)$.

In deterministic MDPs (like **MDP1**), even with constant learning rate $\alpha = 1$, Q-learning converges. In fact, this learning rate is optimal for deterministic MDPs in the sense that it converges the fastest.

In stochastic MDPs (like **MDP2**), with constant learning rate $\alpha = 1$, the $Q(s, a)$s are always equal to the most recent sample for the state-action (s, a). The $Q(s, a)$s will cycle among the possible samples and never converge.