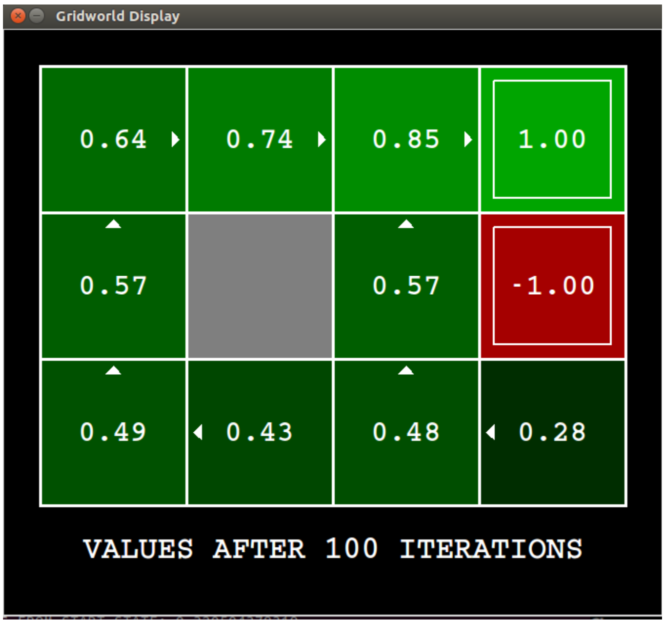
ANTALYA BİLİM
ÜNİVERSİTESİ

Assignment 3: **Reinforcement Learning**

Ahmed Amine
TALEB BAHMED
130201120

Dr. Hilal KAZAN

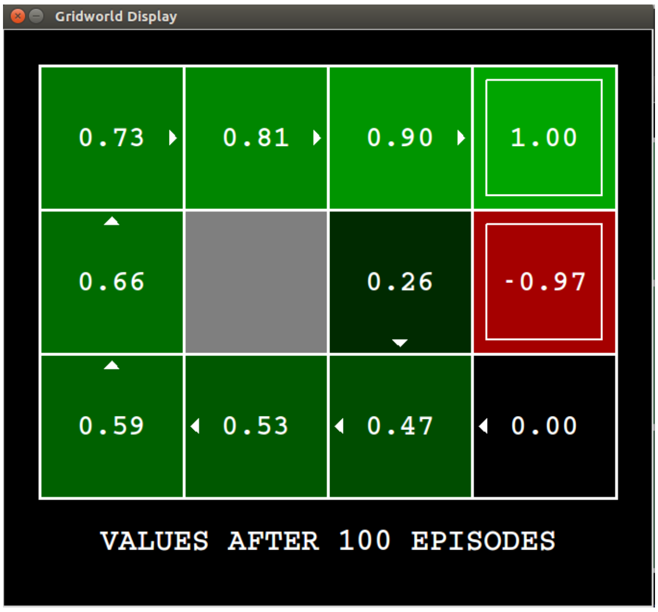
Value Iteration:



0.64 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.57		▴ 0.57	-1.00
▴ 0.49	◂ 0.43	▴ 0.48	◂ 0.28

VALUES AFTER 100 ITERATIONS

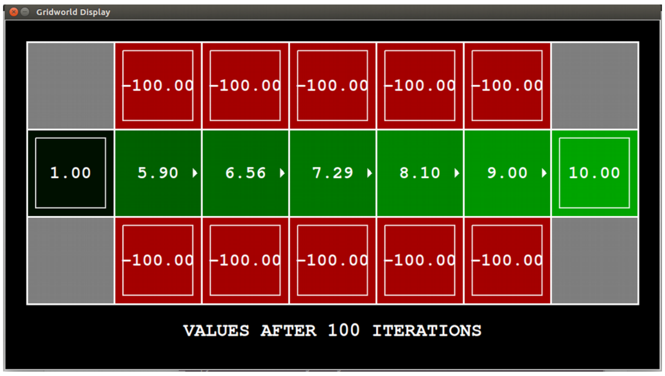
1-
The learned values are smaller then those learned by value iteration, since the agent does not have the transition probabilities and the reward for every transition in Q_learning. To get to the optimal value we make the learning rate small enough.



0.73 ▸	0.81 ▸	0.90 ▸	1.00
▴ 0.66		0.26 ▾	-0.97
▴ 0.59	◂ 0.53	◂ 0.47	◂ 0.00

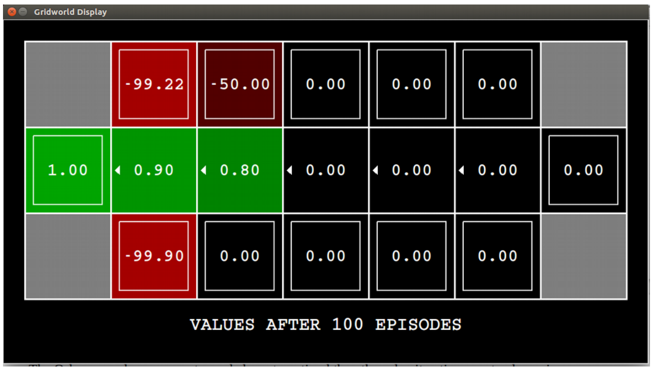
VALUES AFTER 100 EPISODES

2- Train your Q-learner on the BridgeGrid with no noise (-n 0.0) for 100 episodes. How do the learned q-values compare to those of the value iteration agent? Why will your agent usually not learn the optimal policy?
The Q-learner values are smaller and insufficient to reach the optimal than the value iteration agent values, we need more episodes



	-100.00	-100.00	-100.00	-100.00	-100.00	
1.00	5.90 ▸	6.56 ▸	7.29 ▸	8.10 ▸	9.00 ▸	10.00
	-100.00	-100.00	-100.00	-100.00	-100.00	

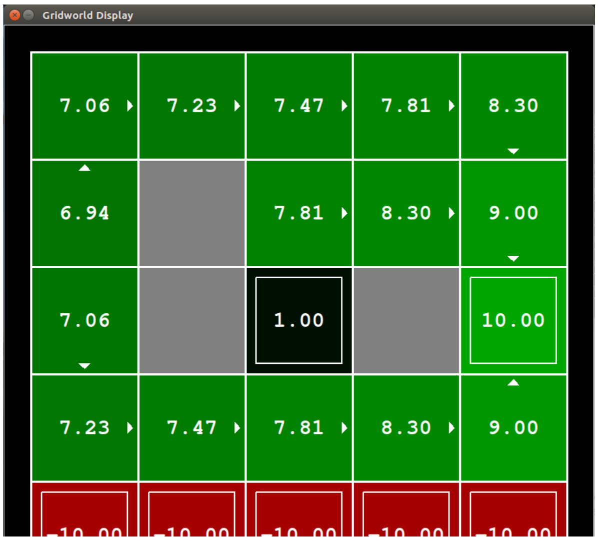
VALUES AFTER 100 ITERATIONS



	-99.22	-50.00	0.00	0.00	0.00	
1.00	◂ 0.90	◂ 0.80	◂ 0.00	◂ 0.00	◂ 0.00	0.00
	-99.90	0.00	0.00	0.00	0.00	

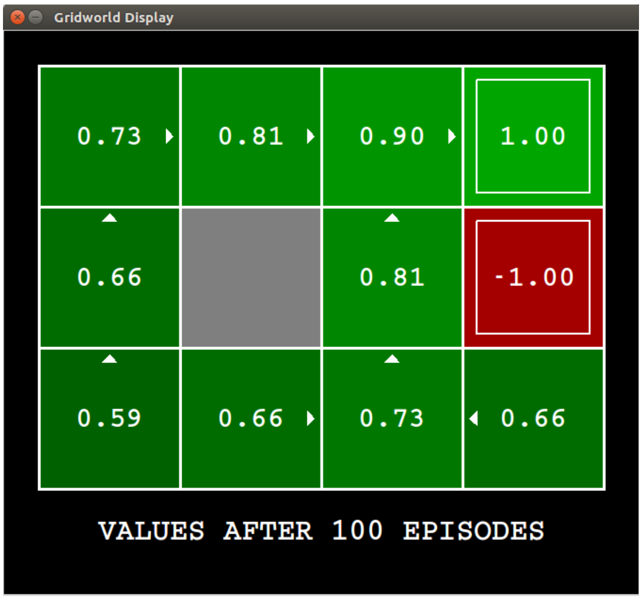
VALUES AFTER 100 EPISODES

3-
Train your Q-learner on the CliffGrid for 100 episodes. Compare the value it learns for the start state with the average returns from the training episodes (printed out automatically). Why are they so different?



7.06 ▸	7.23 ▸	7.47 ▸	7.81 ▸	8.30 ▾
▴ 6.94		7.81 ▸	8.30 ▸	9.00 ▾
7.06 ▾		1.00		10.00
7.23 ▸	7.47 ▸	7.81 ▸	8.30 ▸	▴ 9.00
-10.00	-10.00	-10.00	-10.00	-10.00

6-
When epsilon is -e 0.9 it learns faster than e 0.1, it acts on current policy while, when epsilon is -e 0.1 learns too slow since it acts randomly.



0.73 ▸	0.81 ▸	0.90 ▸	1.00
▴ 0.66		▴ 0.81	-1.00
▴ 0.59	0.66 ▸	▴ 0.73	◂ 0.66

VALUES AFTER 100 EPISODES