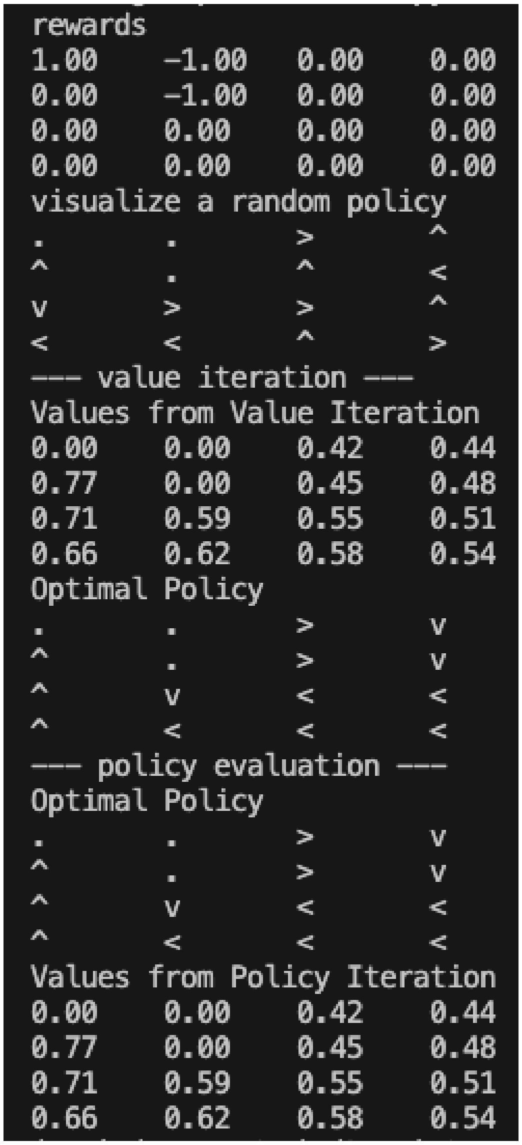
CS 6955 HW3: Markov Decision Process (MDP)

Kutay Eken
u1322888

February 2025

1 Part 1:

Optimal Policy for the Default Configuration MDP



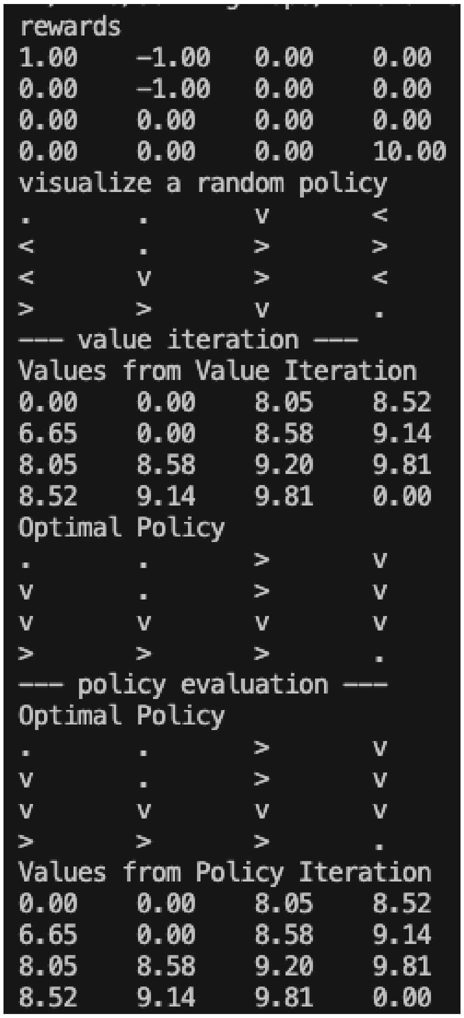
```
rewards
1.00    -1.00   0.00    0.00
0.00    -1.00   0.00    0.00
0.00    0.00    0.00    0.00
0.00    0.00    0.00    0.00
visualize a random policy
.       .       >       ^
^       .       ^       <
v       >       >       ^
<       <       ^       >
--- value iteration ---
Values from Value Iteration
0.00    0.00    0.42    0.44
0.77    0.00    0.45    0.48
0.71    0.59    0.55    0.51
0.66    0.62    0.58    0.54
Optimal Policy
.       .       >       v
^       .       >       v
^       v       <       <
^       <       <       <
--- policy evaluation ---
Optimal Policy
.       .       >       v
^       .       >       v
^       v       <       <
^       <       <       <
Values from Policy Iteration
0.00    0.00    0.42    0.44
0.77    0.00    0.45    0.48
0.71    0.59    0.55    0.51
0.66    0.62    0.58    0.54
```

2 Part 2:

2.1 Custom MDP 1:

For the first custom Markov Decision Process (MDP), I modified the reward structure and terminal states. Specifically, I assigned a reward of +10 to the bottom-right corner cell and designated it as a terminal state. All other aspects of the environment remained unchanged from the default configuration. The output I obtained is provided below:

Optimal Policy for Custom MDP 1



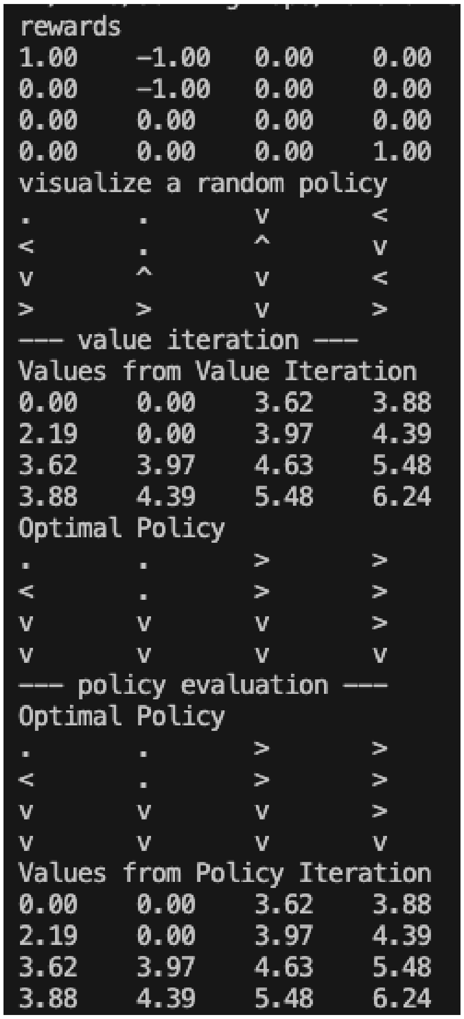
```
rewards
1.00    -1.00   0.00    0.00
0.00    -1.00   0.00    0.00
0.00    0.00    0.00    0.00
0.00    0.00    0.00    10.00
visualize a random policy
.       .       v       <
<       .       >       >
<       v       >       <
>       >       v       .
--- value iteration ---
Values from Value Iteration
0.00    0.00    8.05    8.52
6.65    0.00    8.58    9.14
8.05    8.58    9.20    9.81
8.52    9.14    9.81    0.00
Optimal Policy
.       .       >       v
v       .       >       v
v       v       v       v
>       >       >       .
--- policy evaluation ---
Optimal Policy
.       .       >       v
v       .       >       v
v       v       v       v
>       >       >       .
Values from Policy Iteration
0.00    0.00    8.05    8.52
6.65    0.00    8.58    9.14
8.05    8.58    9.20    9.81
8.52    9.14    9.81    0.00
```

As observed in the output, the optimal policy has changed significantly. From the optimal policy, all the selected actions lead to the +10 reward, meaning the agent now focuses entirely on the +10 reward and completely ignores the +1 reward. This change in the optimal policy is exactly what I expected since the goal was to give the agent a better reward and shift its focus away from the +1 reward, which was the best choice in the default environment. The new optimal policy shows that my agent prioritizes rewards and tries to maximize them.

2.2 Custom MDP 2:

For the second custom Markov Decision Process (MDP), I modified the noise parameter by setting its value to 0.4 and added a +1 reward to the bottom-right corner cell. All other aspects of the environment remained unchanged from the default configuration. The output I obtained is provided below:

Optimal Policy for Custom MDP 2



```
rewards
1.00    -1.00   0.00    0.00
0.00    -1.00   0.00    0.00
0.00    0.00    0.00    0.00
0.00    0.00    0.00    1.00
visualize a random policy
.       .       v       <
<       .       ^       v
v       ^       v       <
>       >       v       >
--- value iteration ---
Values from Value Iteration
0.00    0.00    3.62    3.88
2.19    0.00    3.97    4.39
3.62    3.97    4.63    5.48
3.88    4.39    5.48    6.24
Optimal Policy
.       .       >       >
<       .       >       >
v       v       v       >
v       v       v       v
--- policy evaluation ---
Optimal Policy
.       .       >       >
<       .       >       >
v       v       v       >
v       v       v       v
Values from Policy Iteration
0.00    0.00    3.62    3.88
2.19    0.00    3.97    4.39
3.62    3.97    4.63    5.48
3.88    4.39    5.48    6.24
```

As observed in the output, the optimal policy has changed significantly again. The agent avoids the top part of the environment, as it never chose the 'UP' action. Additionally, it selected the 'LEFT' action only once, while all other actions were 'RIGHT,' indicating an effort to stay away from the left side of the environment.

These changes in the optimal policy align with my expectations. By increasing the noise parameter, I anticipated that the agent would take a safer approach and avoid areas with negative rewards. Even though both positive rewards were +1, the bottom-right side of the environment was more attractive to the agent since it contained no negative rewards nearby.