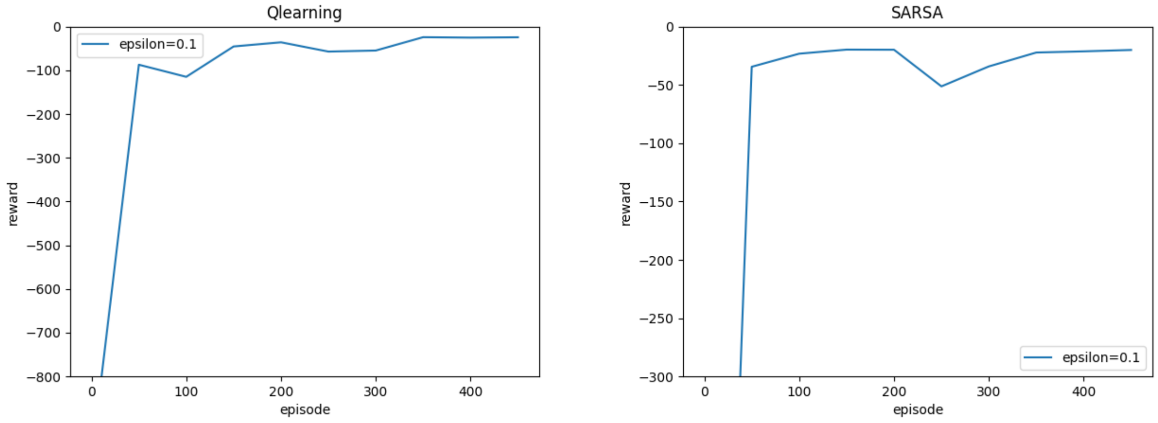
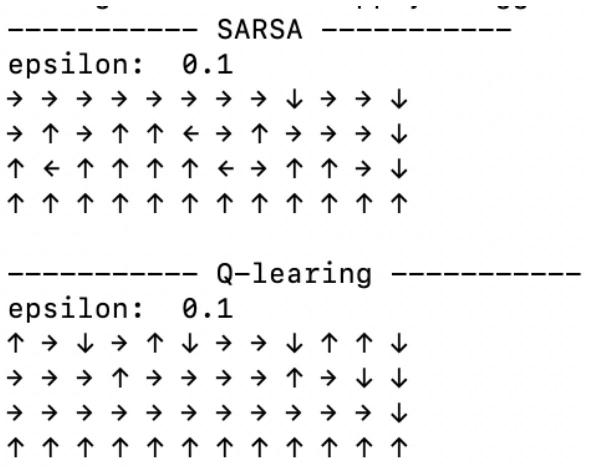
CS471 Project

Yiting Gan gna29@purdue.edu

December 11, 2021

Problem 1

1. 

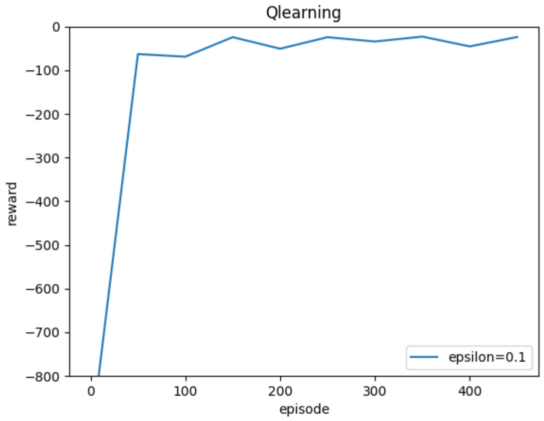
2. 

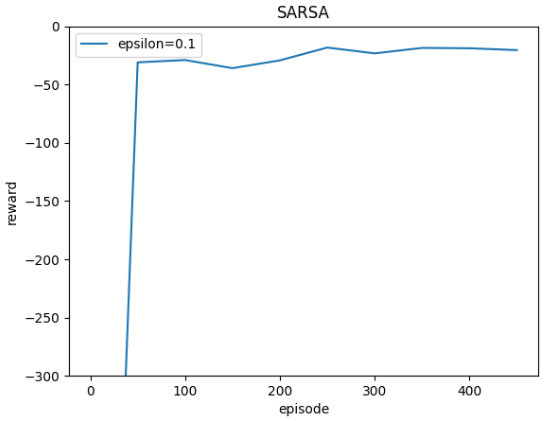
```
----------- SARSA -----------
epsilon:  0.1
→ → → → → → → → ↓ → → ↓
→ ↑ → ↑ ↑ ← → ↑ → → → ↓
↑ ← ↑ ↑ ↑ ↑ ← → ↑ ↑ → ↓
↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

----------- Q-learing -----------
epsilon:  0.1
↑ → ↓ → ↑ ↓ → → ↓ ↑ ↑ ↓
→ → → ↑ → → → → ↑ → ↓ ↓
→ → → → → → → → → → → ↓
↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑
```

3. The policy are different. While Q-Learning agent tends to choose shorter but dangerous paths, SARSA agent tends to take safer but longer paths. Considering average reward at the end of the training, since SARSA cares more about safety, it has a greater average than Q-Learning.

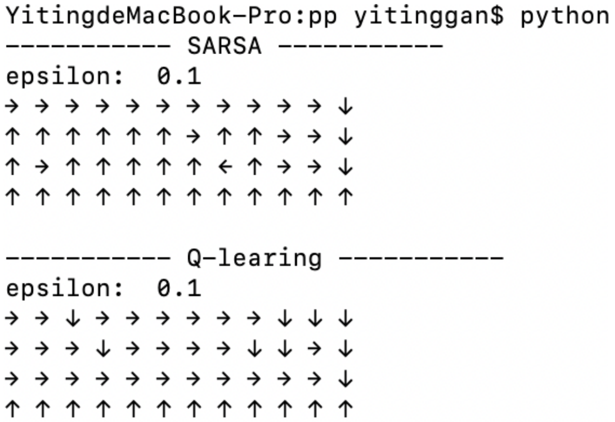
Problem 2: Bonus

Problem 3





1.



```
YitingdeMacBook-Pro:pp yitinggan$ python
----------- SARSA -----------
epsilon:  0.1
→ → → → → → → → → → → ↓
↑ ↑ ↑ ↑ ↑ ↑ → ↑ ↑ → → ↓
↑ → ↑ ↑ ↑ ↑ ↑ ← ↑ → → ↓
↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

----------- Q-learing -----------
epsilon:  0.1
→ → ↓ → → → → → → ↓ ↓ ↓
→ → → ↓ → → → → ↓ ↓ → ↓
→ → → → → → → → → → → ↓
↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑
```

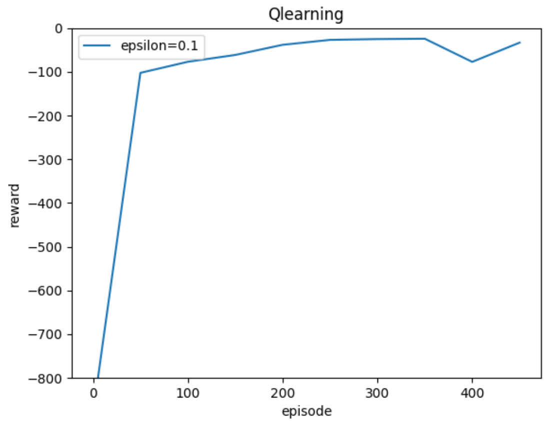
2.

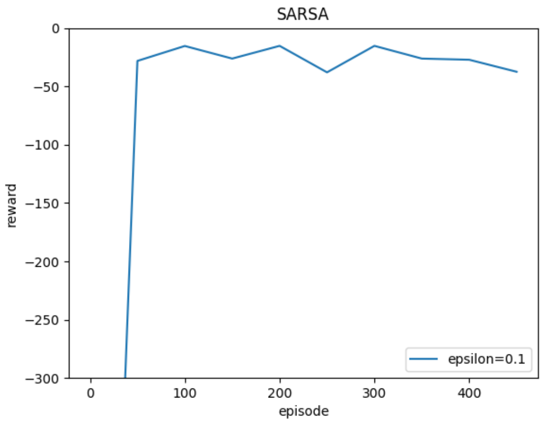
3. The policy does not change with varying ϵ and fixed $\epsilon = 0.1$. The reward increases as ϵ increases. This happens because Q-learning is off-policy and does not learn the action-value function from the policy.

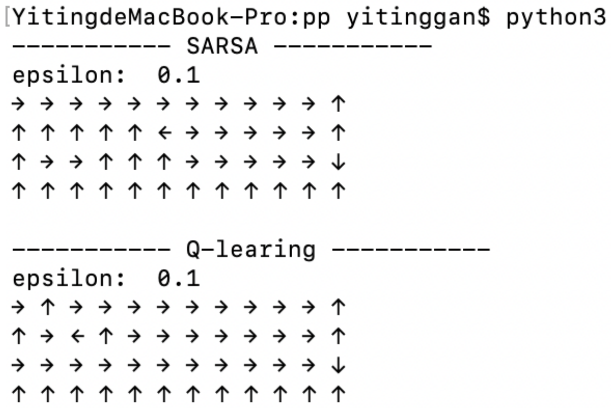
4. The policy varied. With $\epsilon = 0.1$, SARSA will choose path far from the cliff. Since ϵ increases means that the probability to fall off the cliff increases, that SARSA, on-policy, tends to avoid risk results a longer path.

5. The policy is different. Since on-policy and off-policy have different properties, SARSA will try to stay away from the cliff as ϵ changes from 0 to 0.1, while Q-learinng will still take the risk to get the shortest path.

Problem 4

1. 



2. 

```
[YitingdeMacBook-Pro:pp yitinggan$ python3
 ----------- SARSA -----------
 epsilon:  0.1
 → → → → → → → → → → → ↑
 ↑ ↑ ↑ ↑ ↑ ← → → → → → ↑
 ↑ → → ↑ ↑ ↑ → → → → → ↓
 ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

 ----------- Q-learing -----------
 epsilon:  0.1
 → ↑ → → → → → → → → → ↑
 ↑ → ← ↑ → → → → → → → ↑
 → → → → → → → → → → → ↓
 ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑
```

3. The policy are different. SARSA tends to find the upper destination. It will try to avoid going down, since the cliff is at the bottom of the grid. Yet, Q-Learning tends to have a evenly distributed policy with both destinations.