CS 4789 PA 1 Discussion

Chad Yu

February 10, 2024

1 Question 1

Firstly, we notice the value function gets higher for all states as H increase except the goal state, and the further away the state is from goal state, the more the value increases. The value function and policy at time 0 are different for different values of H, because the Bellman Optimality Equation for finite horizons uses the value function at time $h+1$ to determine the value function at time h, so that the dynamic programming algorithm is backwards iterative. The value increases for all states because of the recursive definition of the Bellman Optimality Equation, in which the weighted average (expected value) of the value function at the next time step dominates any negative reward from $r(s, a)$, as the reward from the goal state makes its way into the value function. Intuitively, given more steps, one is more likely to end up at the goal state at some point or it becomes possible to reach the goal state from a certain state, so that the expected reward becomes greater.

Another observation relating to the differences in policy is that for $H = 10$, the actions determined by the policy seem to follow more of a pattern than the policy for $H = 1000$; the arrows seem to draw out a clear path towards the goal state and away from the bad states. This can be explained by the fact that the dynamics of the Markov Decision Process becomes more complex as we add time steps, as more time steps simply leads to more number of trajectories. This means the optimal action taken might not be the one that intuitively makes sense.

2 Question 2

The value function and optimal policy for timestep 8 for $H = 10$ and at timestep 0 for $H = 2$ are the same. Because dynamic programming is backwards iterative, and given that we set V_{H-1} and π_{H-1} to be the same vectors for all H, we have that for any k, V_{H-1-k} and π_{H-1-k} will remain the same no matter what H is. This is because the dynamic programming algorithm always takes the greedy choice w.r.t argmax when determining an optimal policy.

3 Question 3

The first observation we note is that the value function associated with the max policy increases for all states as γ increases. By the convergence of VI as shown in class, we have that for smaller γ, $\|V_i - V^\star\|$ decays faster to 0, as it is bounded by $\gamma^i \|V_0 - V^\star\|_\infty$, so for smaller γ, the algorithm converges in less iterations. As described in the first question, we have that since larger γ take more iterations, there are more possible trajectories to the goal state, whose reward dominates any negative reward that is taken at any other state, and in the expectation expression of the Bellman Optimality Equation, this helps create a net positive as the number of iterations are increased.

The second observation we notice is that for $\gamma = 0.5$ and $\gamma = 0.9$, the actions determined by the optimal policy seem to follow a pattern, while for $\gamma = 0.999$, the actions seem more random/chaotic. For the $\gamma = 0.9$ case, the arrows even seem to point out paths away from bad states and towards the goal state. As mentioned previously, for larger γ, convergence takes more iterations, and after many iterations, we have that the dynamics of the Markov Decision Process becomes complex, so that optimal policy might not be intuitive or pattern-like.

4 Question 4

They are similar in that as H and γ both increase, they follow the same trends of value functions and optimal policy. The policies depicted seems to follow a pattern for lower values of γ and H, while the highest value of γ and H both have a policy that seems to be less intuitive. In fact, the policy for $H = 1000$ and the max policy for $\gamma = 0.999$ are exactly the same, as seen in Figure 2 and Figure 7 below. As mentioned in Question 3 and Question 1, the similarity arises from the change in policy is because larger γ requires more iterations to converge, and more iterations creates a more complex dynamics, which may generate an non-intuitive policy. Also, both sets of value functions increase for every state as H and γ increase; as explained before,

this is because a larger number of iterations results in more trajectories that could possibly reach the goal state, which brings a much larger net positive from the reward of the goal state into the expectation.
The main difference that we notice is that the value function for each state is different for each of the mentioned γ, H pairs, but as γ and H increases, the difference between these value functions becomes less. Specifically, the value functions of each state for each value of γ is less than the value functions of each state for the associated horizon length H. This can be attributed simply to the definition of the discounted problem in infinite horizons; the value only considers a γ^k fraction of the reward for k time steps in the future, as opposed to the finite horizon case which considers the total expected reward. Thus, it makes sense that the difference is larger for smaller γ, as γ^k gets exponentially smaller as we step further into the future. Also, it makes sense that we observe little difference for $\gamma = 0.999$, as 0.999 is very close to 1, which would result in the same expected value objective as in the finite horizon case.

5 Figures

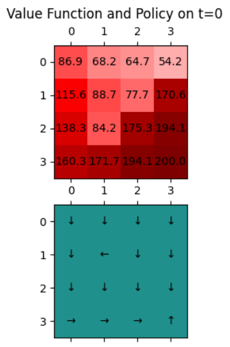

Figure 1: Value Function and Policy for H = 10

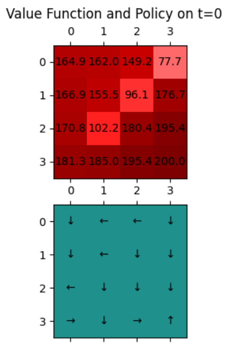

Figure 2: Value Function and Policy for H = 1000

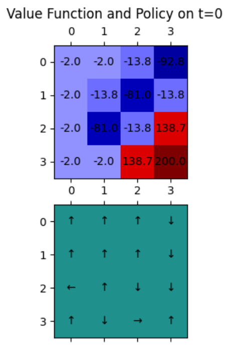

Figure 3: Value Function and Policy for H = 2

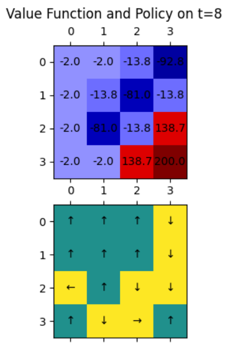

Figure 4: Value Function and Policy for H = 10 and t = 8

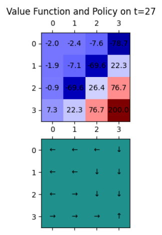

Figure 5: Max Policy and Value Function VI for $\gamma = 0.5$

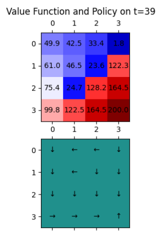

Figure 6: Max Policy and Value Function VI for $\gamma = 0.9$

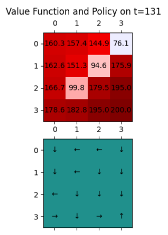

Figure 7: Max Policy and Value Function VI for $\gamma = 0.999$

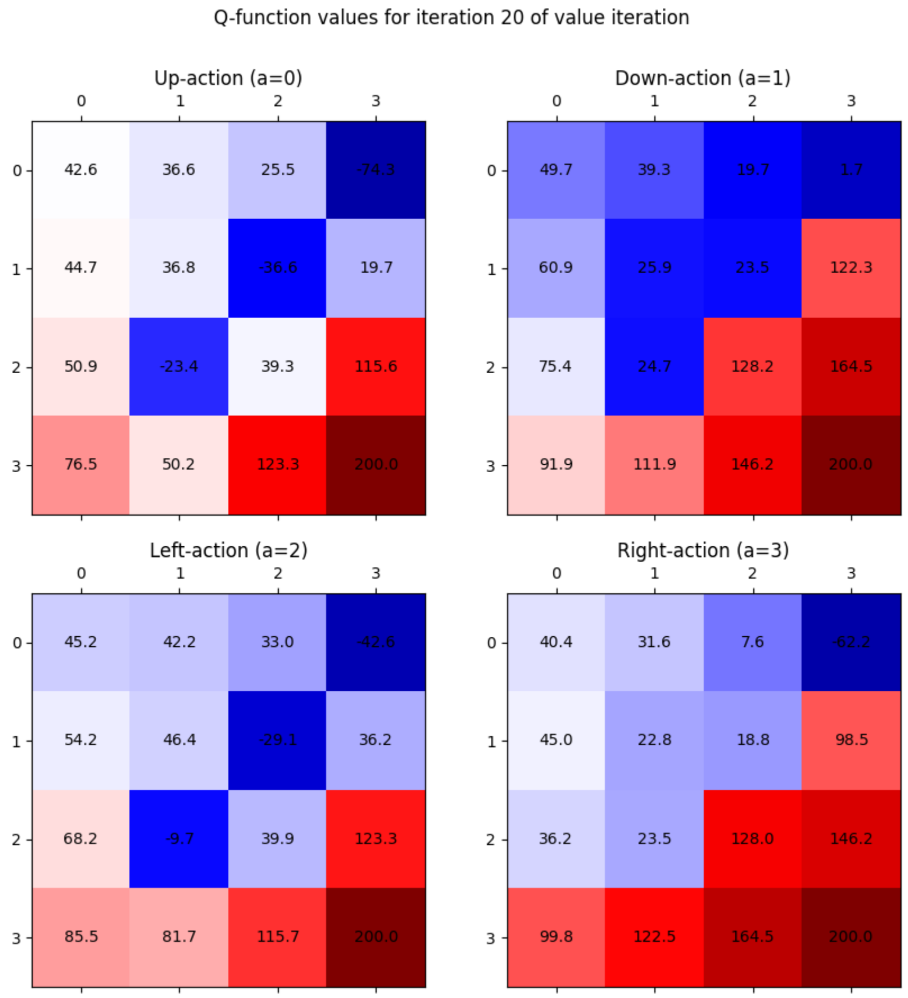

Figure 8: Q-Function VI for 20 iterations for $\gamma = 0.9$

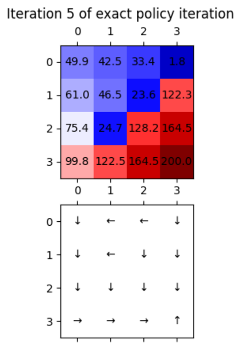

Figure 9: 5 Iterations of PI Exact Value Function and Policy

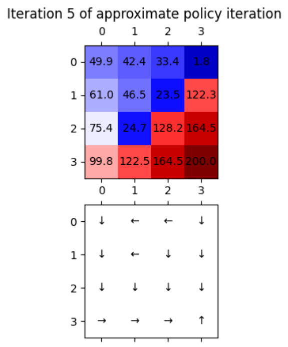

Figure 10: 5 Iterations of PI Approx Value Function and Policy