Reinforcement Learning - Project 2

Fridolin Paiki

March 4, 2024

1 Problem 1

Given:
Inital policy $\pi^0(S) = Left$ and $V_0(S) = 0$ for all $S \in S$. The map for state ID can be seen in Figure 1

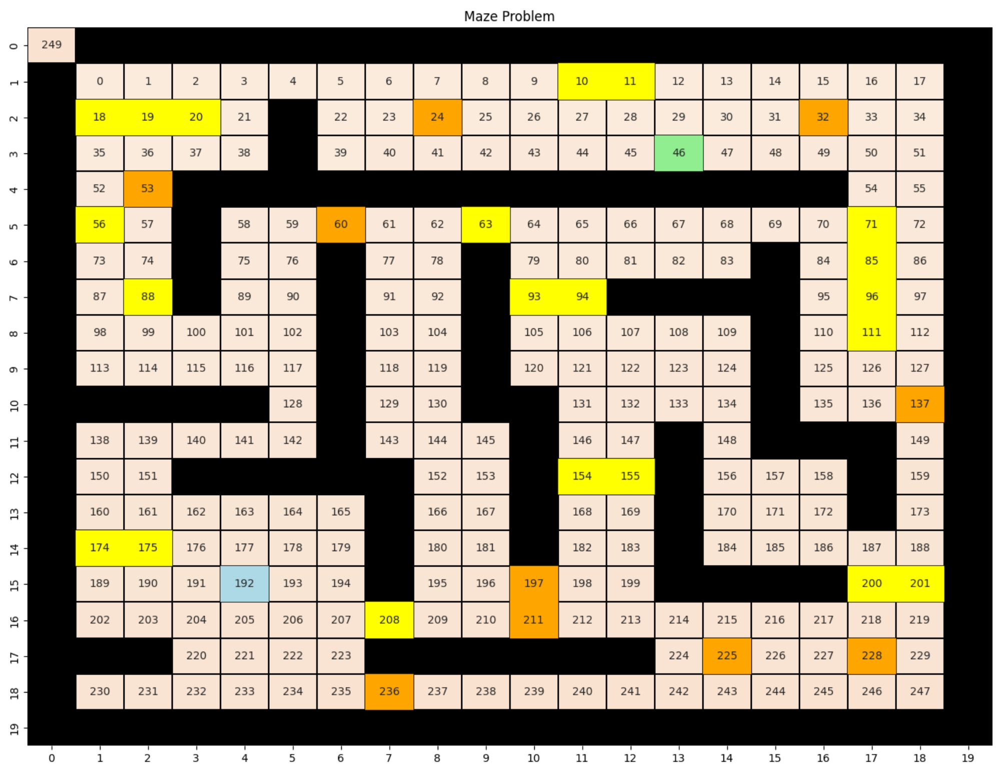

Figure 1: State IDs for all states

1.1 Part 1 - Policy Iteration

1.1.1 Base Scenario

$$p = 0.02, \gamma = 0.95, \theta = 0.01$$

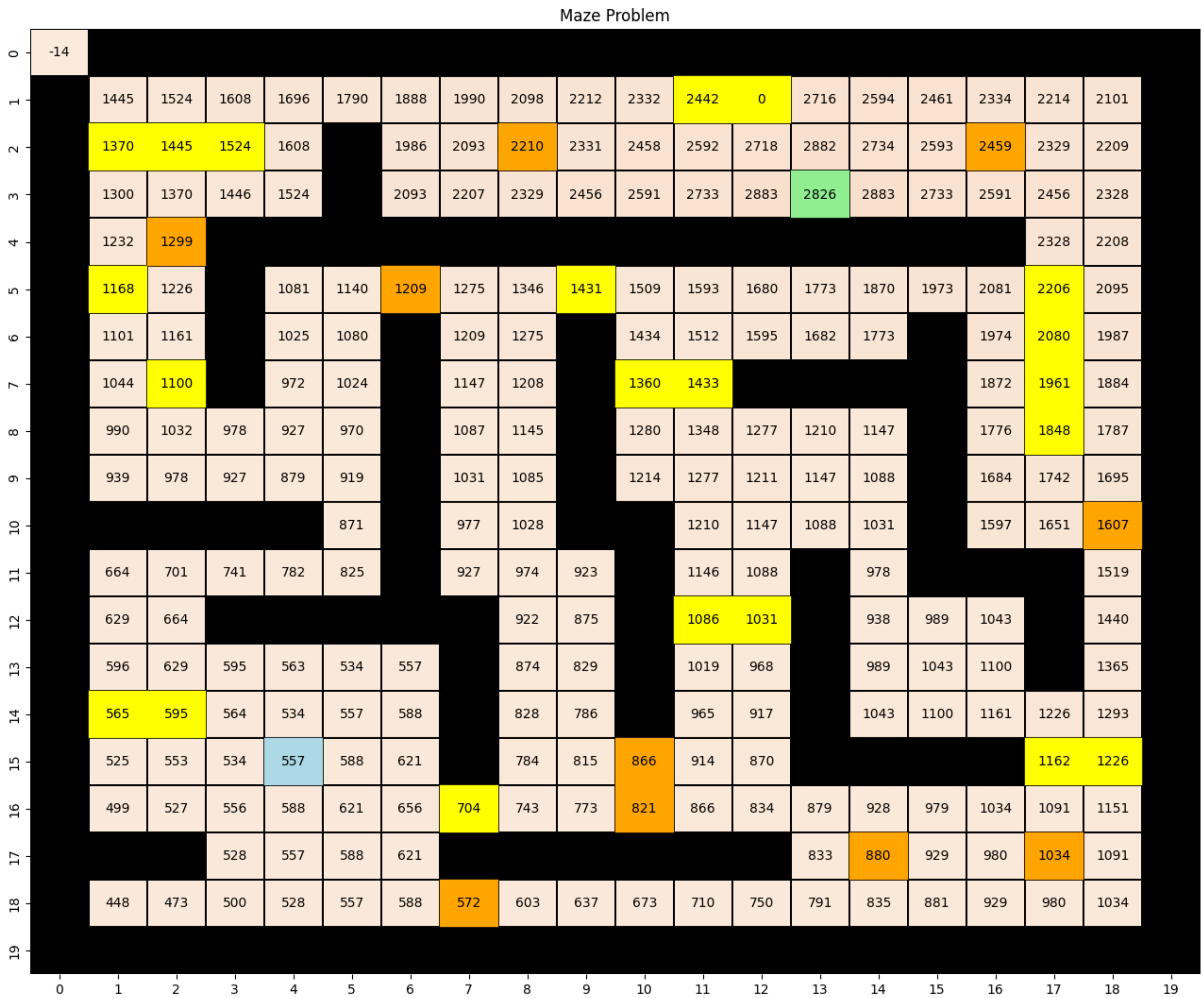

Figure 2: The Optimal Value Function values at all states for the Base Scenario under the Policy Iteration method

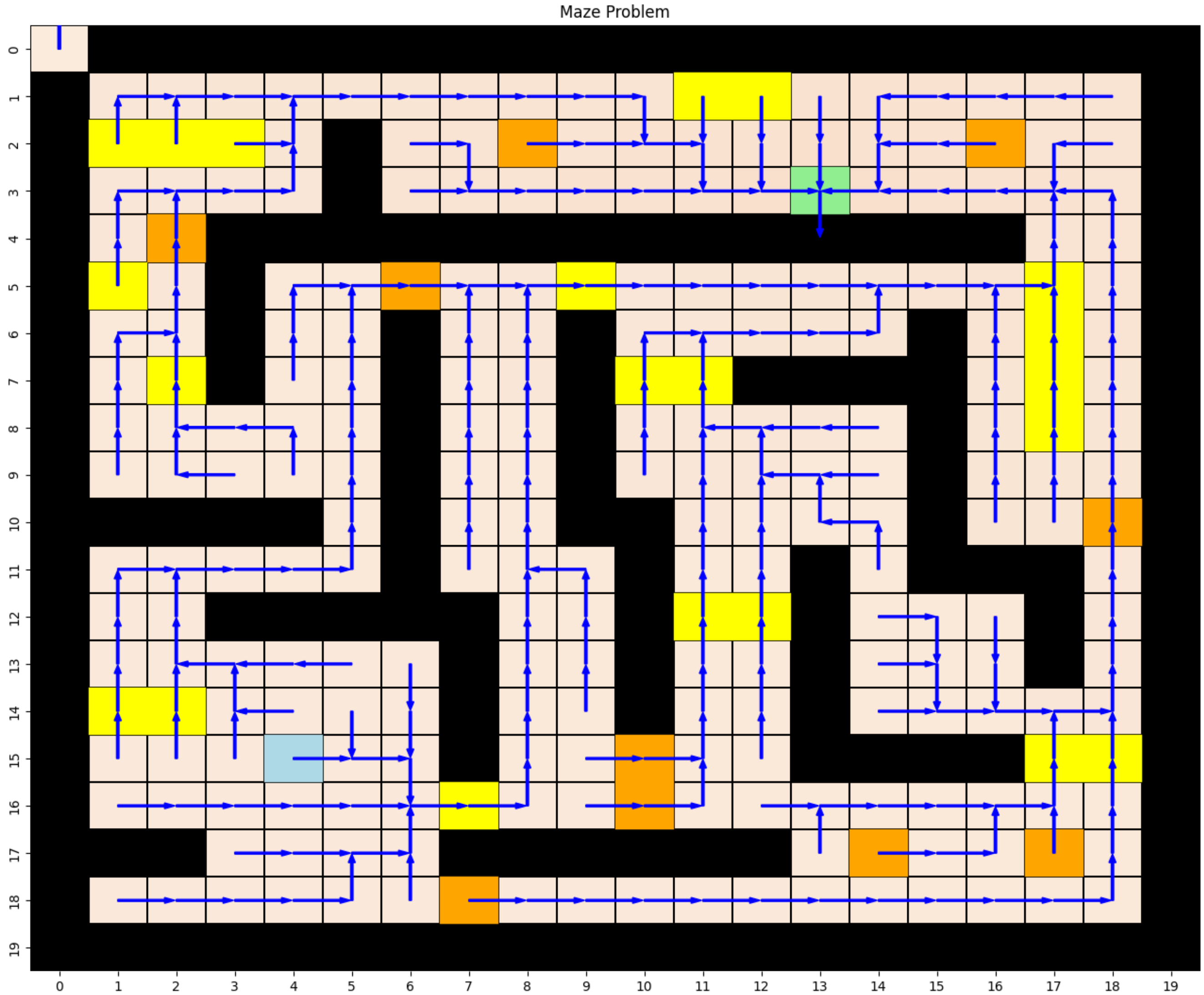

Figure 3: The Optimal Policy values at all states for the Base Scenario under the Policy Iteration method

1.1.2 Large Stochasticity Scenario

$$p = 0.5, \gamma = 0.95, \theta = 0.01$$

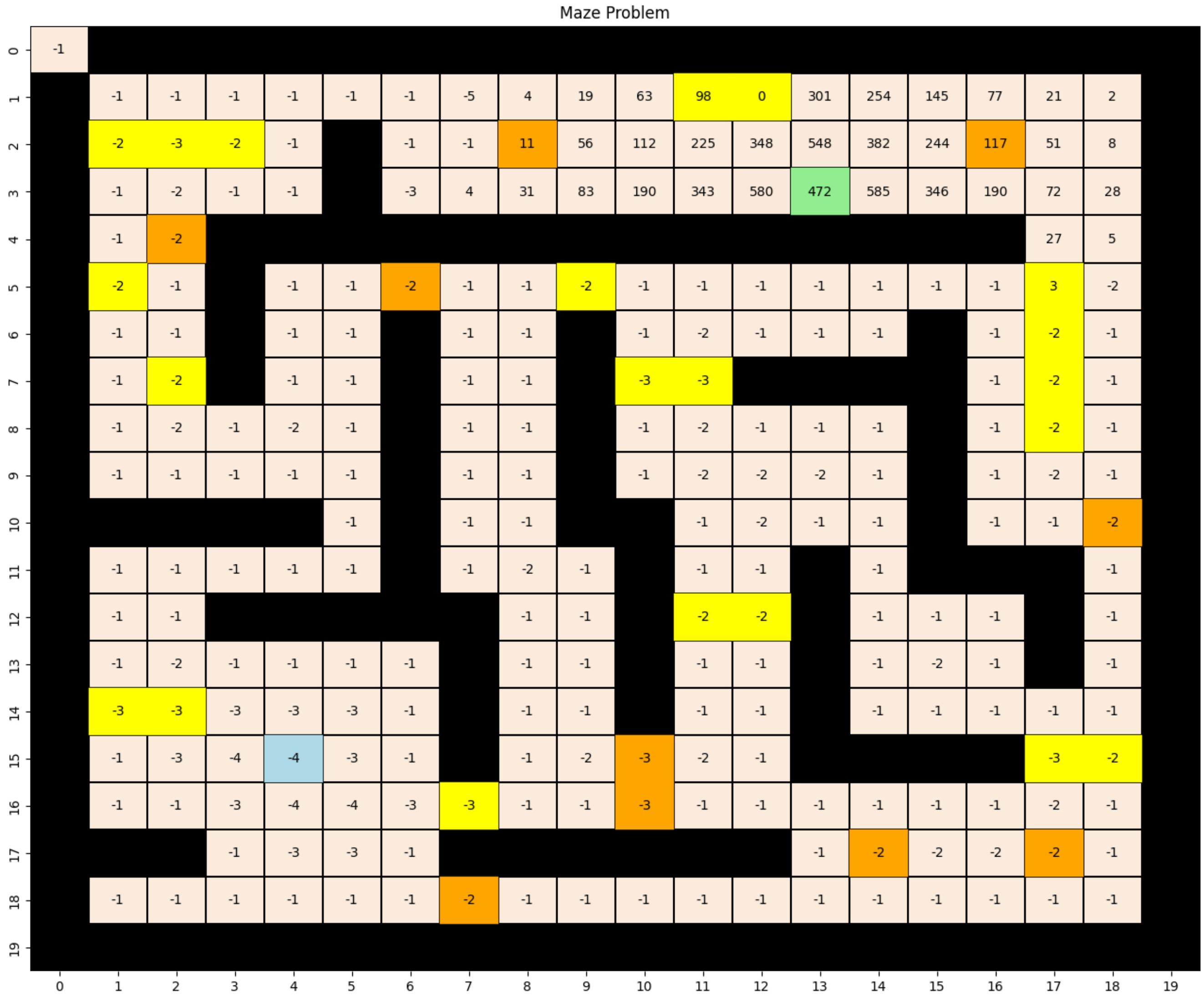

Figure 4: The Optimal Value Function values at all states for the Large Stochastic Scenario under the Policy Iteration method

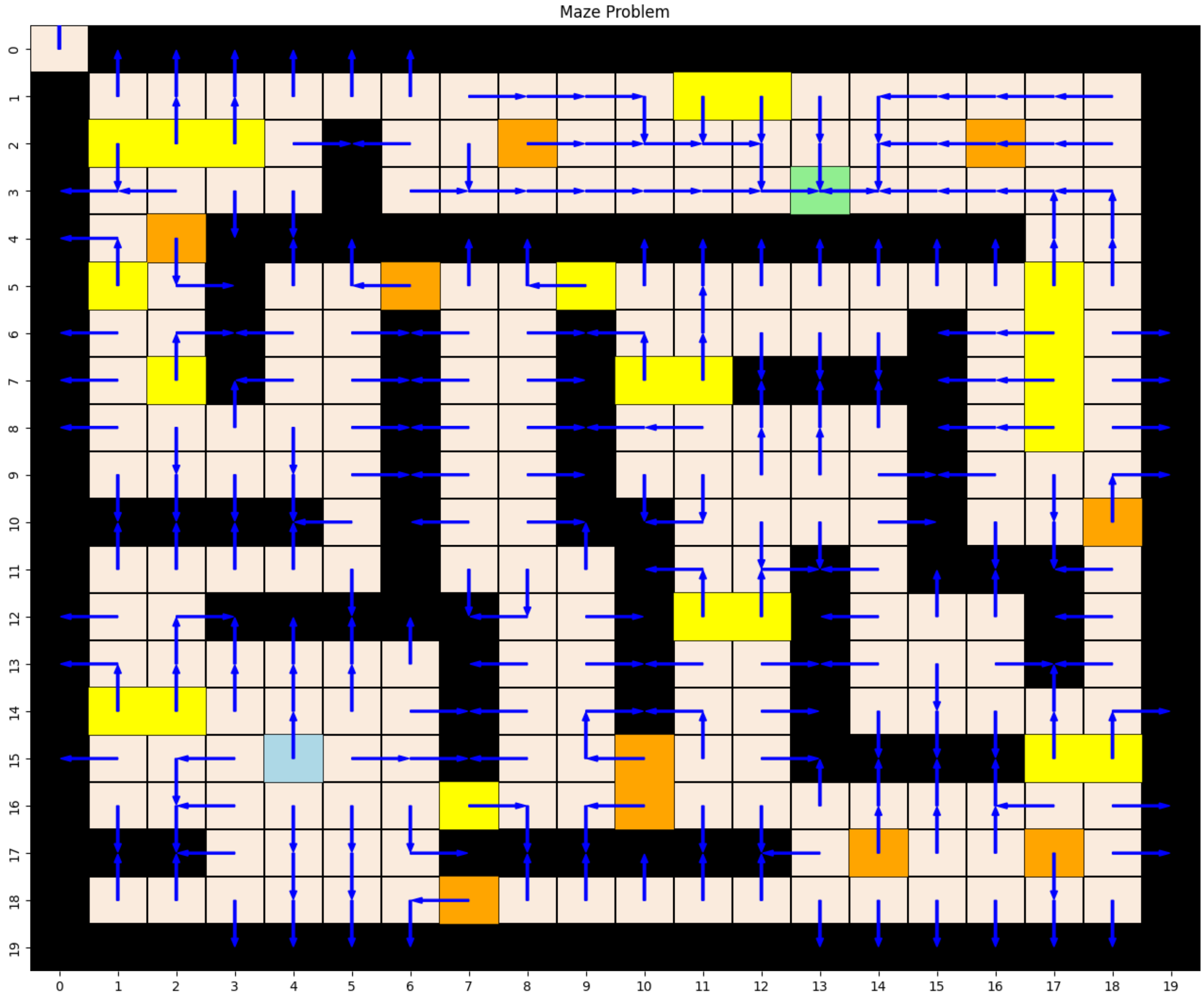

Figure 5: The Optimal Policy values at all states for the Large Stochastic Scenario under the Policy Iteration method

1.1.3 Small Discount Factor Scenario

$$p = 0.02, \gamma = 0.55, \theta = 0.01$$

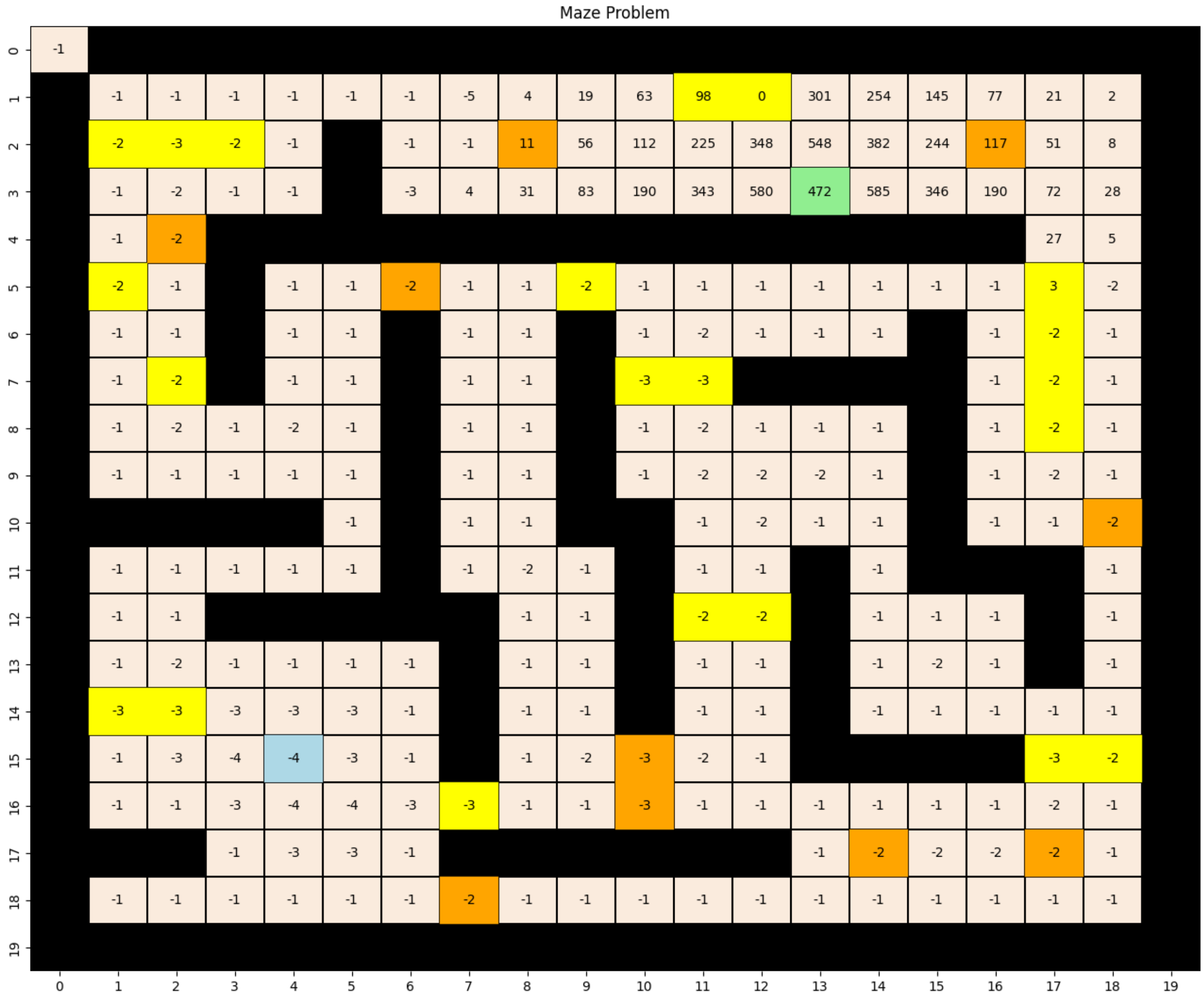

Figure 6: The Optimal Value Function values at all states for the Small Discount Factor Scenario under the Policy Iteration method

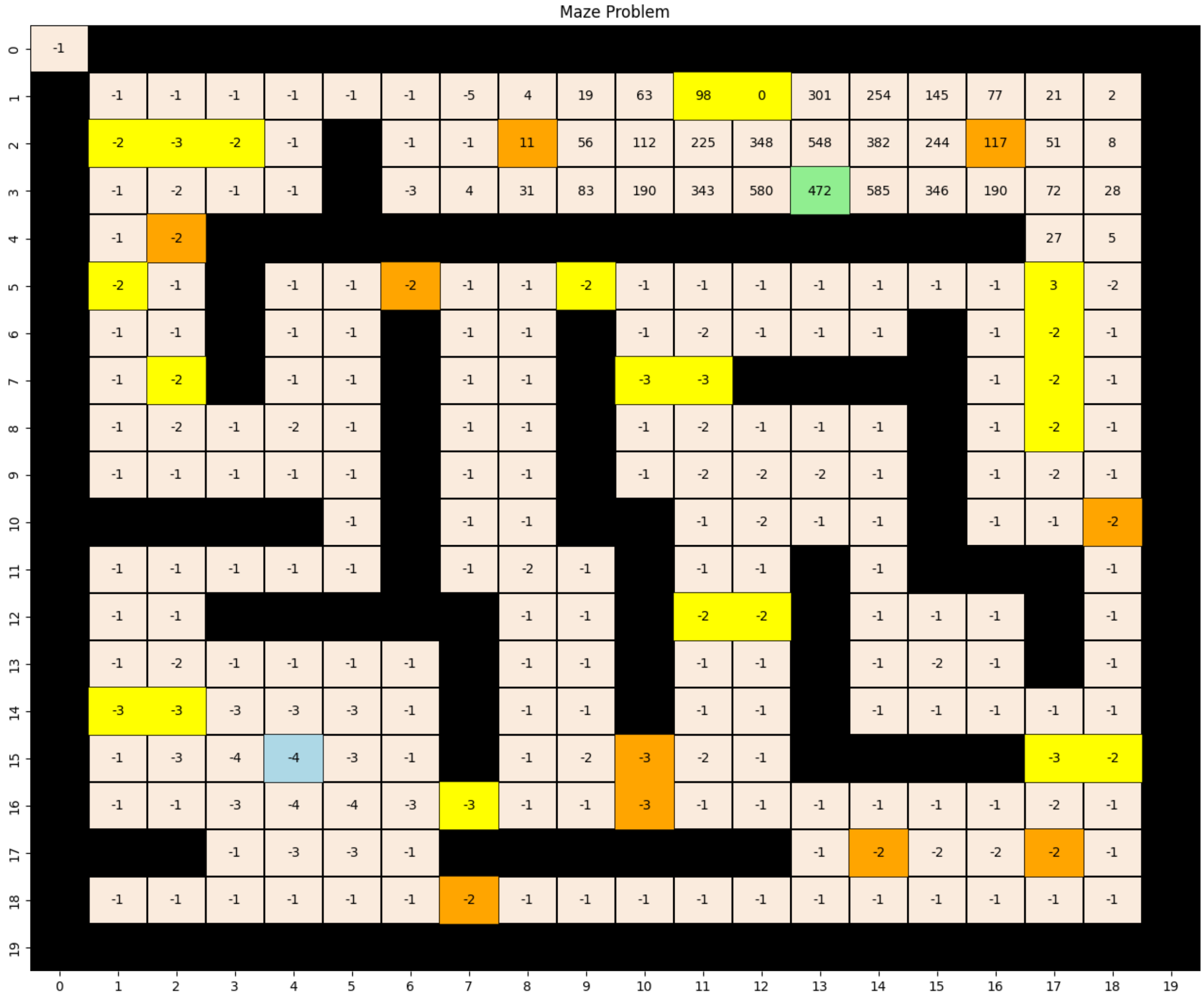

Figure 7: The Optimal Policy values at all states for the Small Discount Factor Scenario under the Policy Iteration method

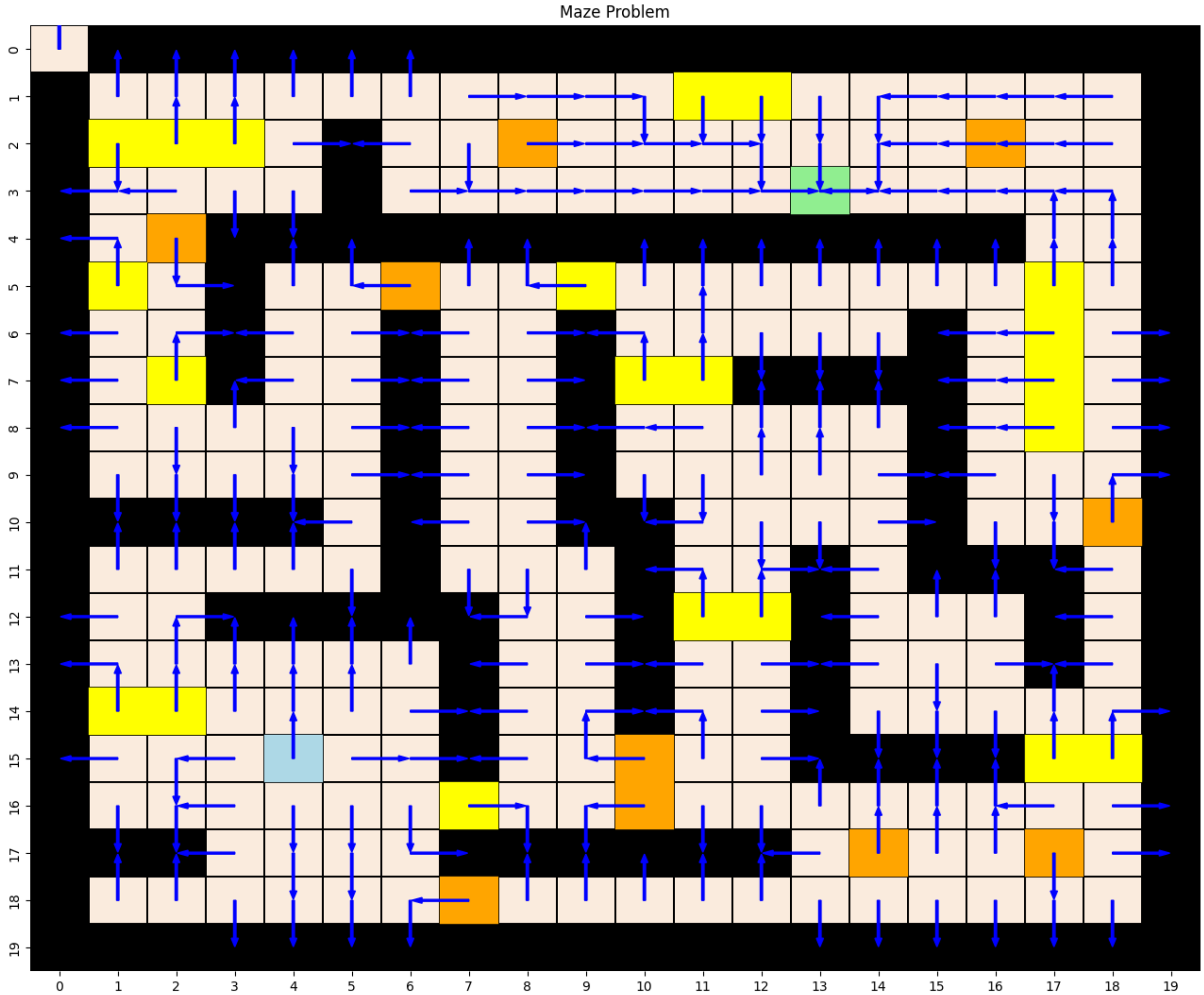

Figure 8: Enter Caption

1.2 Part 2 - Value Iteration

Base Scenario

$$p = 0.02, \gamma = 0.95, \theta = 0.01$$

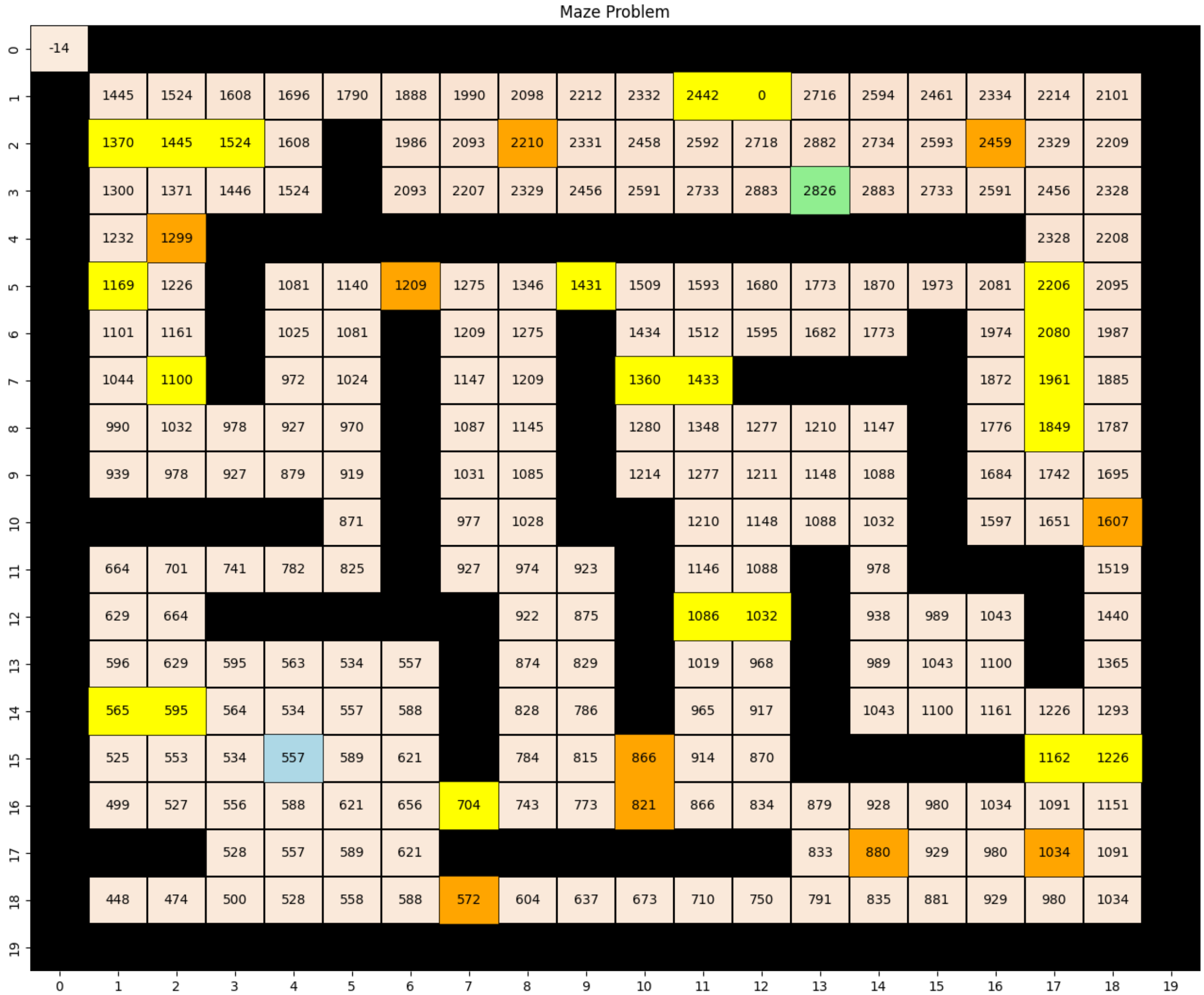

Figure 9: The Optimal Value Function values at all states for the Base Scenario under the Value Iteration method

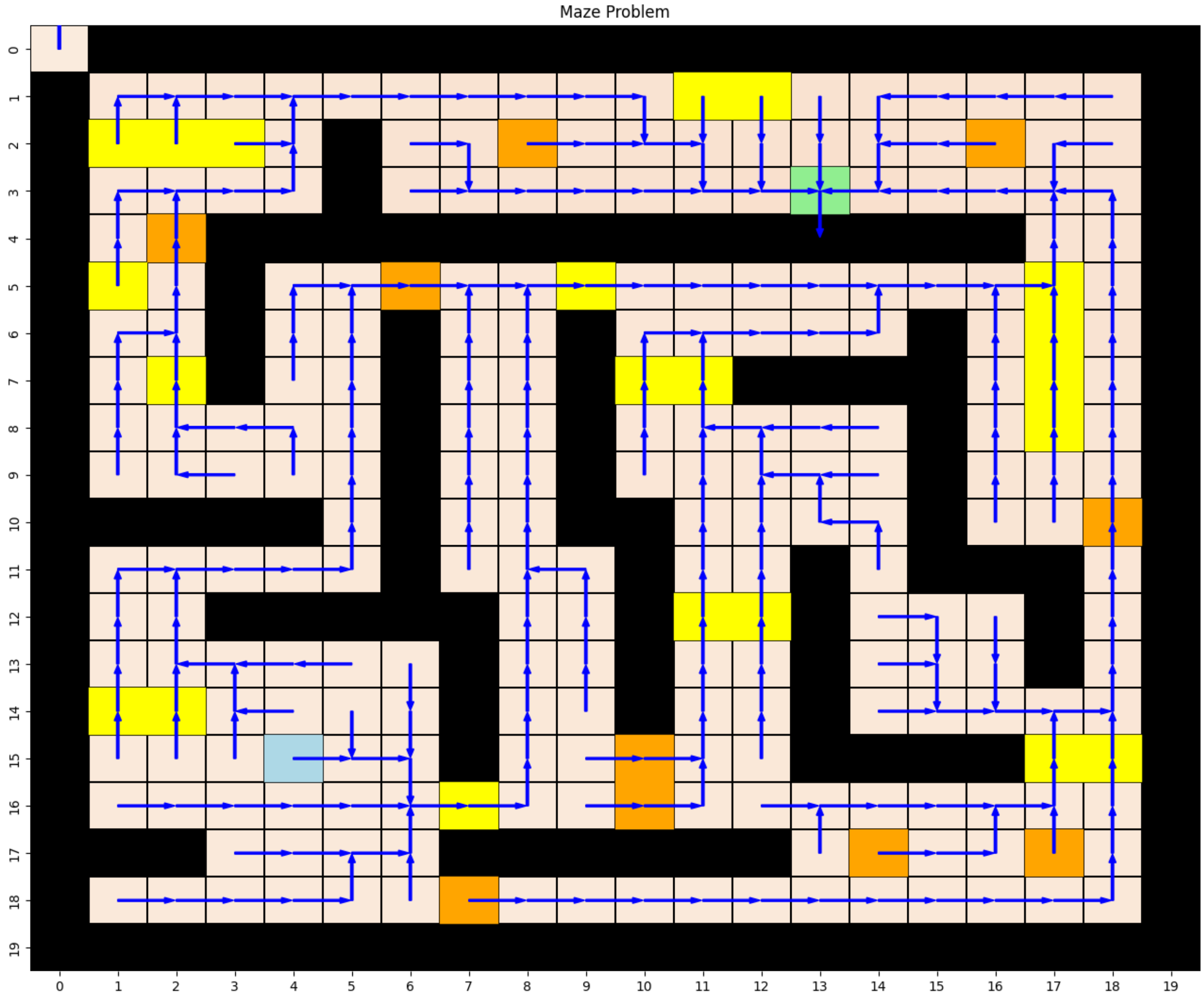

Figure 10: The Optimal Policy values at all states for the Base Scenario under the Value Iteration method

1.2.1 Large Stochasticity Scenario

$$p = 0.5, \gamma = 0.95, \theta = 0.01$$

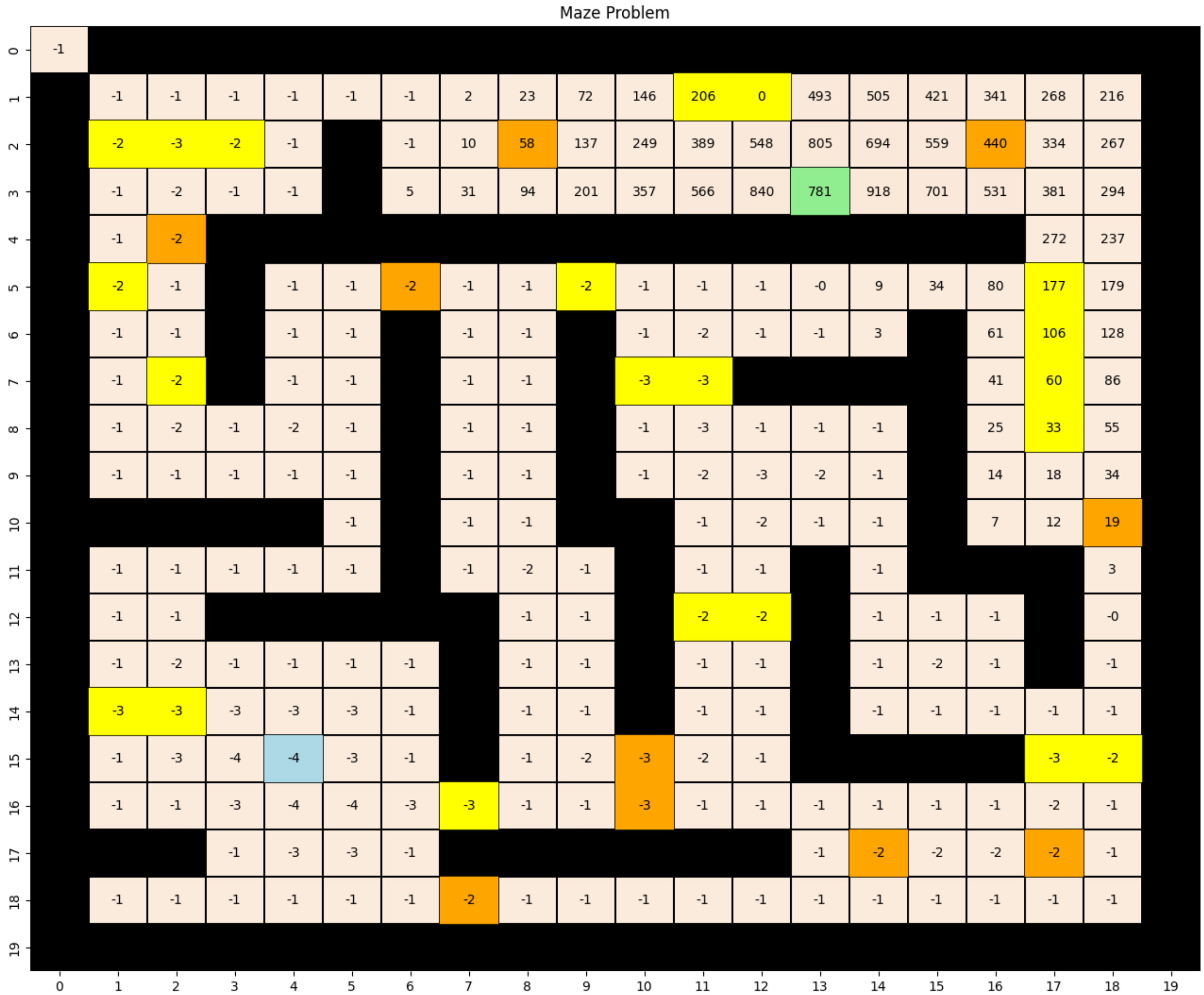

Figure 11: The Optimal Value Function values at all states for the Large Stochastic Scenario under the Value Iteration method

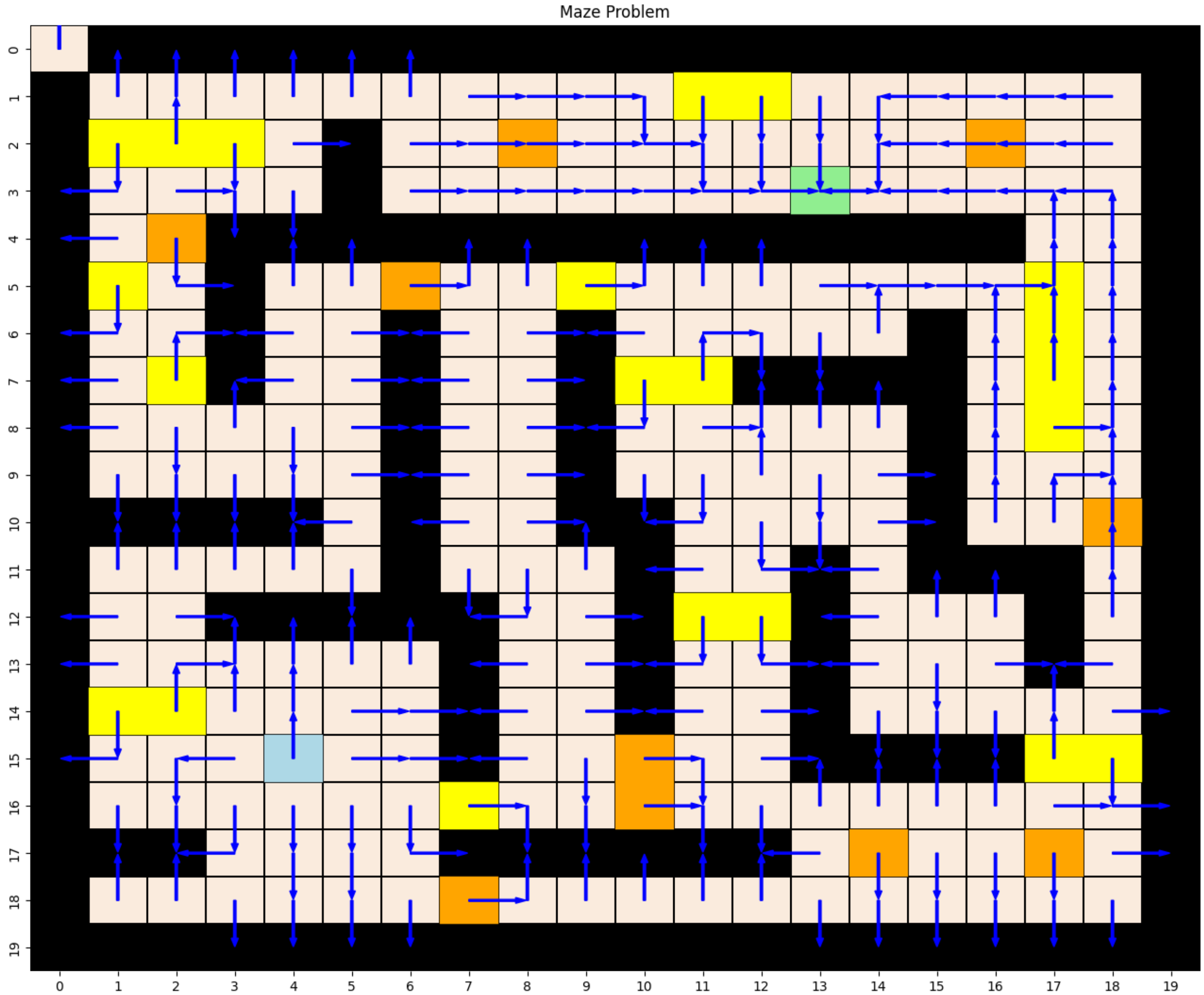

Figure 12: The Optimal Policy values at all states for the Large Stochastic Scenario under the Value Iteration method

1.2.2 Small Discount Factor Scenario

$$p = 0.02, \gamma = 0.55, \theta = 0.01$$

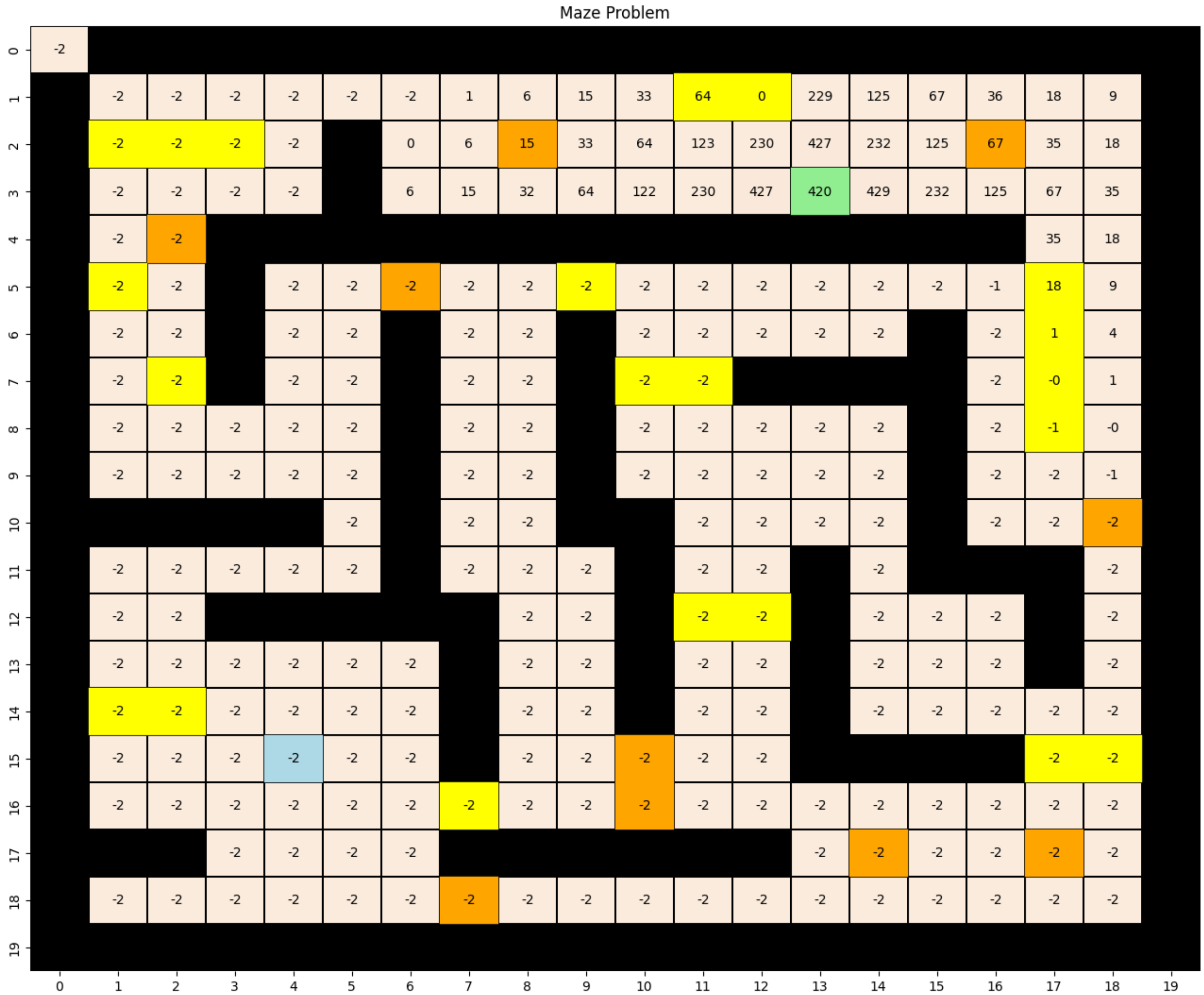

Figure 13: The Optimal Value Function values at all states for the Small Discount Factor Scenario under the Value Iteration method

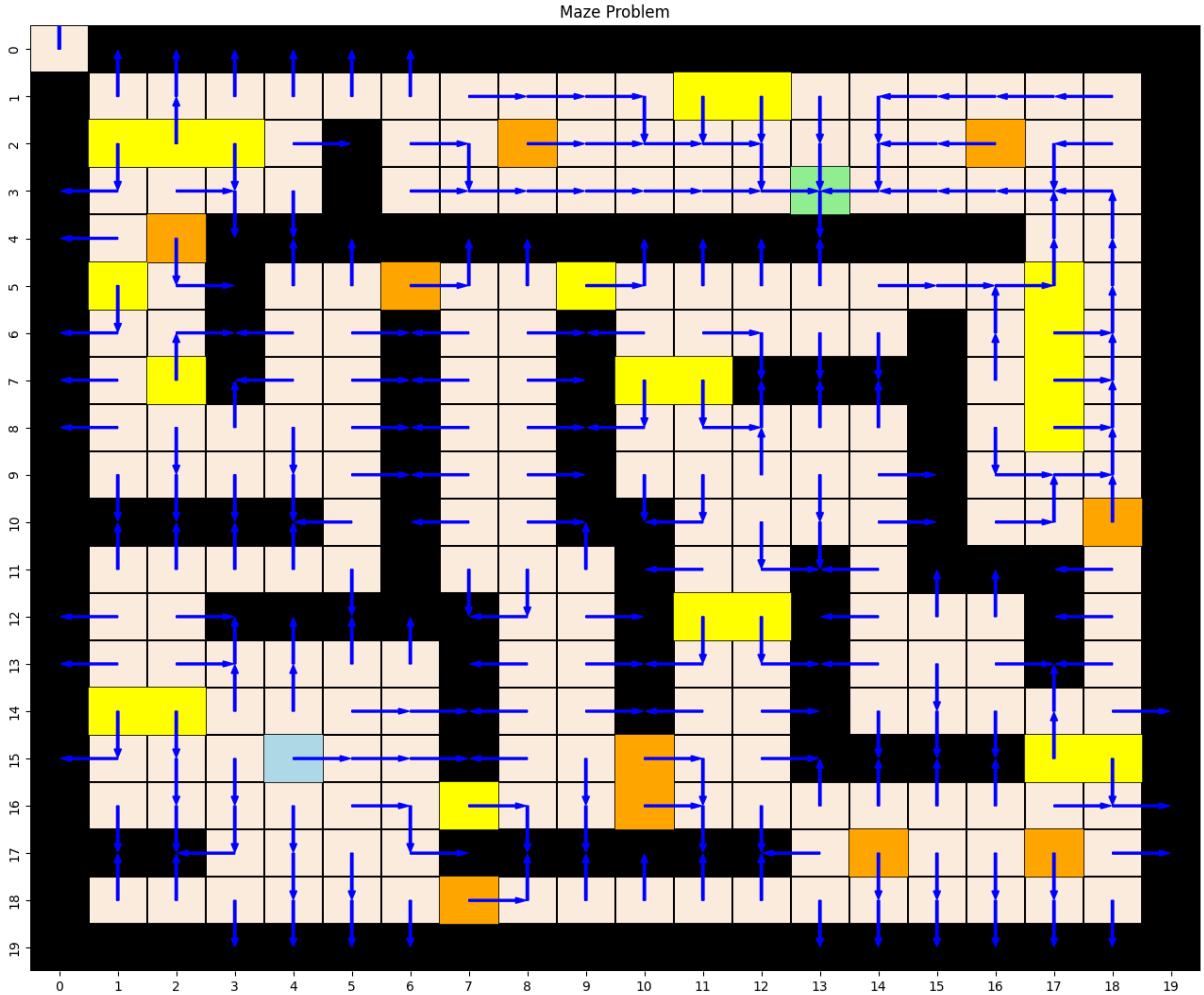

Figure 14: The Optimal Policy values at all states for the Small Discount Factor Scenario under the Value Iteration method

Just like in the policy iteration case, the agent in the ideal scenario finds the fastest route to the goal with a high overall value. However, when randomness (stochasticity) is introduced, the agent might deviate from the optimal policy, leading to a longer path. In the case of a very small discount factor, the agent prioritizes immediate rewards so much that it might never reach the goal at all, getting stuck chasing short-term gains.

2 Problem 2

Given $2^4 = 16$ states and 5 possible actions in action space $a \in A = \{a^1 = [0,0,0,0]^T, a^2 = [1,0,0,0]^T, a^3 = [0,1,0,0]^T, a^4 = [0,0,1,0]^T, a^5 = [0,0,0,1]^T\}$

2.1 Part a - Matrix-Form Value Iteration

$$p = 0.05, \gamma = 0.95, \theta = 0.01$$

Based on the calculation of matrix-form value iteration, I found that the optimal policy is:
2 2 2 2 2 2 2 2 3 2 2 2 4 2 2 2

The average activation genes AvgA are **0.48 for no-control policy** and **2.83 for the optimal policy**.

2.2 Part b - Matrix-Form Value Iteration

$$p = 0.2, \gamma = 0.95, \theta = 0.01$$

Based on the calculation of matrix-form value iteration, I found that the optimal policy is:
$\pi^* = [a^2, a^2, a^2, a^2, a^2, a^2, a^2, a^2, a^3, a^2, a^2, a^2, a^4, a^2, a^2, a^2]$

The average activation genes AvgA are **1.25 for no-control policy** and **2.41 for the optimal policy**.

$$p = 0.45, \gamma = 0.95, \theta = 0.01$$

Based on the calculation of matrix-form value iteration, I found that the optimal policy is:
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

The average activation genes AvgA are **1.92 for no-control policy** and **1.90 for the optimal policy**.

Basically, as the value iteration algorithm gets more random (the higher p value), it struggles to find the best reward. When the randomness parameter (p) is set to 0.45, the algorithm mostly takes random actions, which is no different from having no control strategy at all. In this case, it always picks action zero because it offers the most immediate reward. However, with less randomness, the algorithm can better balance exploring new options and exploiting what it already knows, ultimately resulting in a better strategy than simply doing nothing.

2.3 Part c - Matrix-Form Policy Iteration

$$p = 0.05, \gamma = 0.95, \theta = 0.01$$

Based on the calculation of matrix-form value iteration, I found that the optimal policy is:
2 2 2 2 2 2 2 2 3 2 2 2 4 2 2 2

The average activation genes AvgA are **0.48 for no-control policy** and **2.85 for the optimal policy**.

Both policy iteration and value iteration converge to the same optimal policy. Likewise, the average activation rate they achieve is very similar, with slight variations due to the inherent randomness of the problem. The key difference lies in the number of iterations required. Value iteration takes 137 iterations, while policy iteration takes 3 iterations. This efficiency boost in policy iteration comes from performing the policy evaluation step in a single iteration, essentially doing one big matrix calculation.