Problem 5: Value-Iteration Analysis

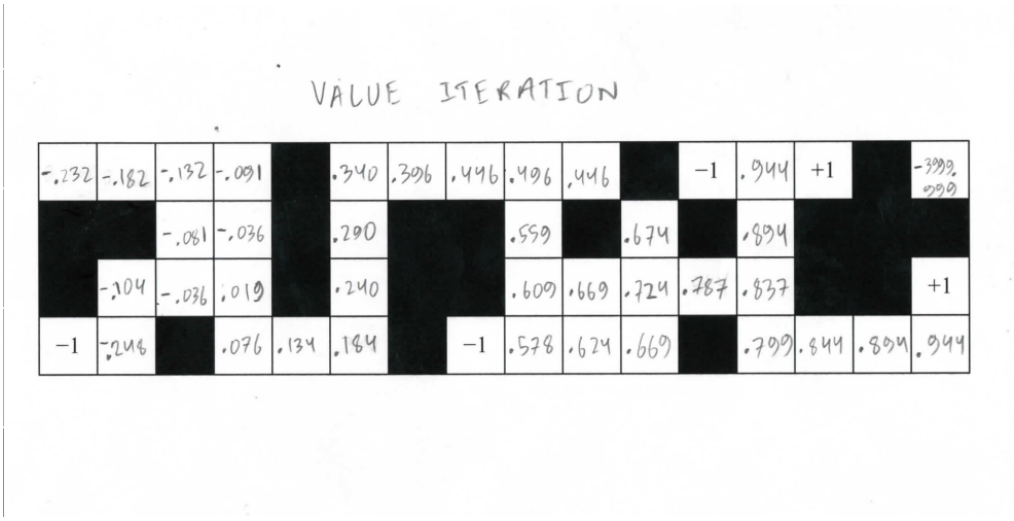
Test

We tested the completed and compiled versions of the algorithm implemented by Jerod Weinman on the 4x3 grid world which he also provided to us. We held gamma to be .99999 and epsilon .001. We kept the discount factor high so that our algorithm would take into account as much of the future expected utility as possible. This decision was motivated by the instructors note in the Piazza group for CSC-261 S2018 (post title: Values for sigma and gamma). We decided to restrict the margin for error so we got as close to the convergence values as possible as seen in figure 17.3 (Russell & Norvig, pp651).

Predict

We predict that the states that are very close to the terminal state with the highest utility will have a higher utility. At the outset, the agent will likely value paths that keep clear of states which are associated with negative utilities. The numbers we see should show a clear tendency to avoid negative states and weave a path through the state space with utilities getting higher as we get closer to the highest-utility-terminal-state. Because we kept the gamma value high, the states which are right next to our highest-utility-terminal-state will exert a greater influence on the utility values and this will ensure we get an efficient path.

Experiment



Reflect

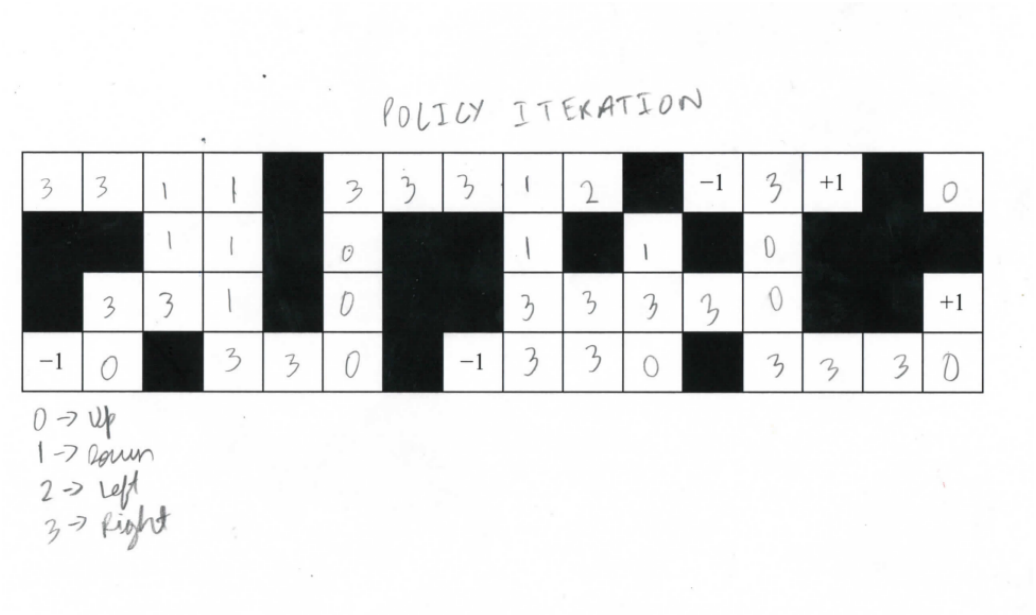
The results of this experiment largely reflect our expectations. States further away from highest-utility-terminal-states on the x-axis have progressively lower utility values. Furthermore, if two states are equally far away from the highest-utility-terminal-state on the x-axis but one is near a state with -1 utility then the utility of that state will be lower. It was also interesting for us to see just how nuanced the utility values ended up being, which demonstrates the way that the bellman update slowly converges.

Problem 6: Policy-Iteration Analysis

Test
We tested the completed and compiled versions of the algorithm implemented by Jerod Weinman on the 4x3 grid world which he also provided to us. We held gamma to be .99999 and epsilon .001, as earlier, in an attempt to ensure we could recreate the values found in Fig. 21.1 (Russell & Norvig, pp832).

Predict
Based on our experience with the 4x3 world we expect the final policy produced by the instructors implementation of policy iteration algorithm to provide the shortest path to the highest-utility-terminal-state. In the case of the 16x4 world this would mean snaking towards both highest-utility-terminal-states. In the squares immediately adjacent to a +1 square we expect the action to be exact one move required to reach that highest-utility-terminal-state.

Experiment



Reflect
The results confirmed most of our expectations. The policy from all squares from which there was a legal move towards a highest-utility-terminal-state was that move. Initially, we did not know how the policy would look for (13,4) because it is sandwiched in between a highest-utility-terminal-state and negative-utility-terminal-state. It makes sense that the algorithm does not need to avoid that path to highest-utility-terminal-state simply because it is next to a negative terminal state. This is because when it is caught in between (14, 4) and (12,4) there is one clear answer of what to do.