HW 1: RL Writeup

Lucas McKamey, Matthew Chen

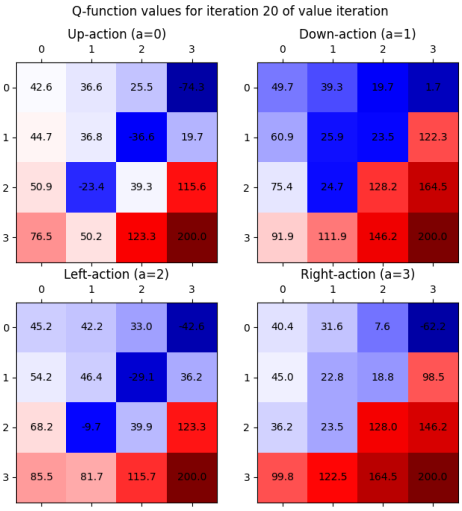


Figure 1: Value iteration (20 iterations; gamma = 0.9) Q-values

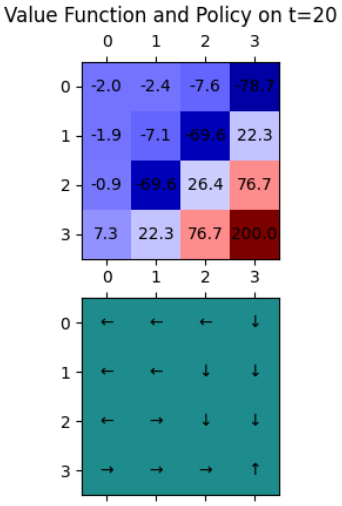


Figure 2: Max policy and corresponding V value (gamma = 0.5)

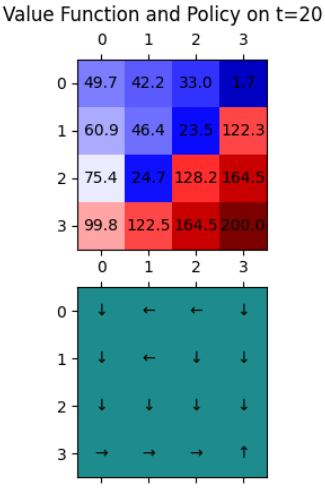


Figure 3: Max policy and corresponding V value (gamma = 0.9)

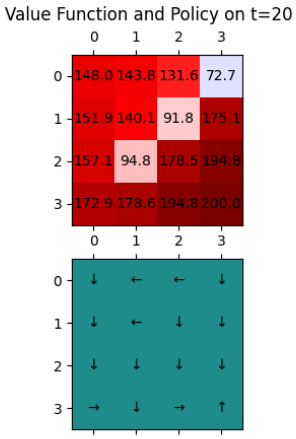
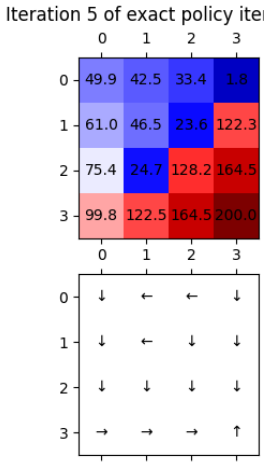
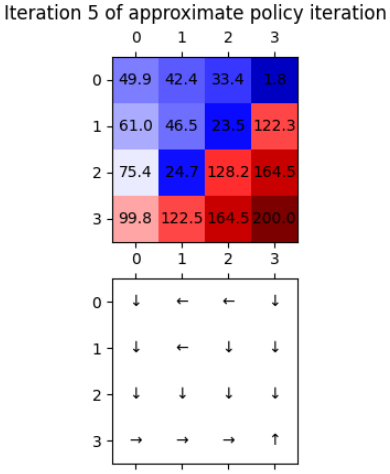


Figure 4: Max policy and corresponding V value (gamma = 0.999)

Policy Iteration (5 iterations)





Discussion Answers

Increasing gamma increases the importance or consideration of future rewards. As a result, when we increase gamma, we see less difference (variance) between the rewards of different tiles in the MDP in the gamma=.999 visualization than we do in the gamma = .9/.5 visualizations.

The exception of the tiles (0,3), (1,2), and (2,1) is visible, as reaching those states yields a -80 penalty, so reaching those states will always be much worse to land on than any other state. This can be seen in the optimal policy when in the state (3,1) which is the state one right to the bottom left corner.

At first glance, logically we would optimally try to move right to reach the bottom right corner (the goal state). However, between figures 3 and 4 we see differing strategies. When gamma is set to 0.9, we see the logical, common-sense strategy, but the optimal policy when gamma is set to 0.999 tells us to go down. This is because reaching the final reward state of the bottom right square faster is much less incentivized with a higher gamma (as future rewards are weighted almost identically to current rewards).

As a result, when considering future rewards almost identically to current rewards, the policy doesn't really care how long it'll take to get to the goal state, so long as it reduces the penalties on the way. Therefore, the optimal policy is to go down, which minimizes the probability of slipping laterally up to the penalty state and receiving a penalty. In summary, the optimal policy on tile (3,1) is to take the .15 chance of slipping to the right when moving down as opposed to incurring a .15 chance of slipping up and incurring a big penalty. Since gamma is almost one, this minimization of penalty state chance is a more beneficial policy than maximizing speed/probability of reward state. In contrast, figures 2 and 3 value distant rewards less and thus attempt to maximize speed/probability of reward state and so "rush" to the finish line.

Conversely, in the gamma = 0.5 case, the policy chooses to go left on tiles (0,0) and (1,0). This is an unintuitive strategy, as there is no risk of slipping into a bad state from moving down on these tiles. You'll also notice that these board states are some of the farthest away from the goal state. Therefore, it appears the policy values future rewards too lightly and so doesn't have a major incentive to start moving towards the reward (as much as it wants to avoid any possibility of incurring penalties from the bad states). In this case, it's almost as though the policy is too scared of the bad states that it fails to make intuitive forward progress towards the goal state.