Reinforcement Learning Homework 4

Deadline 12th November 2024
Students Sahiti Chebolu, Surabhi S Nath, Xin Sui

1 Q-Learning and SARSA

(a)

Q-learning is considered an off-policy control method because it enables learning about policy π from experience sampled with μ (as opposed to just from π as in SARSA). It uses the best local next action A' for value backup instead of the $Q(S', A')$ in SARSA:

$$Q(S, A) \leftarrow Q(S, A) + \alpha(R + \gamma \max_{a'} Q(S', a') - Q(S, A)) \tag{1}$$

(b)

If Q-learning action selection is greedy, this doesn't make it equivalent to SARSA. In SARSA, action selection (and the policy being improved) is ϵ-greedy.
There's still the on vs. off policy distinction in terms of value backup/learning. They general do not make the same action selections and weight updates because they learn from different Q values.

(c)

(a) Action going right should always be taken at state A because it results in on average higher rewards (0 as opposed to -0.1).
(b) We'd expect Q-learning to yield an optimal policy of going left at state A instead. This is because Q-learning uses the best next Q for value backup and is therefore 'biased' towards the best possible final reward (> 0) instead of the average final reward (0).

2 Hands-on in Gridworld and Q-Learning

(a)

See Fig 1 below. Q learning predicted higher expected returns in general because it uses the best local next action A' for value backup, therefore it is (at least largely) unaffected by noise. However, there's an exception with the state in the upper left corner - Q learning produced smaller expected return there. This is probably a result of the ϵ-greedy policy in Q learning as compared to the greedy policy in value iteration.
To make the values closer to optimal values, we need to decay ϵ towards 0 and train for more episode. The optimal values should also be produced in an environment with 0 noise instead of the default 0.2.

(b)

See Fig 2 below. Here, the learned q-values are very small compared to results from value iteration. This is a result of the ϵ-greedy policy in Q learning in addition to the highly negative reward associated with falling off the bridge. Even for 10000 episodes (Fig 3), the values obtained for value-iteration and Q-learning are wildly different. This is because Q-learning agent's value estimates are swayed by the large, negative rewards obtained randomly due to the ϵ-greedy policy.

(c)

The value estimate for the start state from 300 Q-learning iterations is 2.74 (Fig 4). The average returns from these episodes is -26.88. For comparison, the value estimate from value iteration is -4.14 (Fig 5). The average returns is -2.95. The discrepancy in the Q-learning case is because the action selection is ϵ-greedy, leading to negative reward of -100 sometimes due to random actions (instead of greedy action selection)

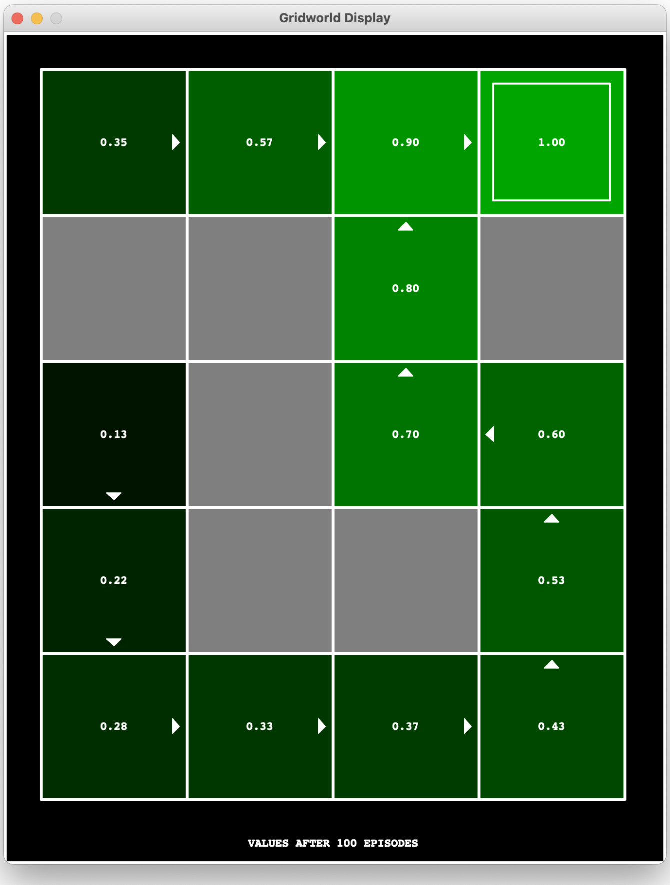

Figure 1: Value function obtained with Q learning with default parameters on MazeGrid after 100 episodes

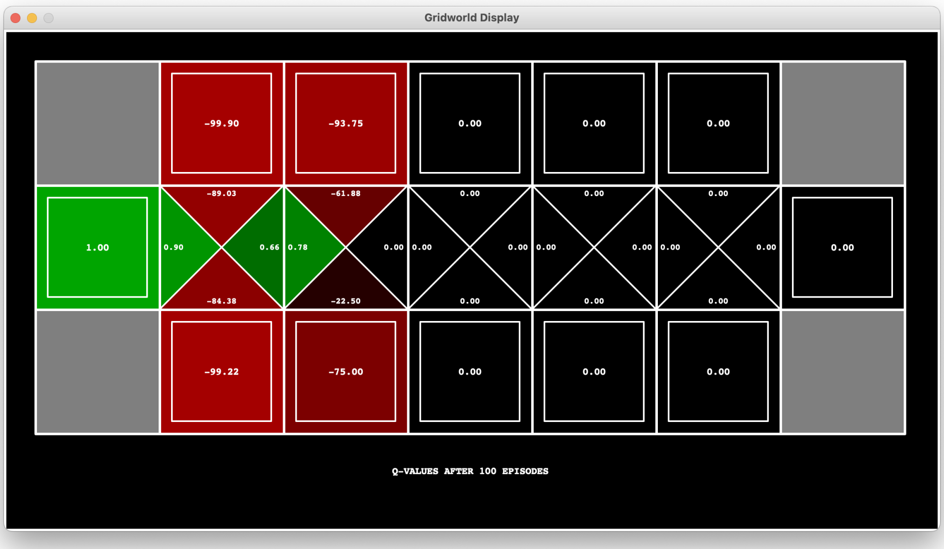

Figure 2: Q-function obtained with Q learning on BridgeGrid without noise after 100 episodes

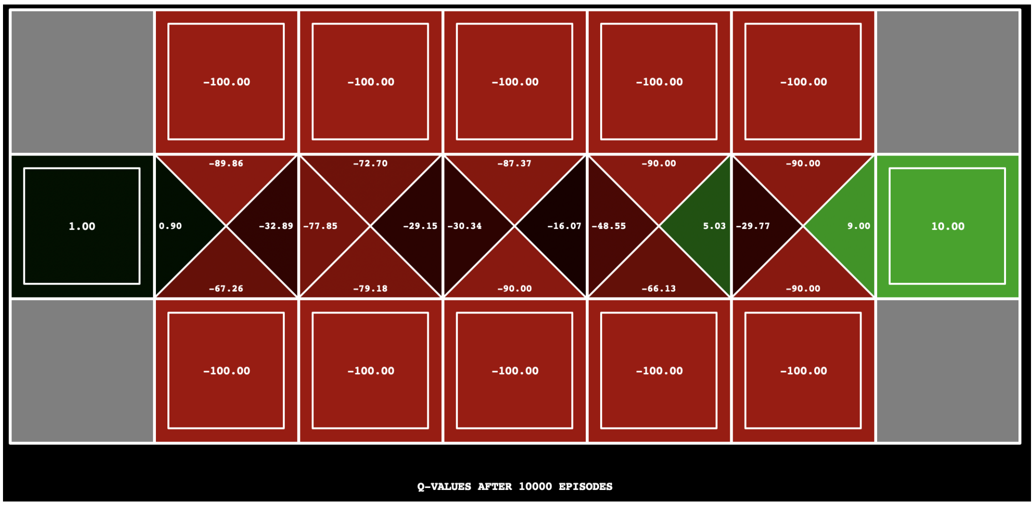

Figure 3: Q-function obtained with Q learning on BridgeGrid without noise after 10000 episodes

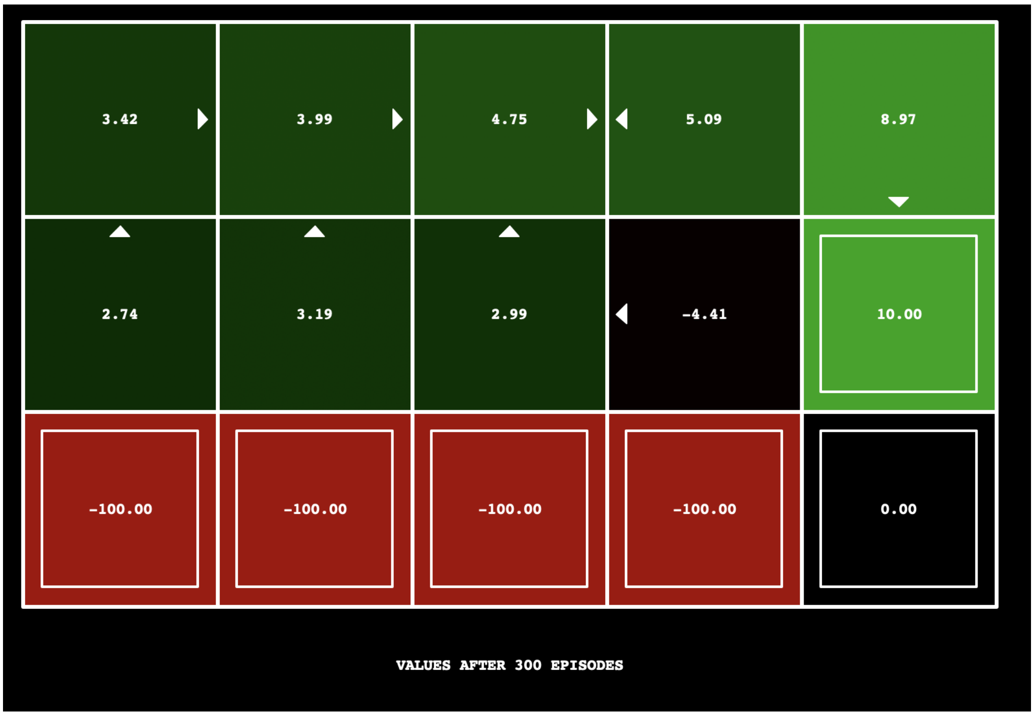

Figure 4: Value estimates obtained with Q learning on CliffGrid after 300 episodes

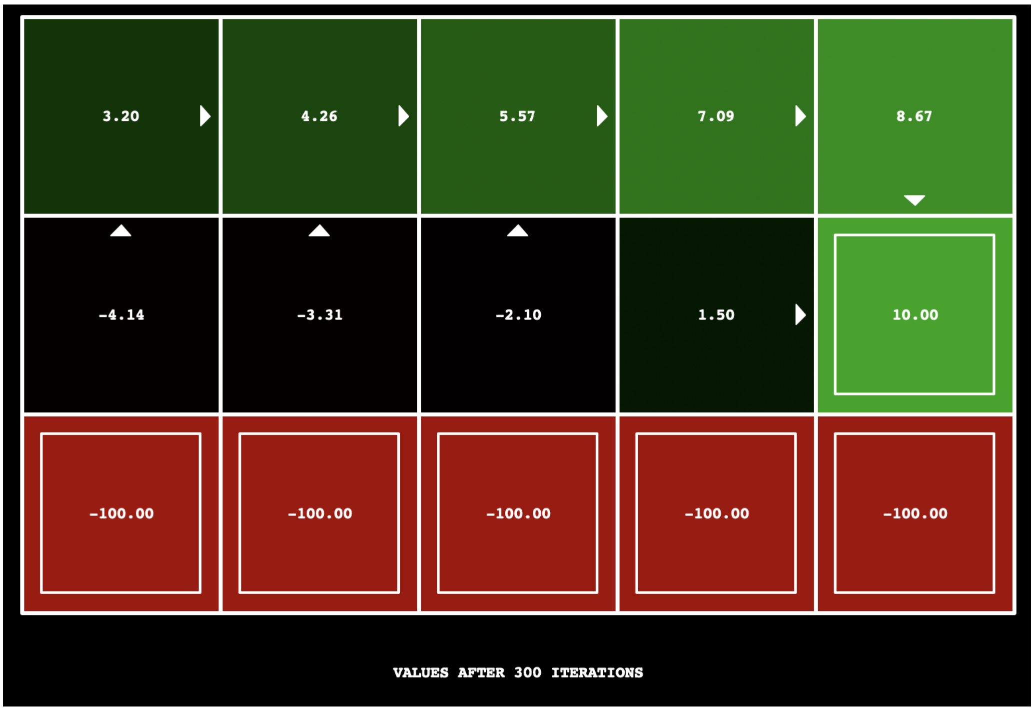

Figure 5: Value estimates obtained with value iteration on CliffGrid after 300 iterations

(d)

We show value estimates for book grid with 300 episodes of Q-learning (Fig 6) and 300 iterations of value iteration (Fig 7). The Q-learning agent is sensitive to the terminal state with negative reward (-1) obtained due to sub-optimal actions of the ϵ-greedy policy, hence leading to avoidant behavior around the negative state. This persists even after 10000 episodes (Fig 8).

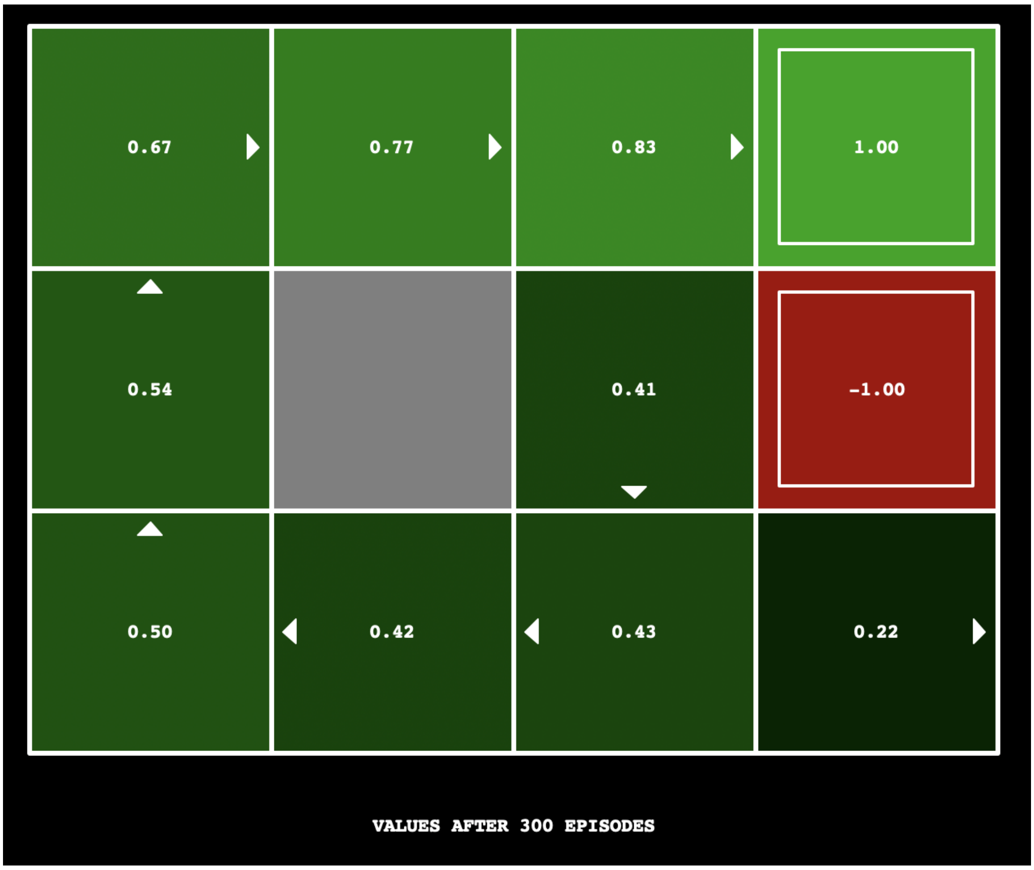

Figure 6: Value estimates obtained with Q learning on BookGrid after 300 episodes

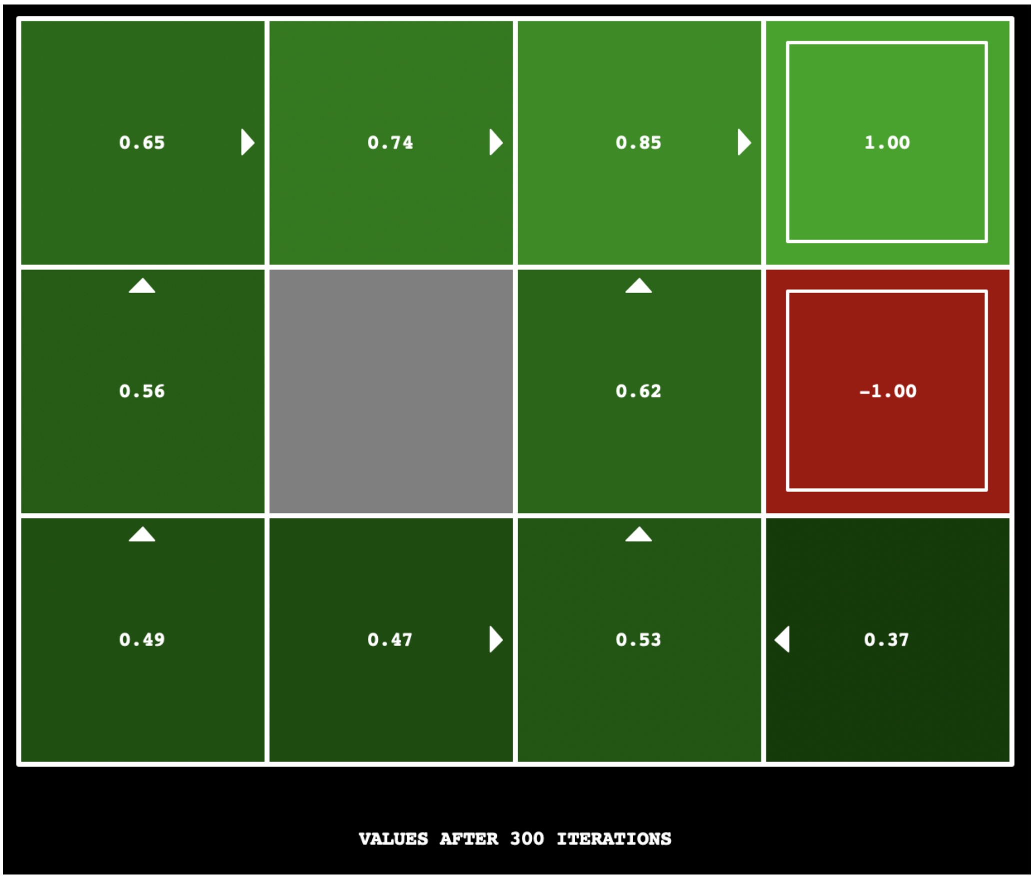

Figure 7: Value estimates obtained with value iteration on BookGrid after 300 iterations

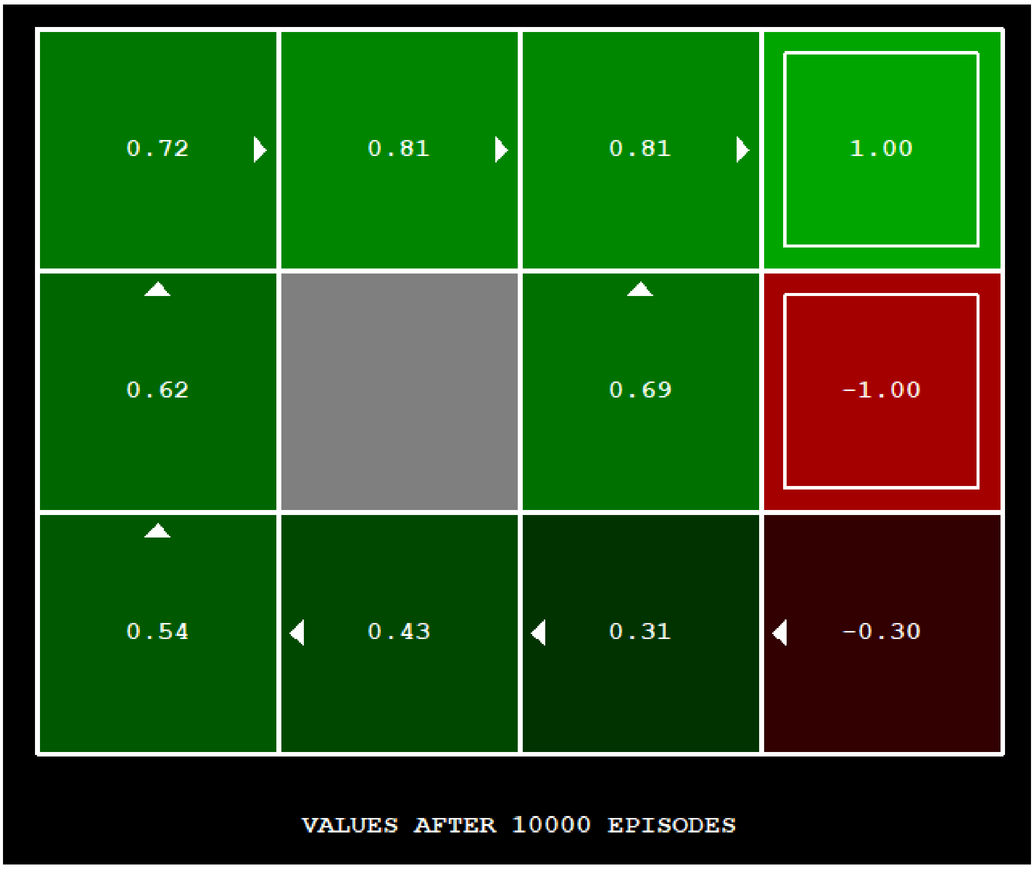

Figure 8: Value estimates obtained with Q learning on BookGrid after 10000 episodes