Markov Decision Processes

Gary Saavedra

gsaavedra3@gatech.edu

Introduction

Small Grid World

*Why is it interesting?*

The small grid world is a small MDP.

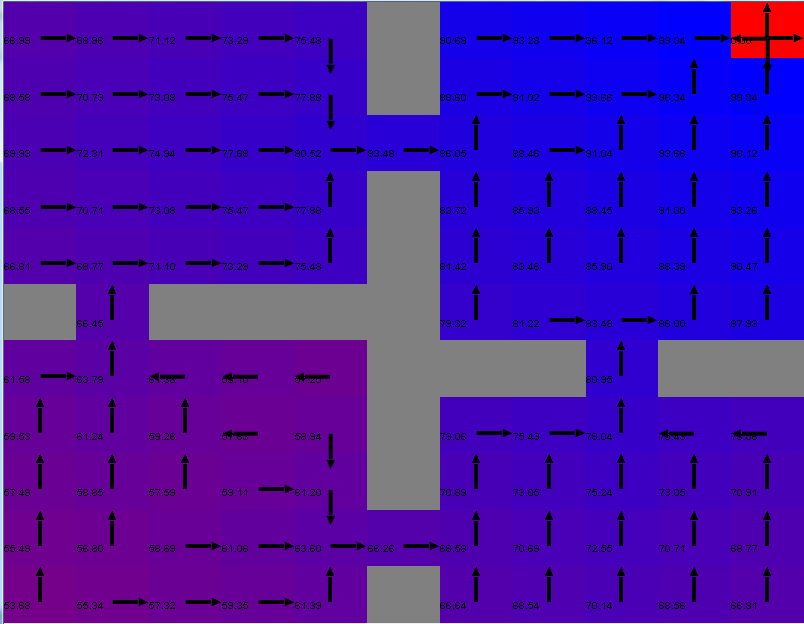
Value Iteration

Decreasing the discount factor decreases the number of iterations to find the optimal policy. But this is because the recursive application of the discount factor is making the difference between the later value functions smaller than delta much earlier than it would with a large discount value. This leads to a weak policy being found which in turn increases the number of iterations it takes for the agent to reach the absorbing state. This is because the optimal policy will only get an optimal value for the first few squares. After that the agent must choose randomly amongst the remaining squares. This can especially be seen with discount=1.

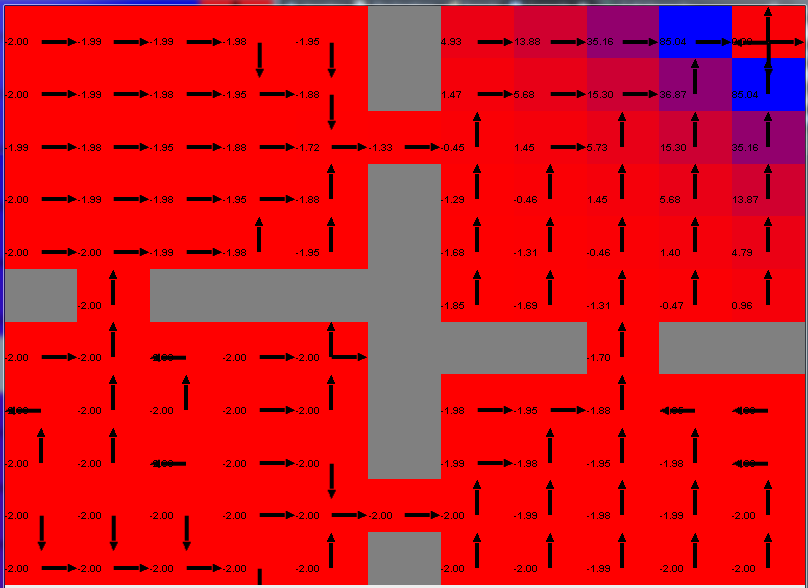
Figure … shows a policy map when gamma = 0.99. Here future reward is not being discounted much and thus the reward is able to propagate out from the goal state. This creates a clear policy for the agent to follow.

Figure … shows a policy map when gamma = 0.5. Since future reward is discounted by such a large factor the utility is not able to propagate out from the goal state. The red indicates that the utility for those states is very low at about 5. The utility of approaching the goal state is not realized until a few squares before as shown by the blue in the top right corner. This is why some parts of the grid can show the agent wandering in non-ideal directions. The agent does not know about the future reward because it is too far away and has been discounted.

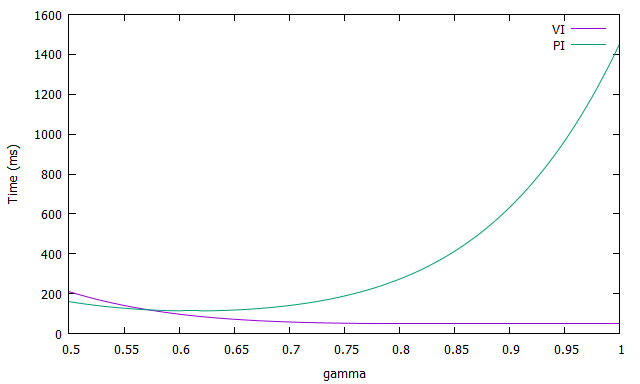
Graph ideas. Passes vs discount factor. Iterations needed to get to goal vs passes or discount factor. Convergence measure vs time



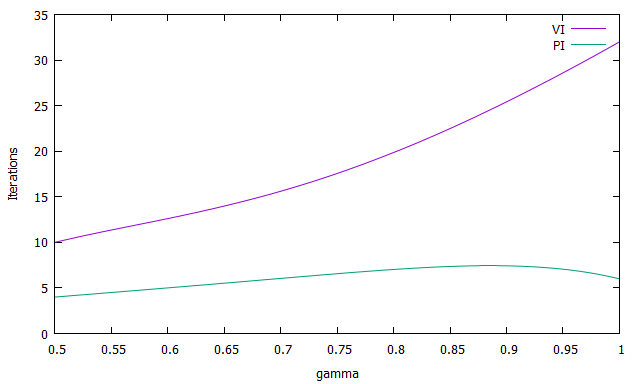
*Figure : VI, Policy map for gamma = 0.99*



*Figure : VI, Policy map for gamma = 0.5*



*Figure : Time vs gamma for PI and VI, Small grid*



*Figure : Iterations vs gamma for PI and VI, Small grid*

**Policy Iteration**

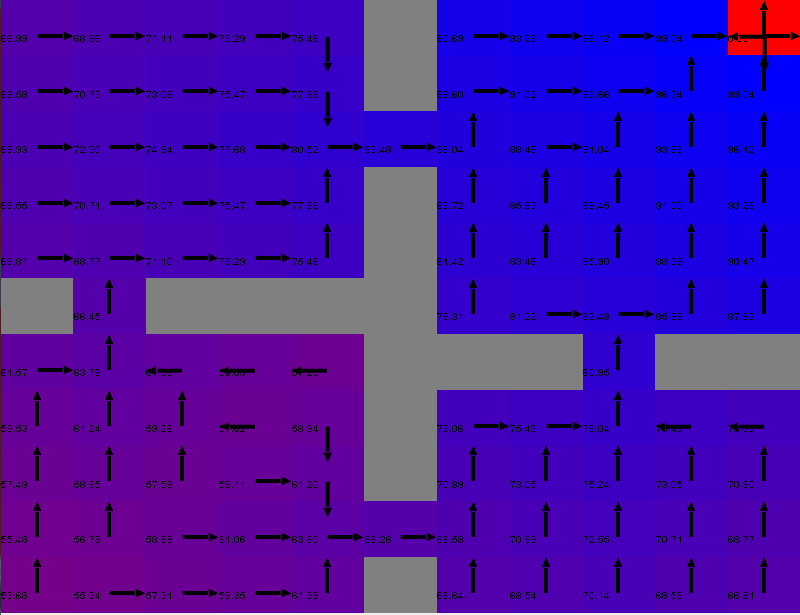
Policy iteration typically takes less iterations to converge than value iteration. However, each iteration tends to be more expensive. Each iteration has an inner value iteration that computes a value based on actions defined by the current policy.

For both PI and VI number of iterations increases as gamma decreases. Good for a graph.

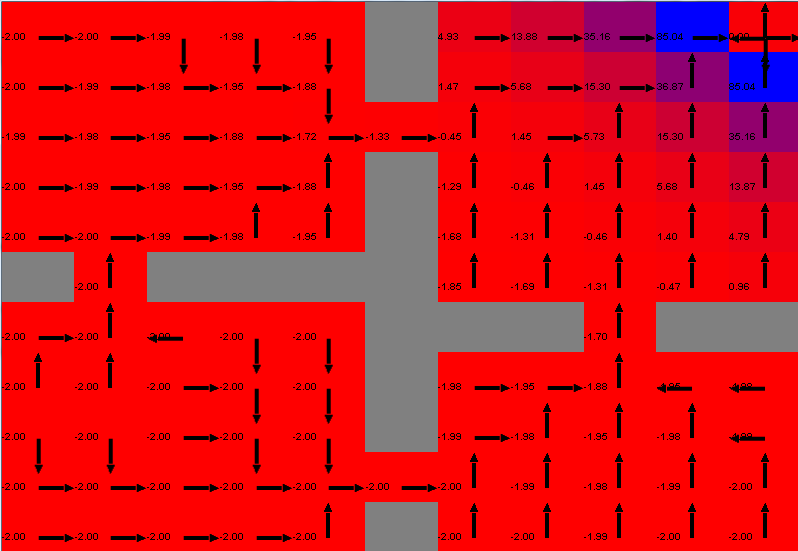
Figure … shows the policy map when gamma = 0.99. The policy is almost identical to the policy found by value iteration. The only difference is in the bottom left corner in the starting room. Explain why.

Figure … shows the policy map when gamma = 0.5. Similar to VI the small discount factor keeps the utilities from propagating out from the goal state. The policy found is different from the policy found with a larger gamma.

Show the average time per iteration for VI and PI. PI iterations take longer and thus have a higher average. However, each iteration takes less time than the previous because it requires less value evaluations.



*Figure : PI, Policy map for gamma = 0.99*



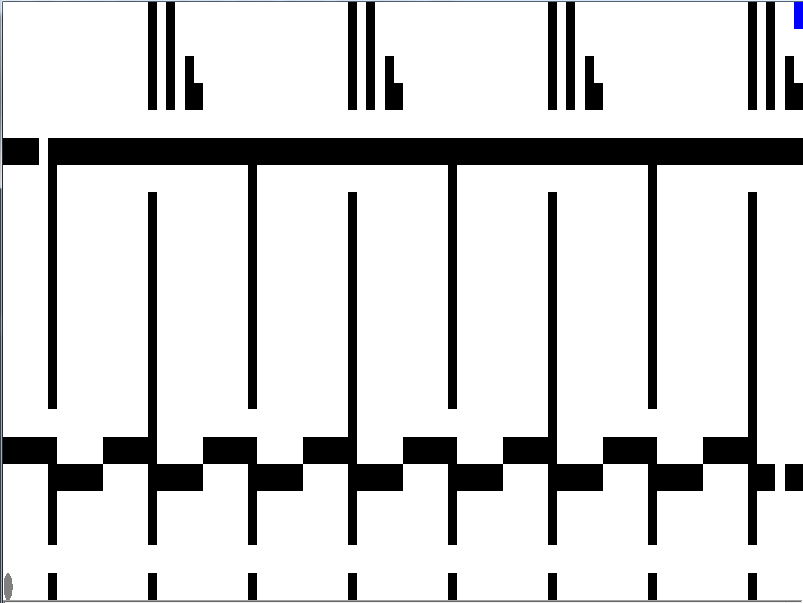
*Figure : PI, Policy map for gamma = 0.5*

**Large Grid World**

***Why is it Interesting?***

The large grid world is shown in Figure… The agent starts in the bottom left corner and is trying to reach the goal state in the top right corner marked by the blue square.

Talk about rewards fir reaching goal state etc.



*Figure : Large grid world*

**Value Iteration**