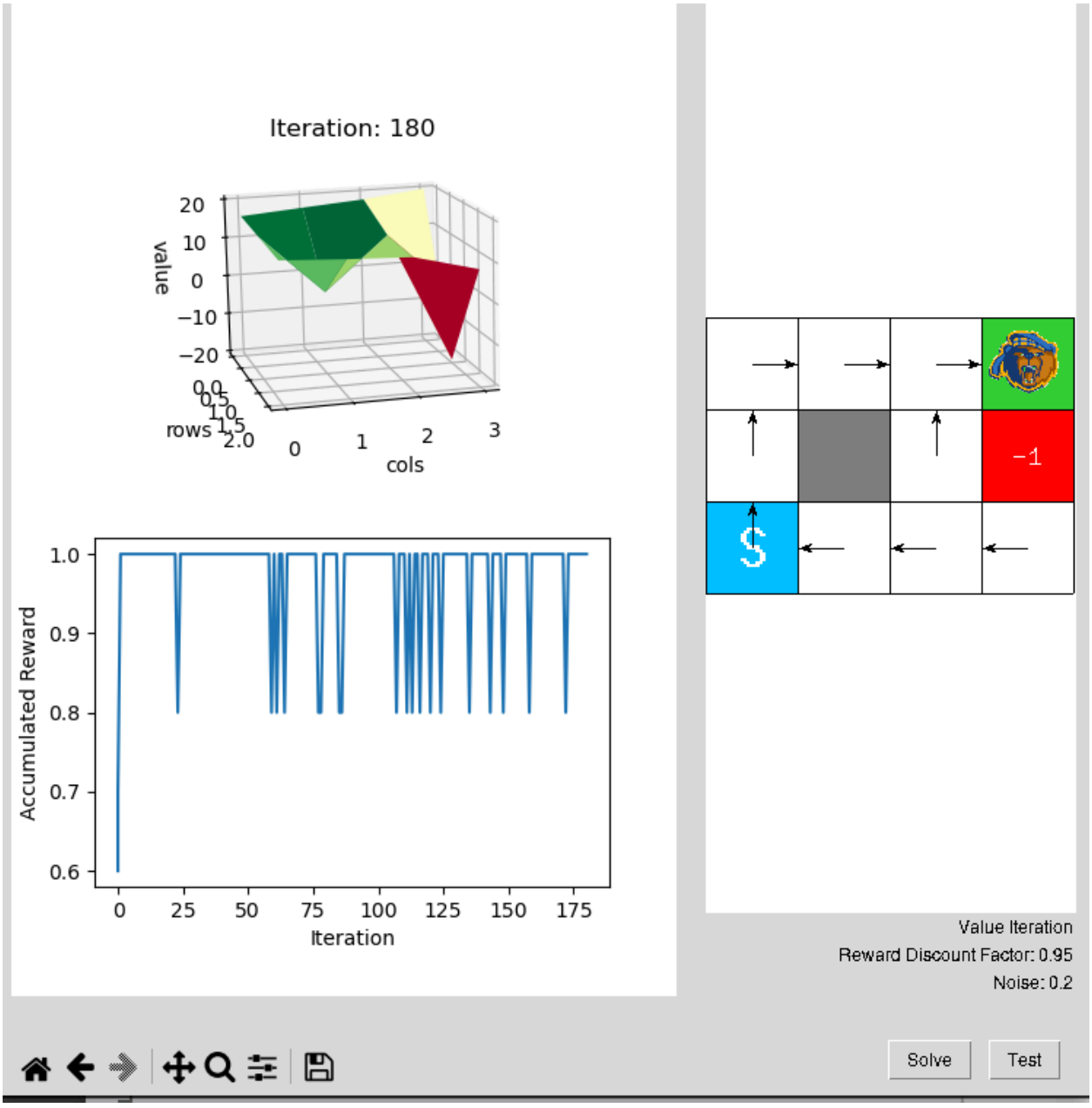
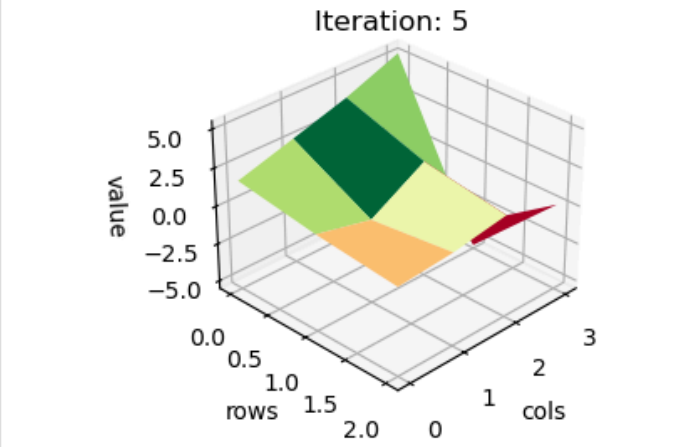
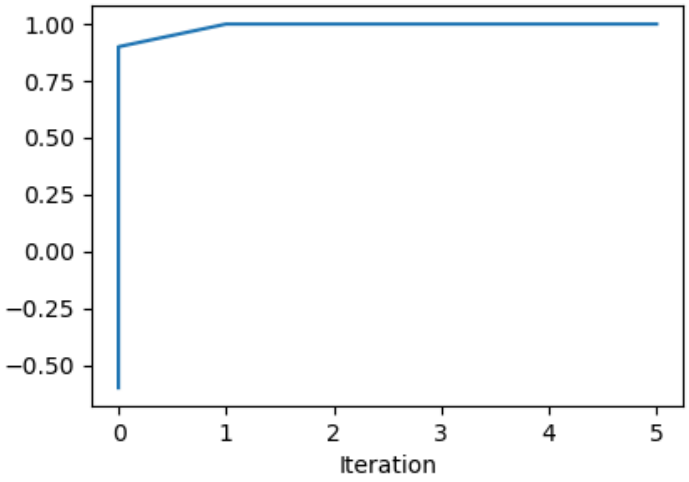
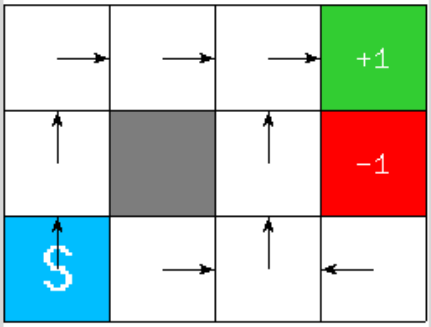

Iteration: 180
value
rows
cols
Accumulated Reward
Iteration
-1
S
Value Iteration
Reward Discount Factor: 0.95
Noise: 0.2
Solve
Test



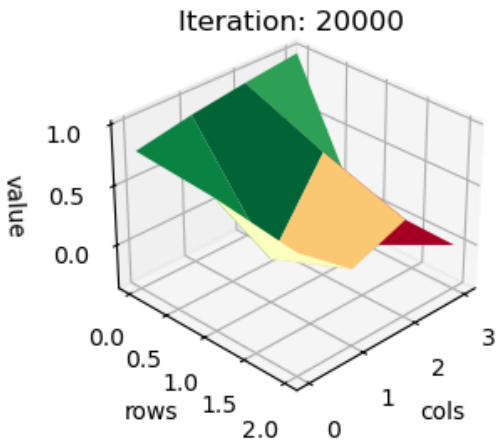


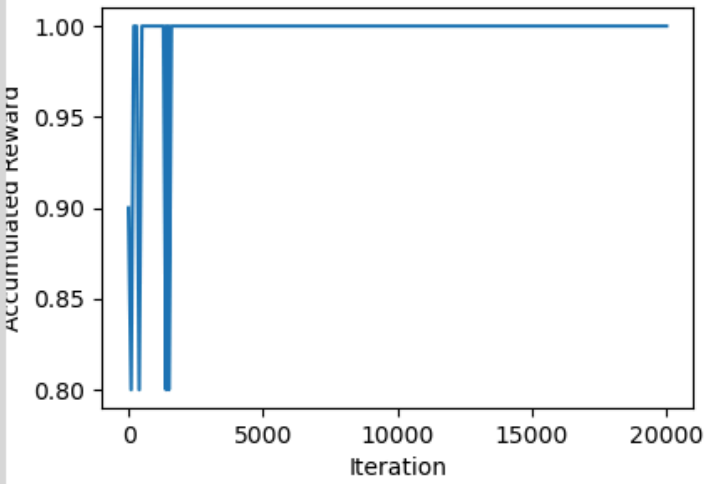


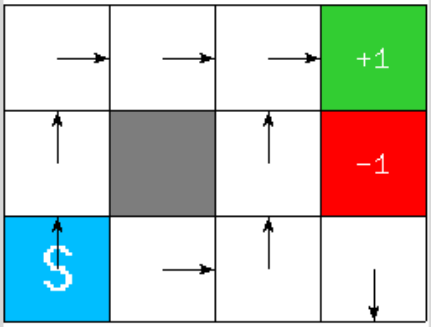
Policy Iteration
Reward Discount Factor: 0.95
Noise: 0.2

(x, y) = (4.82, 0.396)

Solve Test

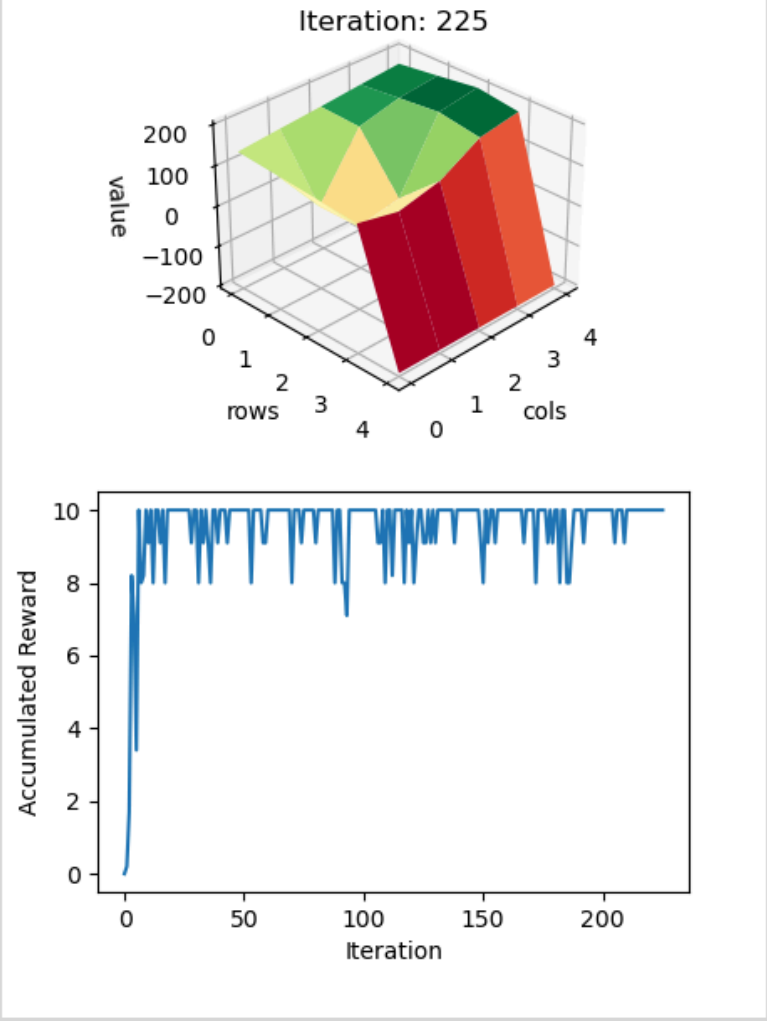


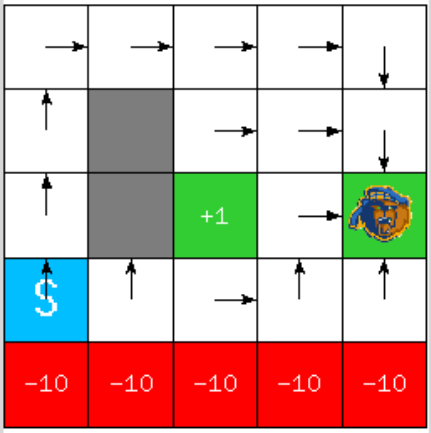




Q-Learning
Reward Discount Factor: 0.95
Noise: 0.2

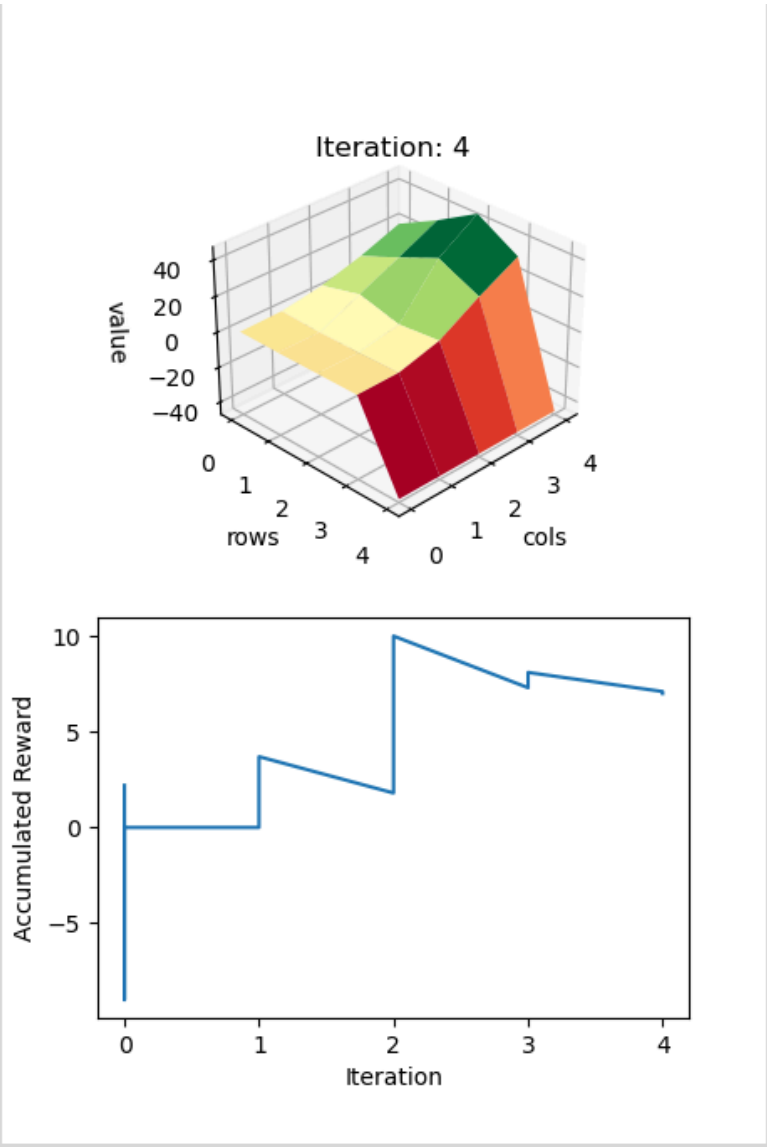
Solve Test

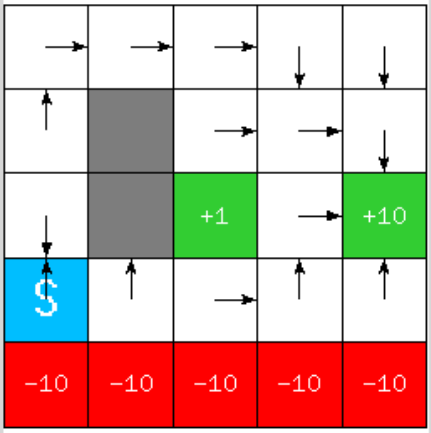




Value Iteration
Reward Discount Factor: 0.95
Noise: 0.2

Solve Test

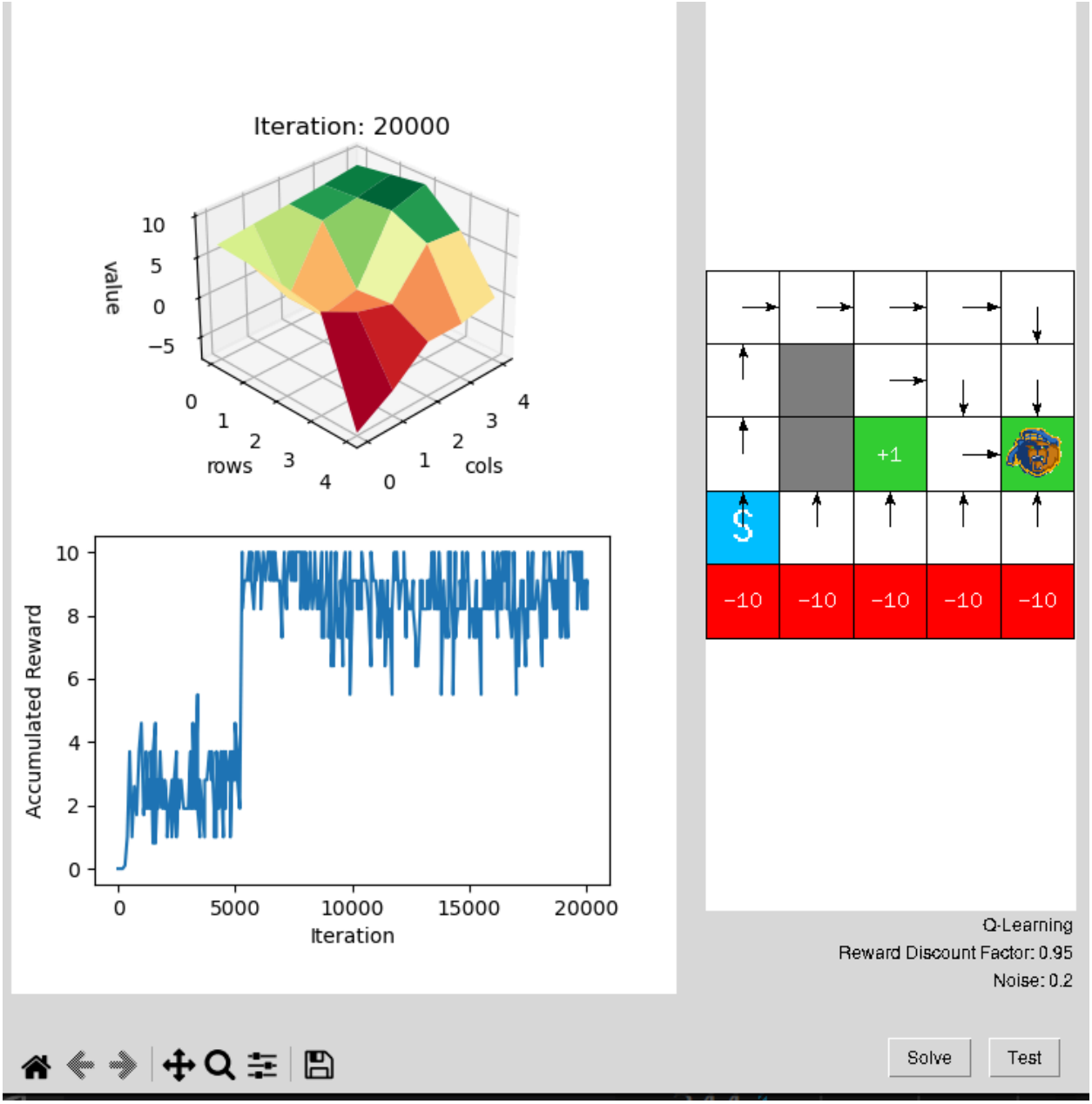




Policy Iteration
Reward Discount Factor: 0.95
Noise: 0.2

(x, y) = (3.83, 1.78)

Solve Test



Notes:

Value Iteration and Policy Iteration converge to the true optimal policy when given enough iterations because they're working with the complete model of the environment (transition probabilities and rewards).

Q-Learning will converge to the optimal if:

- Every state-action pair is visited infinitely often
- The learning rate decreases appropriately over time

- The Markov property holds for the environment

For world 1

Why policies are similar:
- All three algorithms aim to find the same optimal solution
- Small environment with clear optimal path
- Limited number of possible policies

Why Q-Learning differs slightly:
- Learns from samples instead of using the complete model
- Exploration may not visit all states equally
- Randomness in exploration affects learning
- Might take different actions in states rarely visited

For world3, the grid is much more complex and the rewards and penalties are much more pronounced.

Differences for world 3:
- Q-Learning might find slightly different paths to the +10 reward
- States near penalty areas might show different policies
- Rarely visited states will have the most variation
- Q-Learning might be more conservative near dangerous areas

Convergence speed:
- Value/Policy Iteration: Fast (tens to hundreds of iterations)
- Q-Learning: Much slower (thousands of iterations)