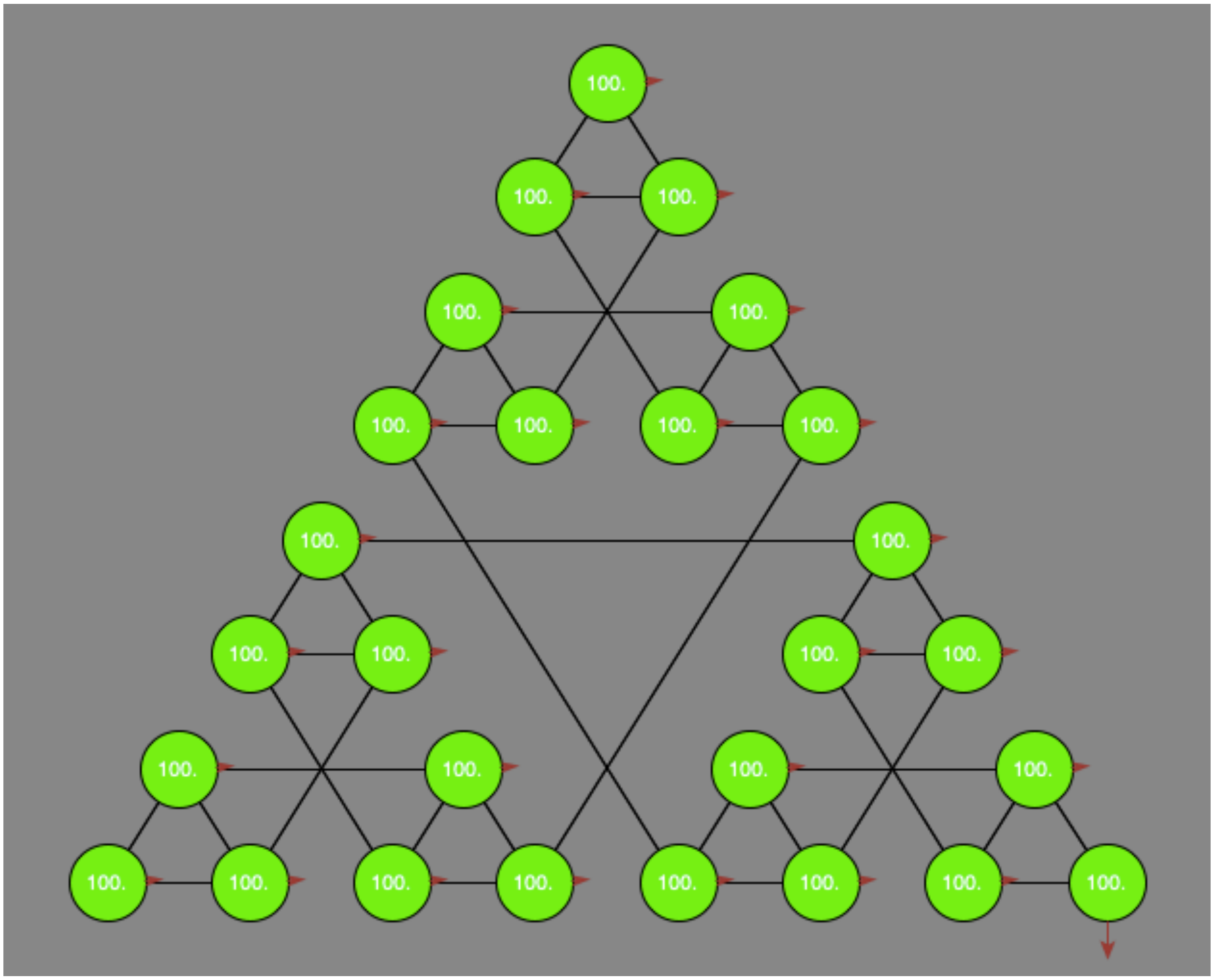
Zachary McNulty
zmcnulty, 1636402
CSE 415 HW5 Report

1) No noise/discounting/living reward
   a) 4 Value Iteration steps to turn 1/3 of the states green.
   b) 8 Value Iterations steps to turn all states green.
   c) No, this policy is quite bad. Many of the policy choices are illegal actions which have high Q-state values simply because they result in staying at the same state. Since all the states have the same value of 100 after 8 steps of value iteration, the same value as the goal as there is no living reward/discounting, pretty much any action is deemed optimal: they all end at a node of value 100. As there is no noise in this scenario, an illegal move as our policy will leave us stuck at that node forever. In fact, following the current policy will never bring you to the goal state. As a result, this policy is incredibly bad.

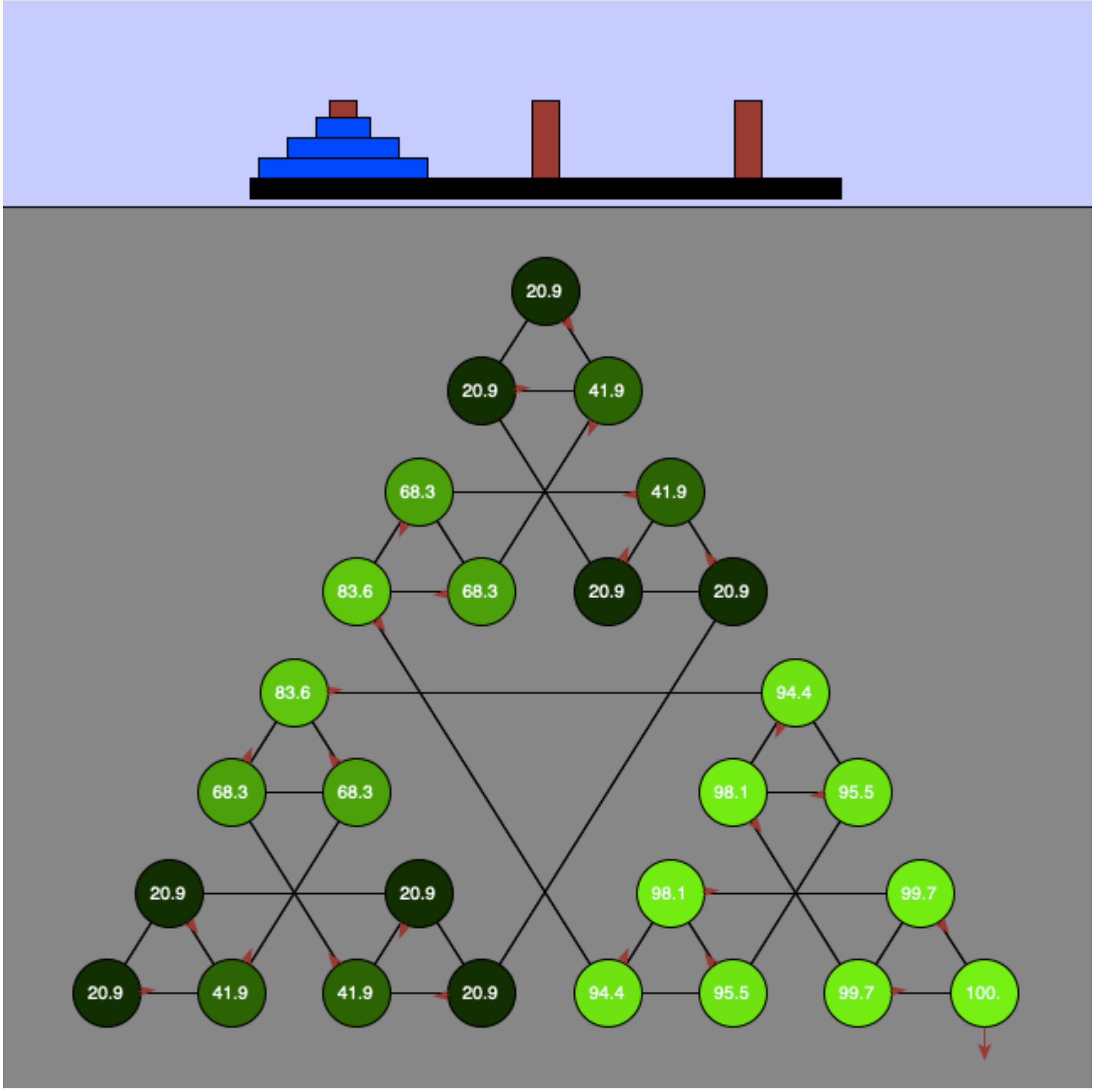


2) 20% Noise; no discounting/living reward

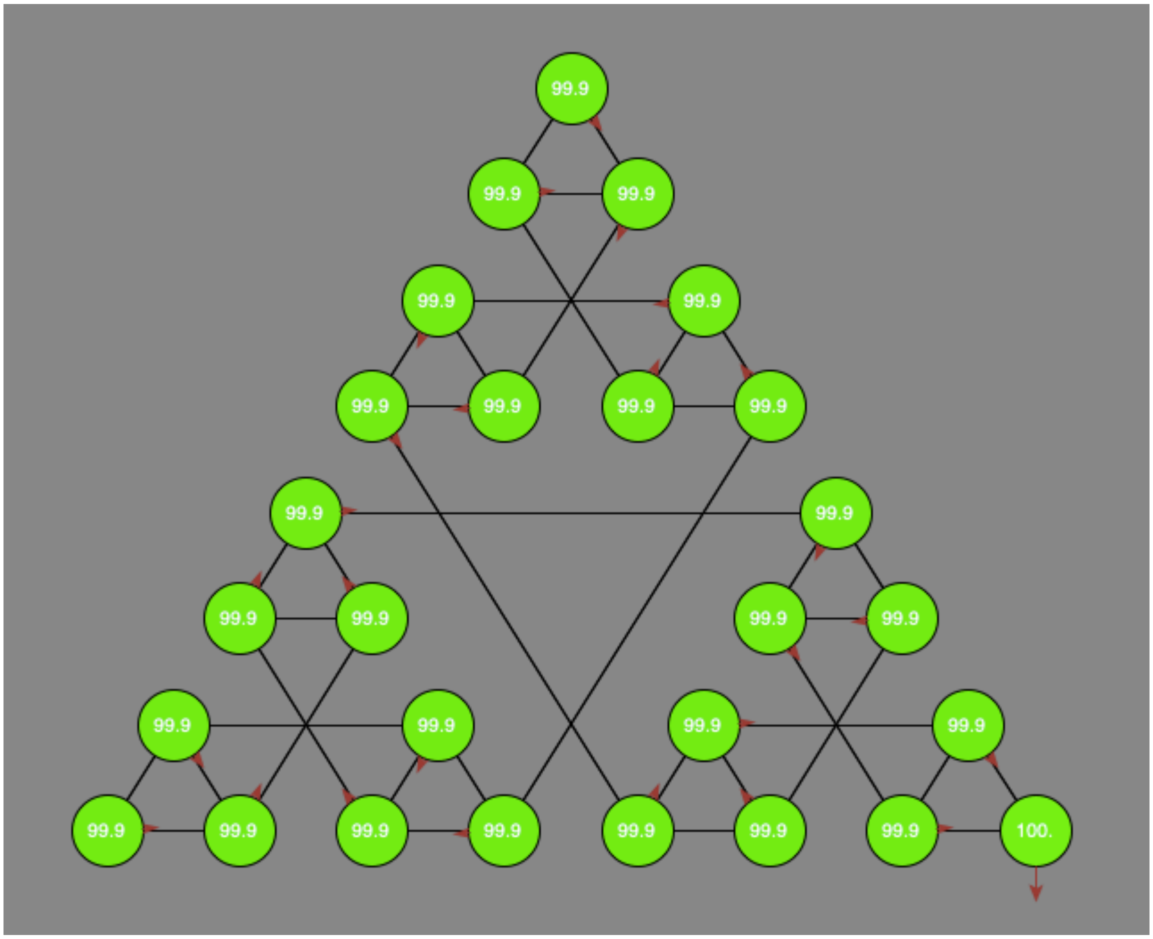
   a) 8 Value Iteration steps until start receives nonzero value.

   b)

   c) The policy seems fairly good. There are no longer any illegal actions and the policy points along the optimal path. There are no cycles in the graph that could lead the agent to get stuck for awhile until noise frees it, and in general it seems that each policy points in the direction of the shortest path to the goal. As there are no “bad” exit states like there were in some of our grid world examples, there is no reason to deviate from the most direct path. Especially since the noise is relatively low, I feel this policy will optimally bring you to the goal state.



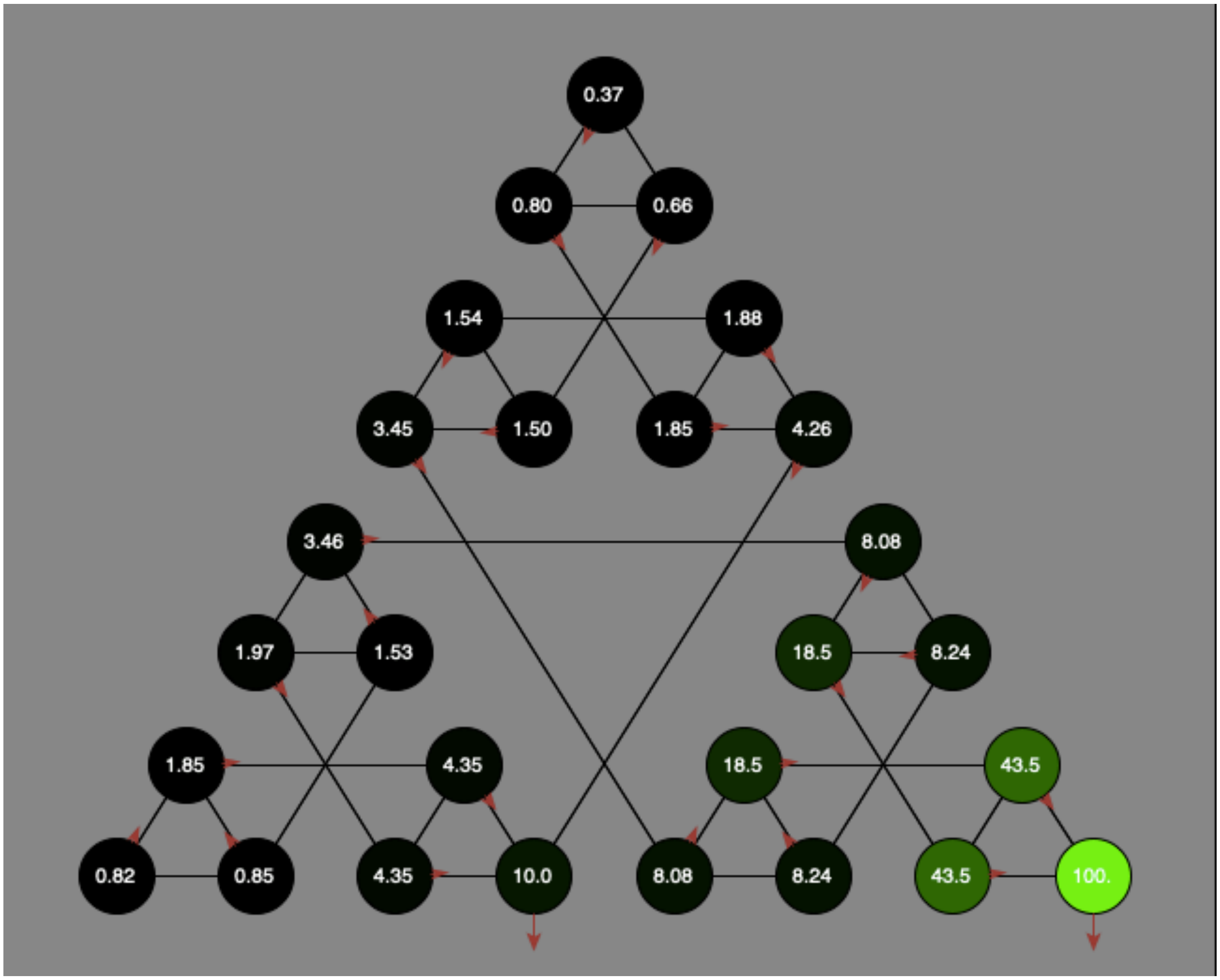
d) 56 VI steps (in total) to converge.



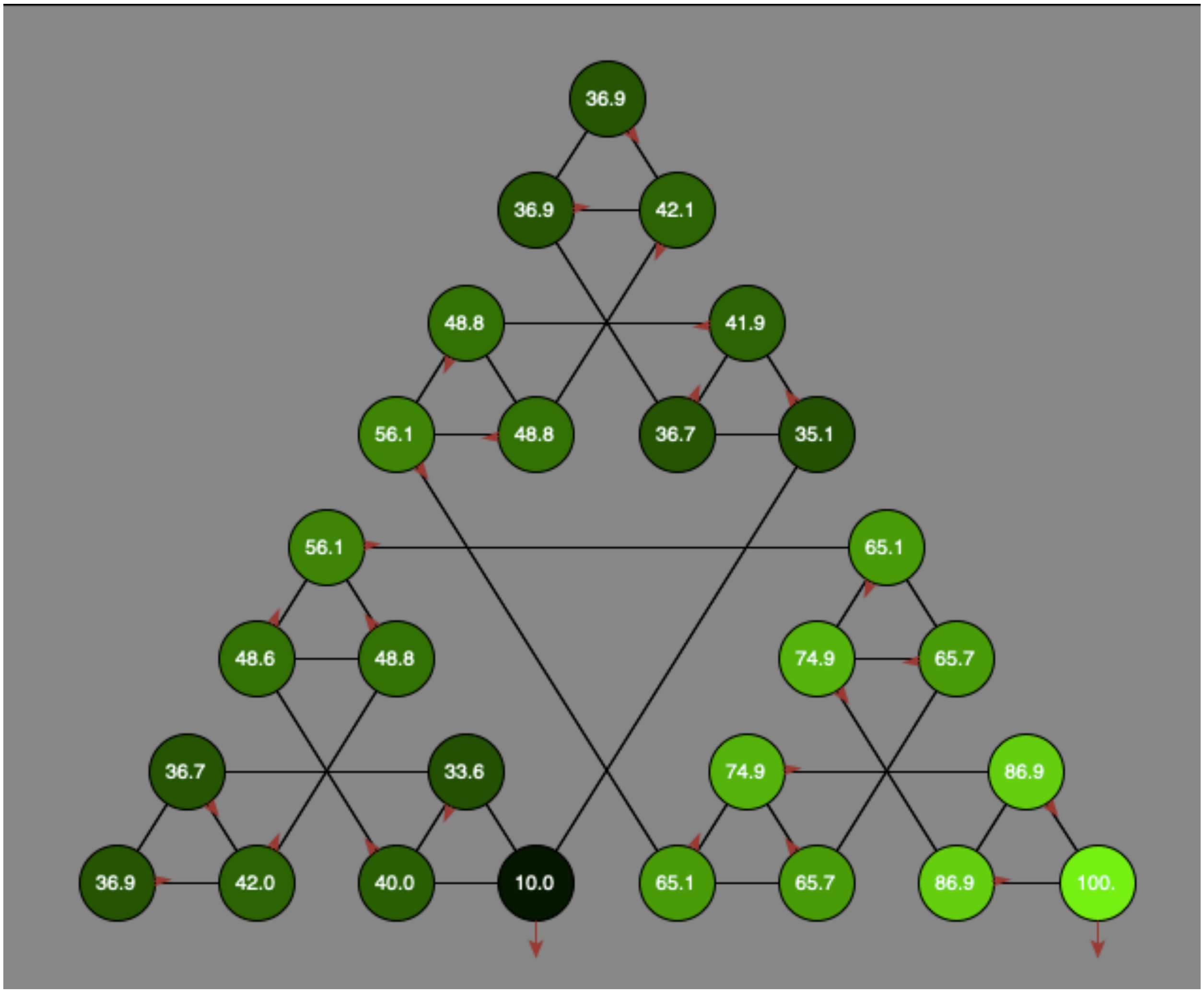
e) The policy has not changed from what we found during step 2c. By the time we had generated our policy during step 2c, we had explored pretty much the entire board and covered the entire optimal route. While exploring the board further increases the general V values at each state, it like likely does not change the relative values of each of the given states, and the maximizing actions are maintained.

3) 20 % Noise, 2 goals, discount of 0.5

   a) Start state has value 0.82. The policy typically points to the nearest goal state, whether it is the 100 or 10 goal state. The 100 state seems to be slightly more favored. This is likely because the discount rate is so high. Even though the 100 reward is ten times greater, if it is more steps away than the 10 goal state, the discounting quickly negates the benefit of the higher reward.



b) Start state has value 36.9. The new policy almost exclusively leads to the 100 goal state. Because the discount rate is relatively low and the difference between the two goals states (10 vs 100 reward) is so high, the 100 reward state is almost always prefered even if it takes many more steps to reach it.



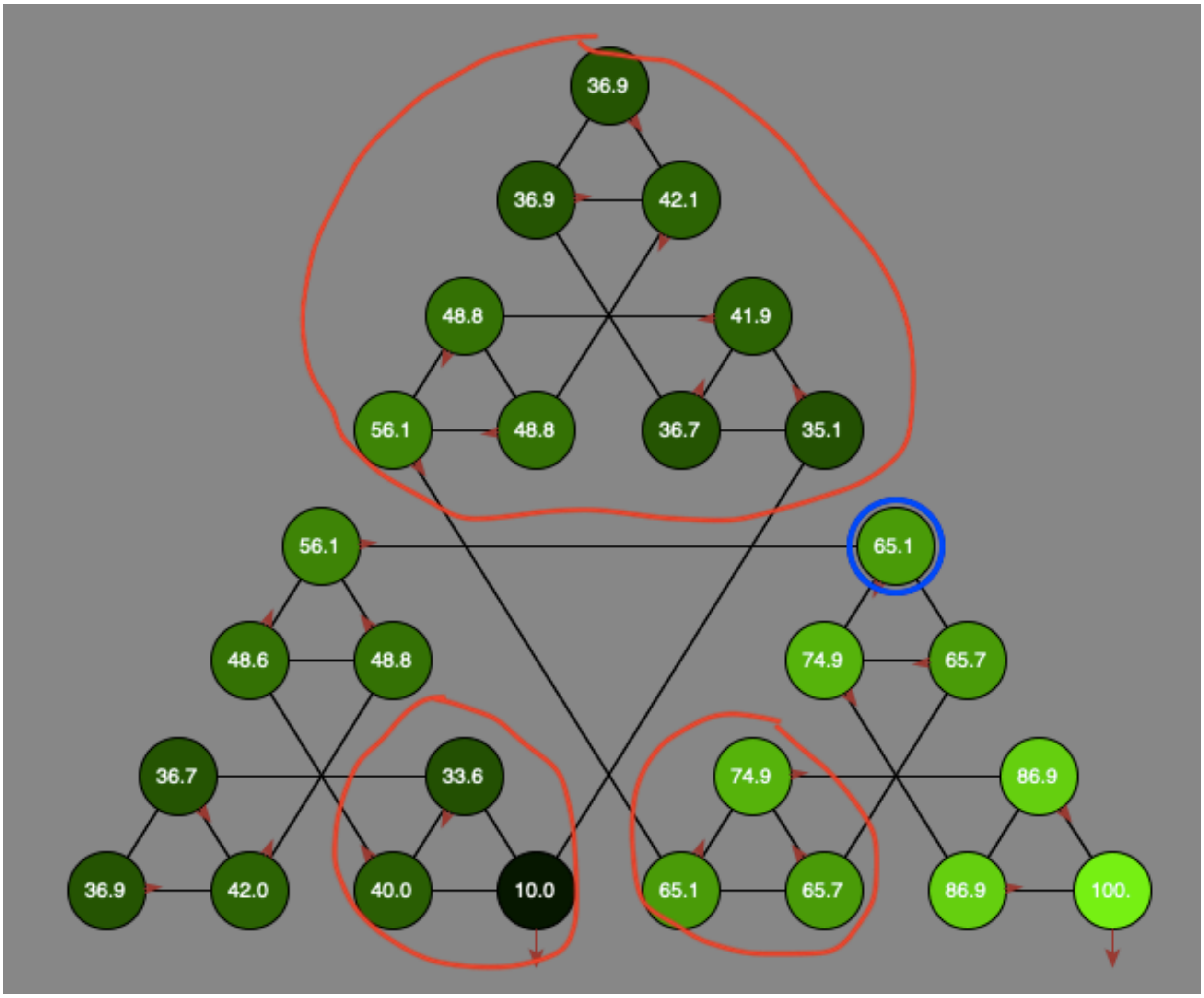
4) Simulating agent

a) 8/10 went off path

b) 6/10 goal state reached;

c) 1 step away; 3 steps away; 3 steps away; 2 steps away

d) The circled areas were never (or only seldom) visited.



5) Overall Reflections

a) No, it is not necessarily essential that the values of the state converge first for a good policy to be generated. As we saw in 2c, as long as the values are a fair approximation of the true value or the states have approximately correct values relative to each other it is possible to generate the optimal policy without convergence. It is this relative value of the states that matter as that entirely dictates the policy when the transition reward is constant across all actions.

b) Not incredibly important. Some values can be inferred from the values of their neighboring states, so as long as the states value is updated eventually you could get a fairly good approximation for its value without visiting it too often. In the case of some of the states we see in 4d, their true value is not so relatively important as these branches of the tree are seldom visited. Essentially, we just have to visit some nodes enough to realize they are a bad state to be in and from there can focus on the other nodes where an optimal path is likely to occur on.