HW 3

Chih-Hsiang Wang

933271081

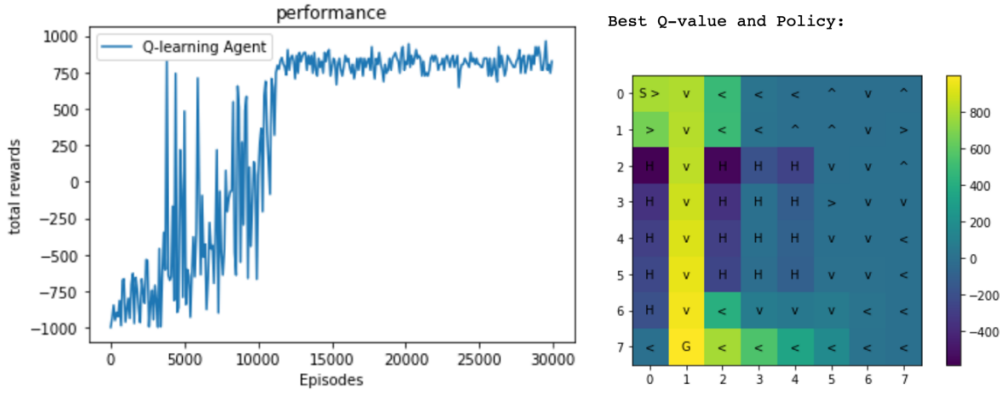
Chieh-Chin Tao

933320456

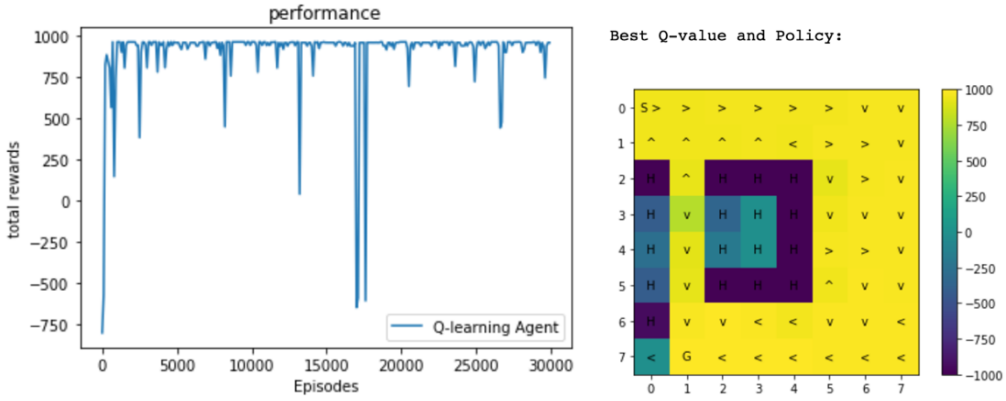
1. Provide the learning curves for the above experiments. Clearly label the curves by the parameters used.

- Dangerous Hallway Map:
    i. Q-learning Agent:

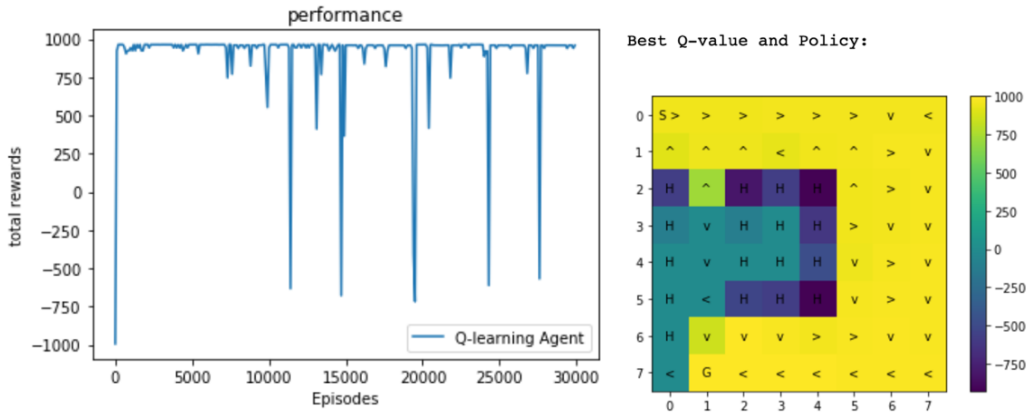
Epsilon = 0.3, Learning rate = 0.001, Learning time: 56.4612



Epsilon = 0.3, Learning rate = 0.1, Learning time: 27.8832

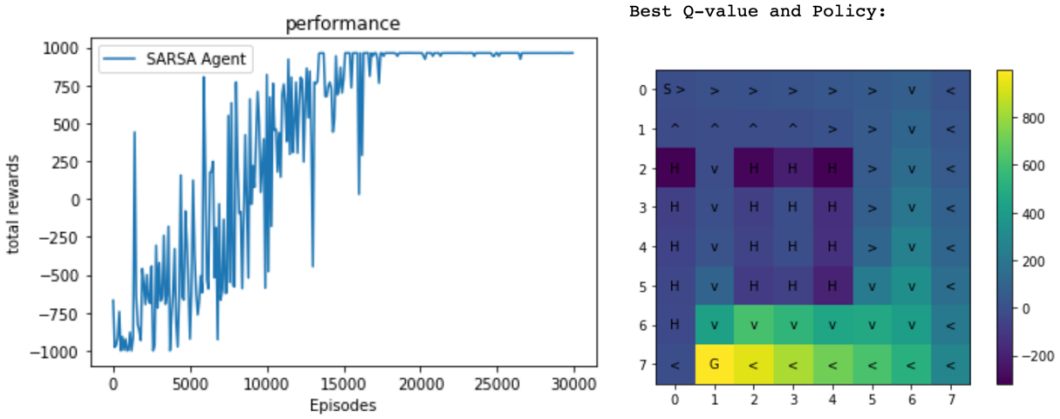


Epsilon = 0.05, Learning rate = 0.1, Learning time: 25.9365

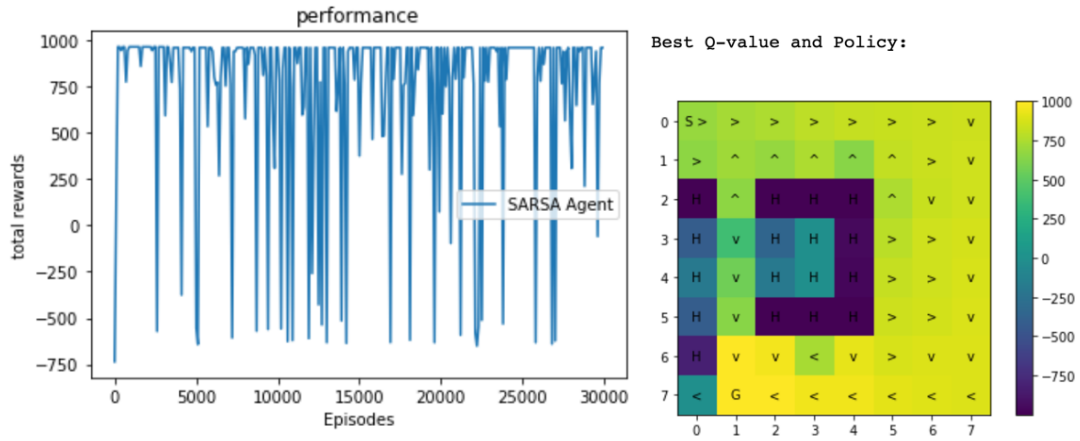


ii. SARSA Agent

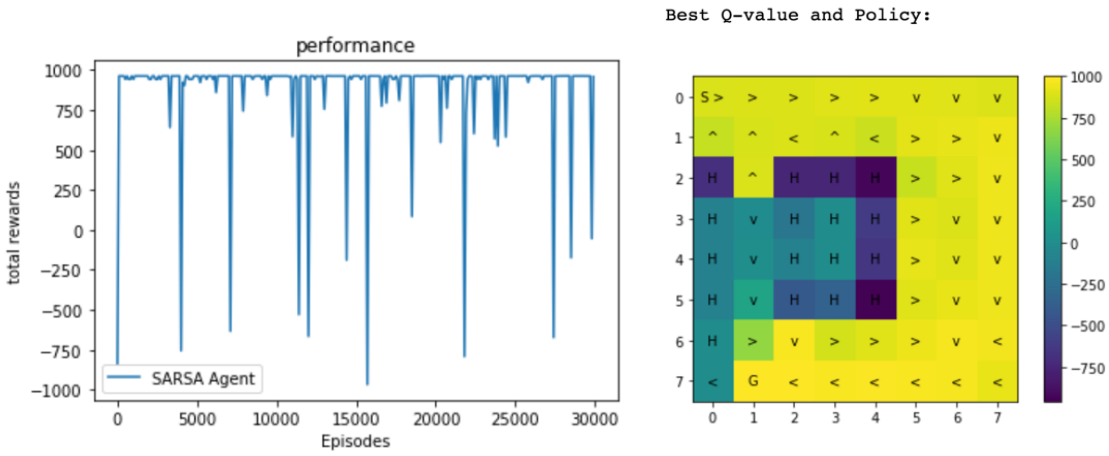
Epsilon = 0.3, Learning rate = 0.001, Learning time: 68.3363



Epsilon = 0.3, Learning rate = 0.1, Learning time: 58.717

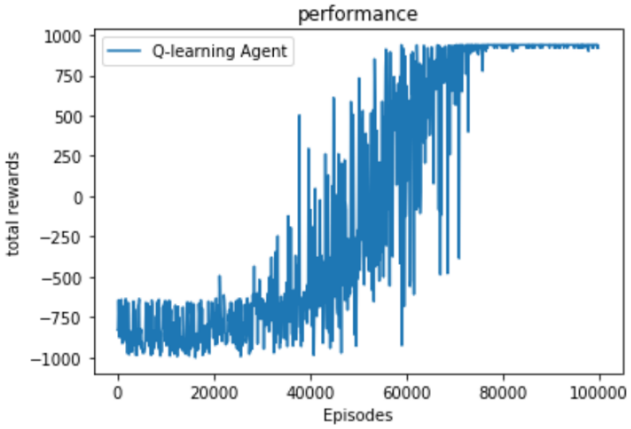


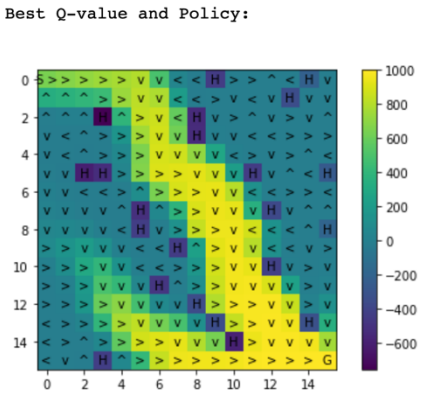
Epsilon = 0.05, Learning rate = 0.1, Learning time: 27.8606



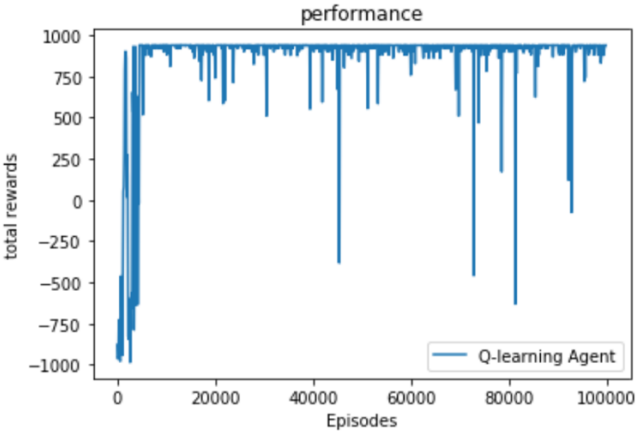
- Size 16x16 Map:
  i. Q-learning Agent:

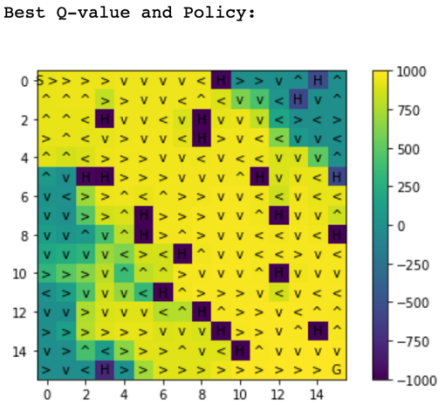
Epsilon = 0.3, Learning rate = 0.001, Learning time: 438.3829



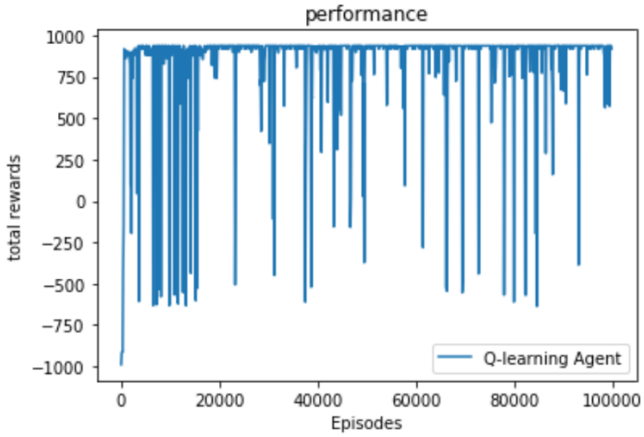


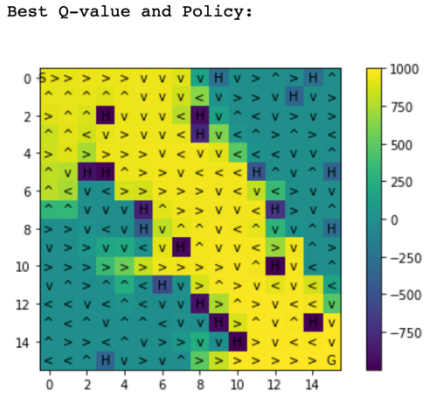
Epsilon = 0.3, Learning rate = 0.1, Learning time: 127.3173





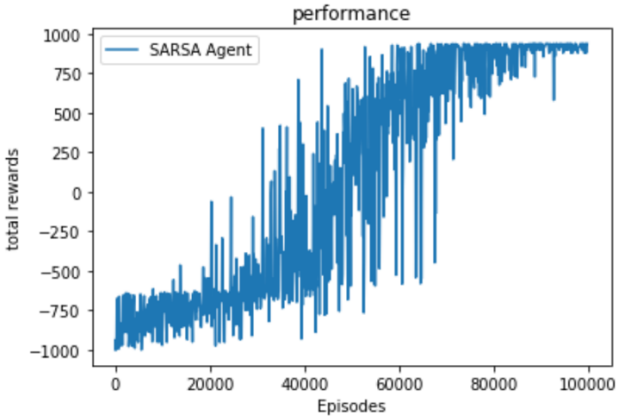
Epsilon = 0.05, Learning rate = 0.1, Learning time: 124.1329

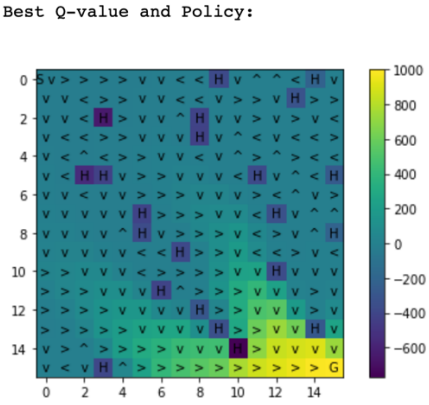




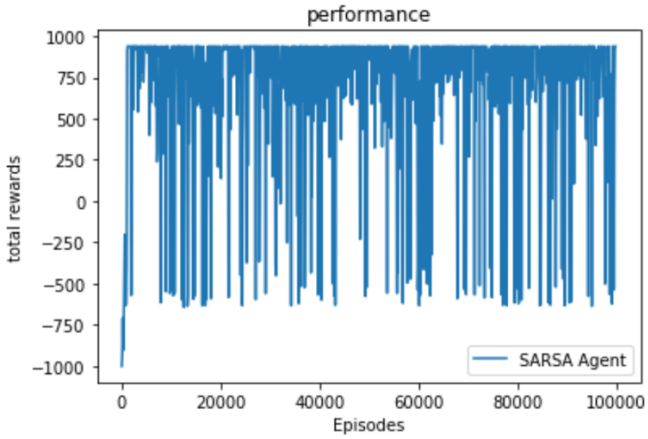
ii. SARSA Agent

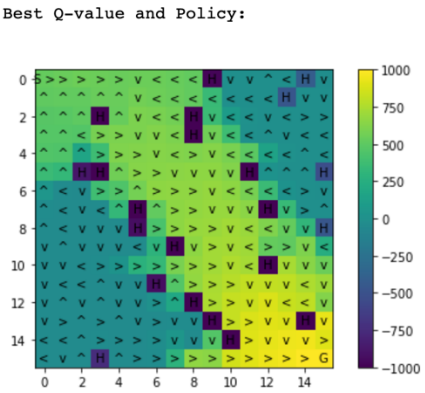
Epsilon = 0.3, Learning rate = 0.001, Learning time: 471.5013



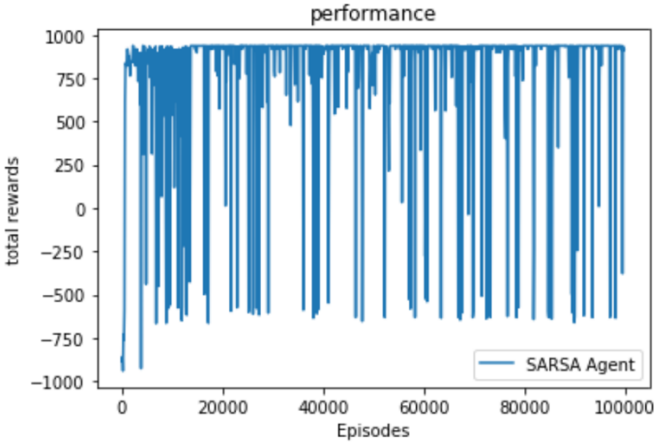


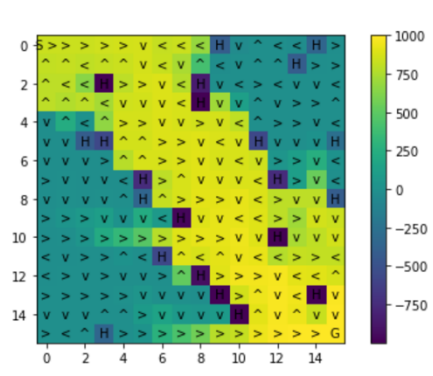
Epsilon = 0.3, Learning rate = 0.1, Learning time: 266.9076



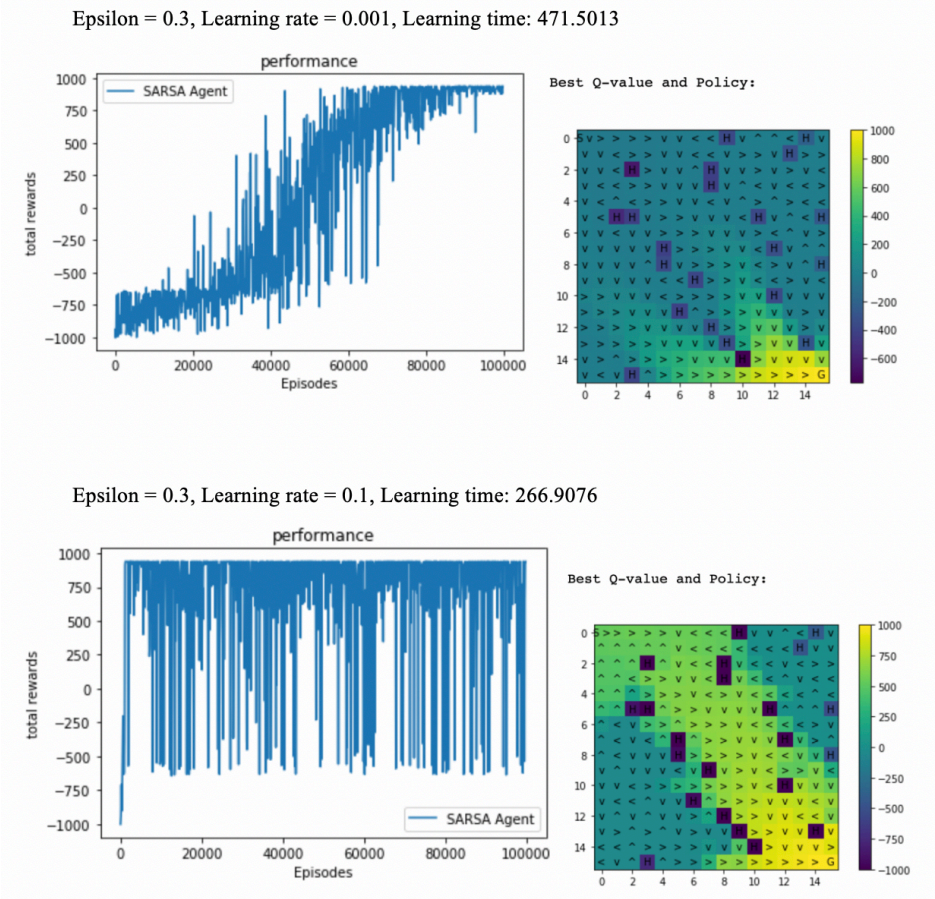


Epsilon = 0.05, Learning rate = 0.1, Learning time: 171.0059





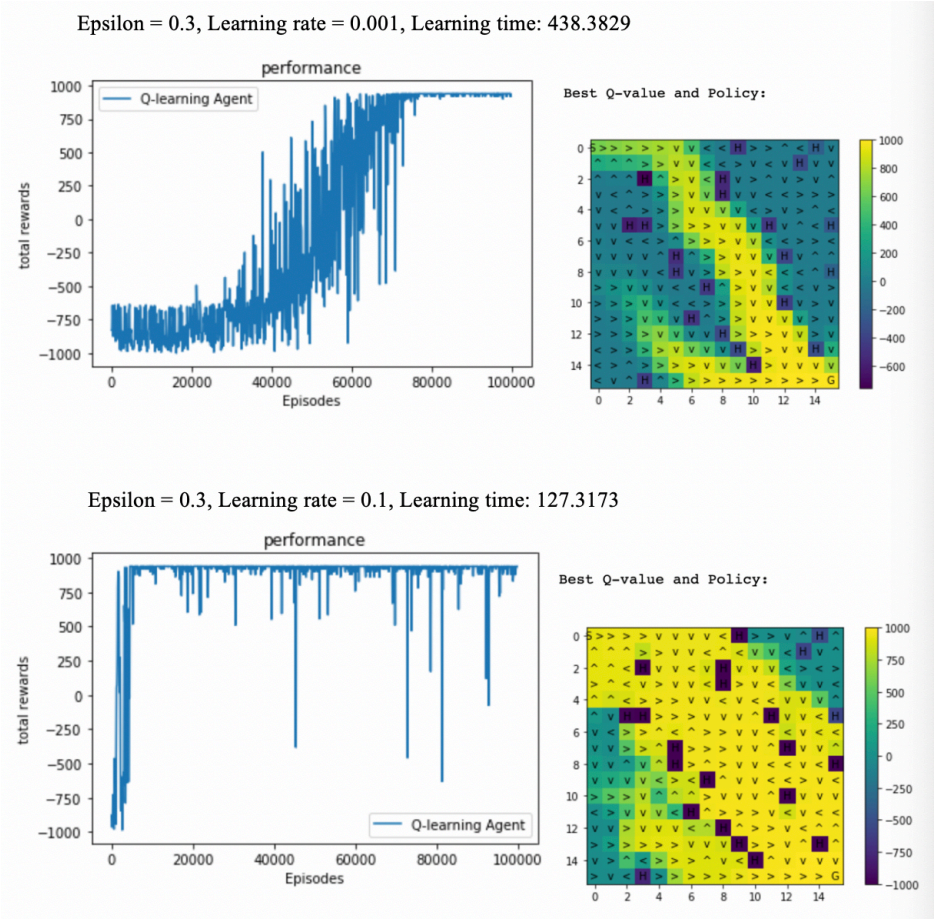
2. Did you observe differences for SARSA when using the two different learning rates? If there were significant differences, what were they and how can you explain them?



When the learning rate equals to 0.001, the total reward will become larger gradually as the episode increases. Besides, the fluctuation between each episode is small compared to higher learning rate. On the contrary, while the learning rate equals to 0.1, the total reward will reach the highest value in a short period and the fluctuation is large. For the Q-value and policy maps, the one with higher learning rate results in higher Q-value in each grid.

According to the formula, $Q(s,a) \leftarrow Q(s,a) + \alpha(r+\beta Q(s',a') - Q(s,a))$, learning rate α is the step size which relates to scaling the difference between target and estimated Q-value. Larger α means the update of old estimate each episode will approach target value faster.

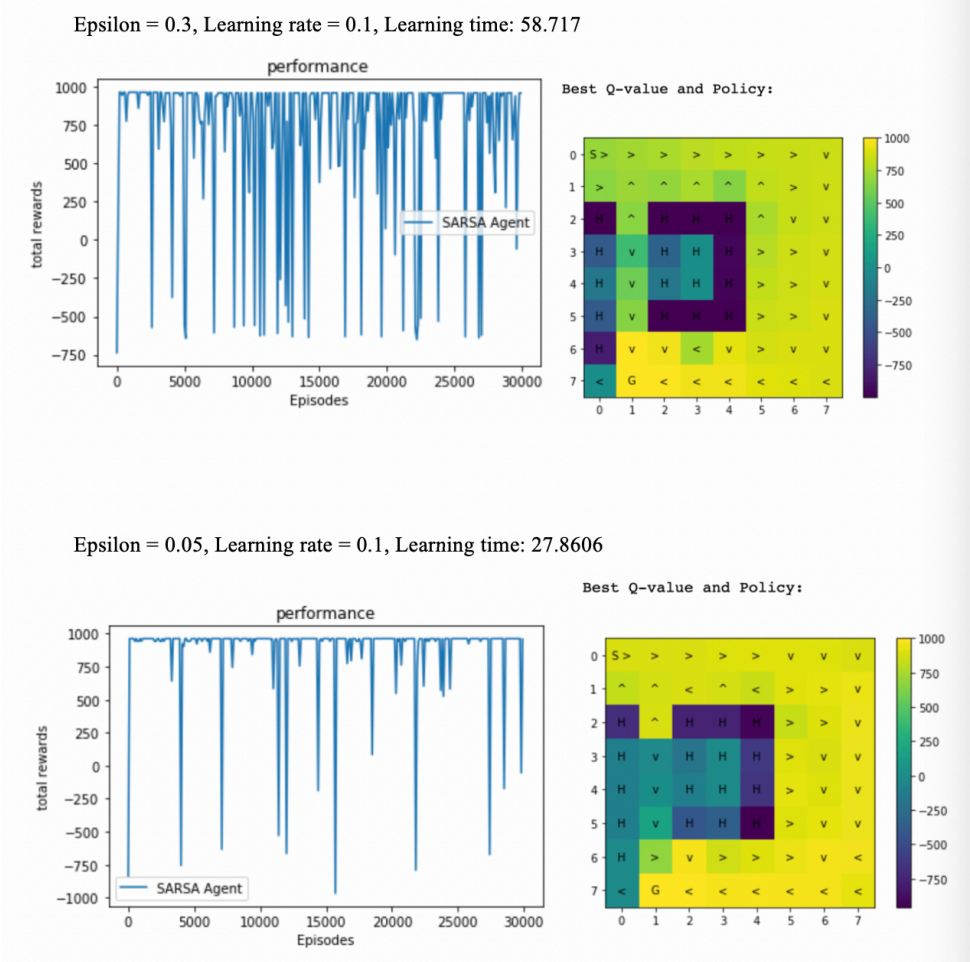
3. Repeat (2) for Q-Learning.



When the learning rate equals to 0.001, the total reward will become larger gradually as the episode increases. On the contrary, while the learning rate equals to 0.1, the total reward will reach the highest value in a short period. Due to the greedy policy in Q-learning, the fluctuation becomes smaller because it already found the largest reward and tends to comply with large Q-value. For the Q-value and policy maps, the one with higher learning rate results in higher Q-value in each grid.

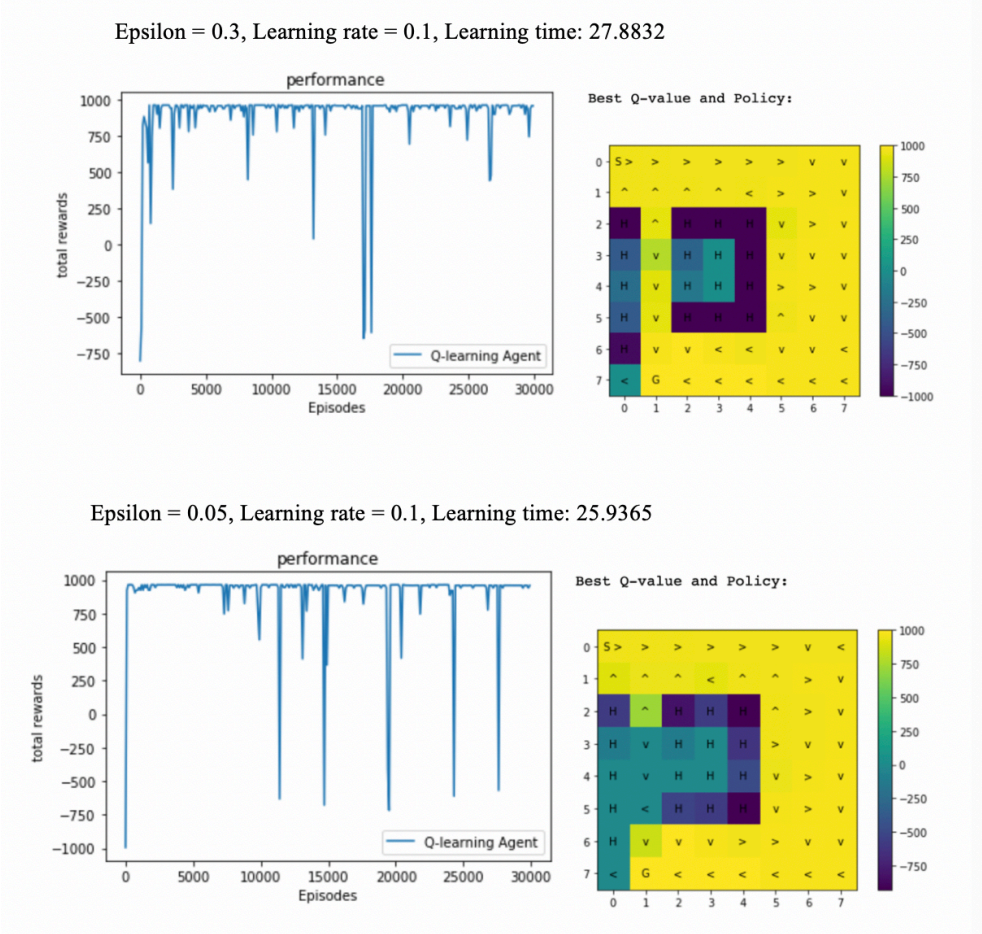
According to the formula, $Q(s,a) \leftarrow Q(s,a) + \alpha(r+\beta Q(s',a') - Q(s,a))$, learning rate α is the step size which relates to scaling the difference between target and estimated Q-value. Larger α means the update of old estimate each episode will approach target value faster.

4. Did you observe differences for SARSA when using different values of ϵ? If there were significant differences, what were they and how do you explain them?



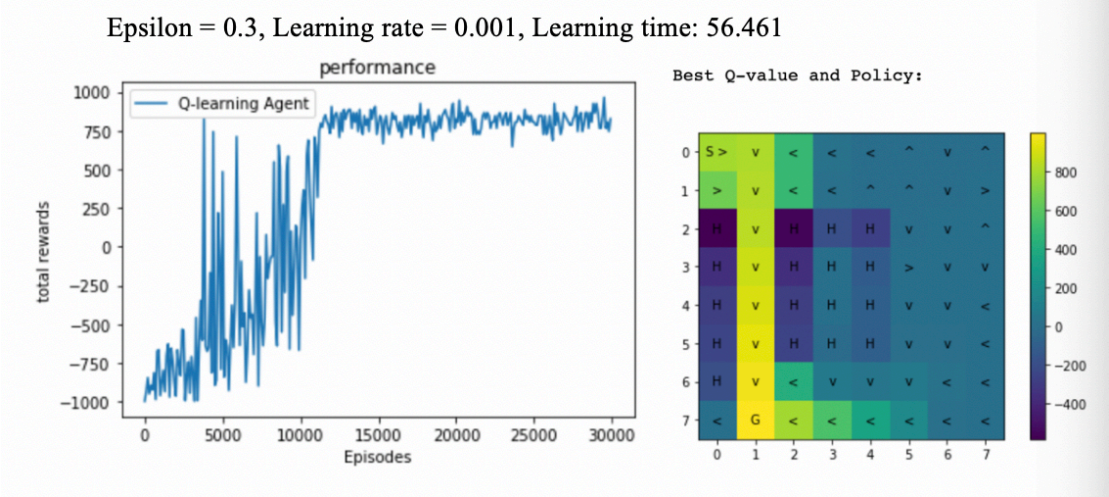
When the ϵ is small, the action selection will depend more on greedy policy, meaning exploit more. In contrary, larger ϵ means explore more that action is randomly assigned. In the maps, the probability of choosing the dangerous path for smaller ϵ is lower because the policy tends to take the action to the grid with high Q-value.

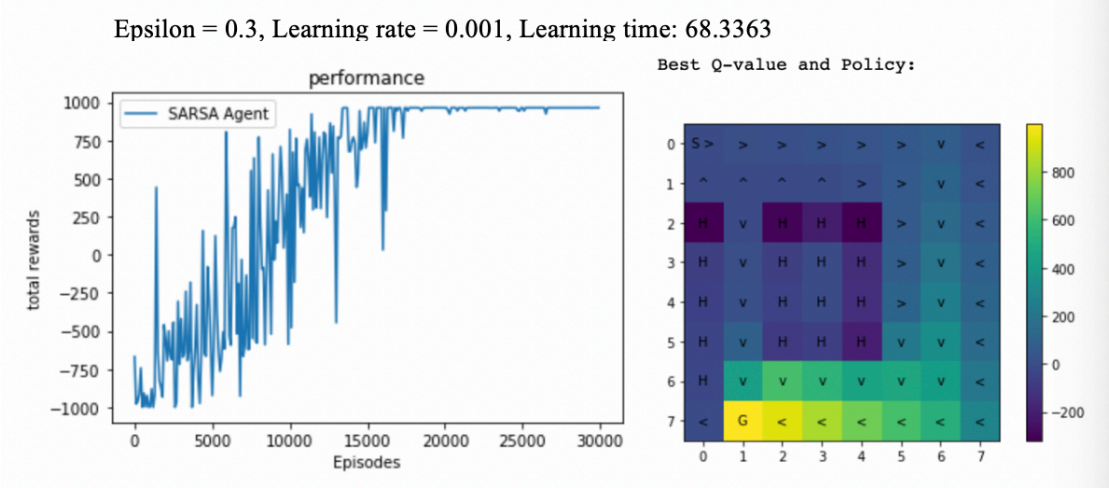
5. Repeat (4) for Q-Learning.



When the ϵ is small, the action selection will depend more on greedy policy, meaning exploit more. In contrary, larger ϵ means explore more that action is randomly assigned. In addition, we also discover that smaller ϵ for Q-learning makes the policy avoid dangerous area more that the Q-value in this area is smaller. The result corresponds to the fact that we exploit the map more frequently.

6. For the map "Dangerous Hallway" did you observe differences in the policies learned by SARSA and Q-Learning for the two values of epsilon (there should be differences between Q-learning and SARSA for at least one value)? If you observed a difference, give your best explanation for why Q-learning and SARSA found different solutions.





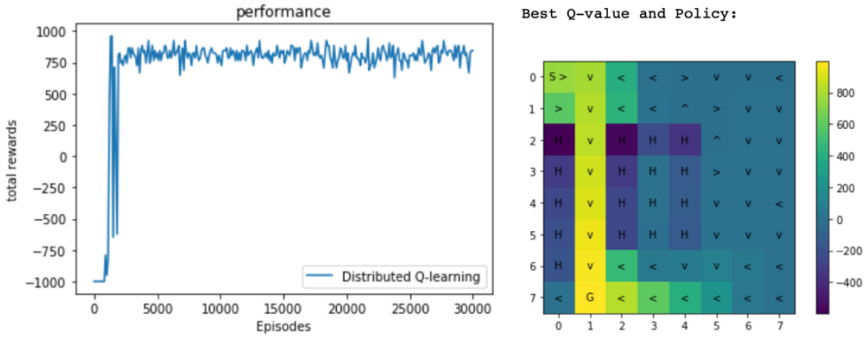
Policy following Q-learning tries to reach the goal as fast as possible and only consider about large Q value. It leads to the outcome to take the nearest but most dangerous path. It is off-policy without knowing the next state and action. In comparison, SARSA is on-policy that the next state and action will become the state and action it takes, meaning it knows the next state and action while in the current state. As a result, Q-learning tends to be greedy and bold that it cares less about resulting in mistakes. SARSA, however, is a conservative algorithm trying to avoid failure while learning. That's the reason why SARSA avoids taking the nearest path to the goal in the map "Dangerous Hallway".

7. Show the value functions learned by the distributed methods for the best policies learned with ϵ equal to 0.3 and compare to those of the single-core method. Run the algorithm for the recommended number of episodes for each of the maps. Did the approaches produce similar results?

- Dangerous Hall Map:

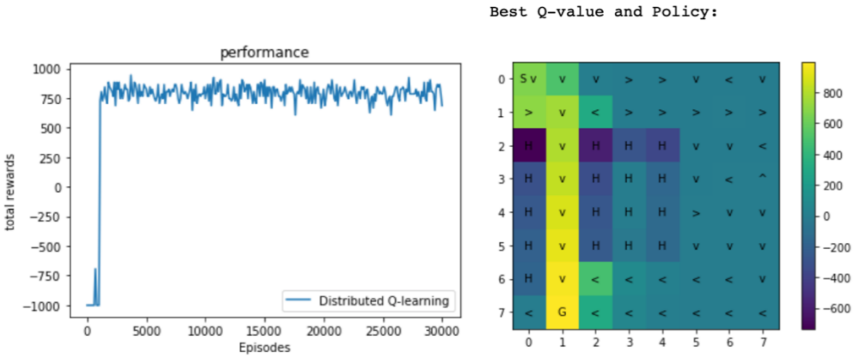
Epsilon = 0.3, Learning rate = 0.001

Collector workers = 2, Evaluator workers = 4, Learning time: 22.5676



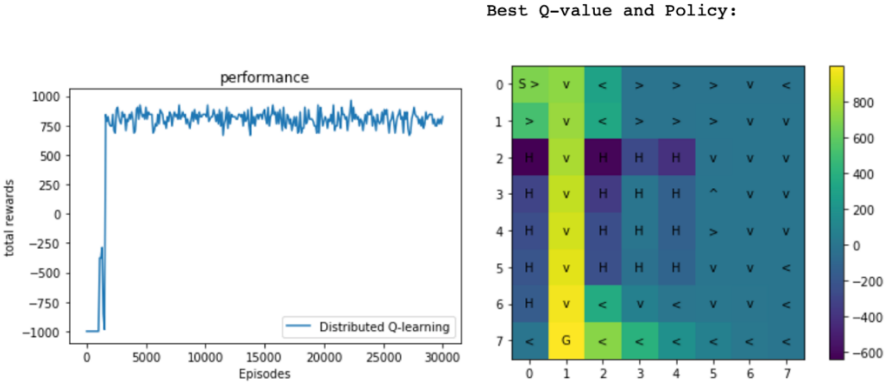
Epsilon = 0.3, Learning rate = 0.001

Collector workers = 4, Evaluator workers = 4, Learning time: 7.87663



Epsilon = 0.3, Learning rate = 0.001

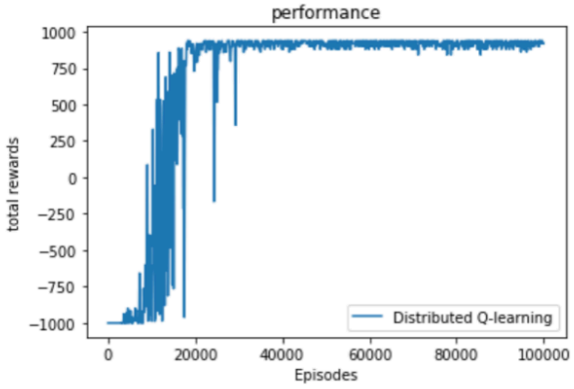
Collector workers = 8, Evaluator workers = 4, Learning time: 7.708254

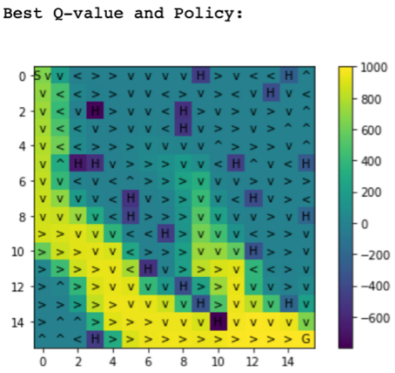


- Size 16x16 Map:

Epsilon = 0.3, Learning rate = 0.001

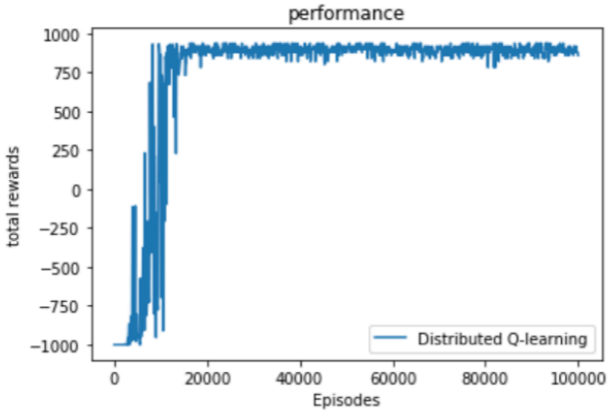
Collector workers = 2, Evaluator workers = 4, Learning time: 133.2313

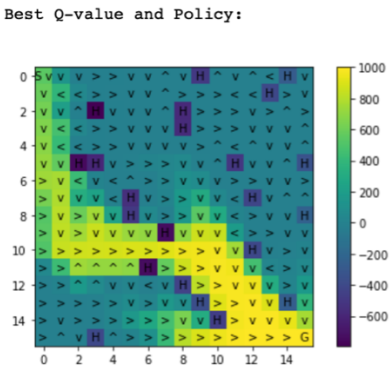




Epsilon = 0.3, Learning rate = 0.001

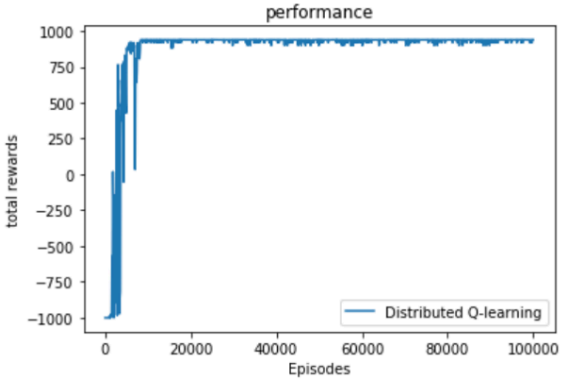
Collector workers = 4, Evaluator workers = 4, Learning time: 45.09626

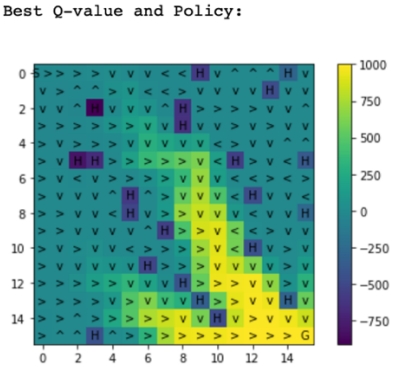




Epsilon = 0.3, Learning rate = 0.001

Collector workers = 8, Evaluator workers = 4, Learning time: 32.4094





The performances of total reward are different. The distributed one approaches the highest value much faster. The Q-value and policy maps are similar, but there is a little difference in the path choices in the 16x16 map for different collectors in the distributed method.

8. Provide and compare the timing results for the single-core and distributed experiments, including the time to do the evaluations during learning. Describe the trends you observe as the number of workers increases.

- Dangerous Hall Map (Epsilon = 0.3, Learning rate = 0.001):

Method	Distributed experiment			Single-core	
	Collector workers = 2	Collector workers = 4	Collector workers = 8	Q-learning	SARSA
Learning Time	22.5676	7.87663	7.708254	56.4612	68.3363

- Size 16x16 Map (Epsilon = 0.3, Learning rate = 0.001):

Method	Distributed experiment			Single-core	
	Collector workers = 2	Collector workers = 4	Collector workers = 8	Q-learning	SARSA
Learning Time	133.2313	45.09626	32.4094	438.3829	471.5013

When the number of collector workers increases, the learning time becomes faster.

9. Provide and compare the timing results for the single-core and distributed experiments with the evaluation procedure turned off. That is, here you only want to provide timing results for the learning process without any interleaved evaluation. Compare the results with (8).

- Dangerous Hall Map (Epsilon = 0.3, Learning rate = 0.001):

Method	Distributed experiment			Single-core	
	Collector workers = 2	Collector workers = 4	Collector workers = 8	Q-learning	SARSA
Learning Time	7.11312	3.6112	2.74975	7.941449	11.200521

- Size 16x16 Map (Epsilon = 0.3, Learning rate = 0.001):

Method	Distributed experiment			Single-core	
	Collector workers = 2	Collector workers = 4	Collector workers = 8	Q-learning	SARSA
Learning Time	57.67671	30.014223	20.29555	77.9968	77.7362

Without the evaluation procedure, the learning time becomes much faster than the procedure with evaluation.