Multi-Agent Grid Coverage Using Q-Learning

By

Sushovit Nanda

21283

Dept. of Electrical Engineering and Computer Science (EECS)

Indian Institute of Science Education and Research Bhopal, Bhopal

Under the Guidance of

Dr. Sujit Pedda Baliyarasimhuni

PROFESSOR

Dept. of Electrical Engineering and Computer Science (EECS)

Indian Institute of Science Education and Research Bhopal, Bhopal

Abstract

Grid coverage by autonomous agents is a critical challenge in robotics and artificial intelligence, with significant applications in search and rescue, agricultural monitoring, and surveillance. While much progress has been made in pathfinding algorithms like A* and Dijkstra's for single-agent navigation, relatively little focus has been placed on enabling multi-agent systems to perform grid mapping in dynamic environments collaboratively. This project addresses this gap by developing a multi-agent reinforcement learning (MARL) framework to achieve optimal grid coverage using Q-learning. The proposed system enables agents to navigate and map a shared 10x10 grid with static obstacles, ensuring that every cell is visited while avoiding collisions and minimizing revisits. The results demonstrate the effectiveness of reinforcement learning in enabling collaborative behavior among agents and highlight the challenges posed by obstacle-laden environments. These findings set the stage for deploying multi-agent systems in real-world scenarios where adaptive, efficient exploration is crucial.

1. Problem Statement

The objective of this project is to design a solution where multiple autonomous agents collaboratively operate in a 10x10 grid to:

1. Navigate dynamic environments with randomly placed obstacles.
2. Ensure every grid cell is visited at least once.
3. Minimize overlaps, collisions, and redundant exploration.
4. Learn efficient navigation policies using reinforcement learning.

This problem presents challenges due to the need for implicit coordination between agents in a shared environment without direct communication.

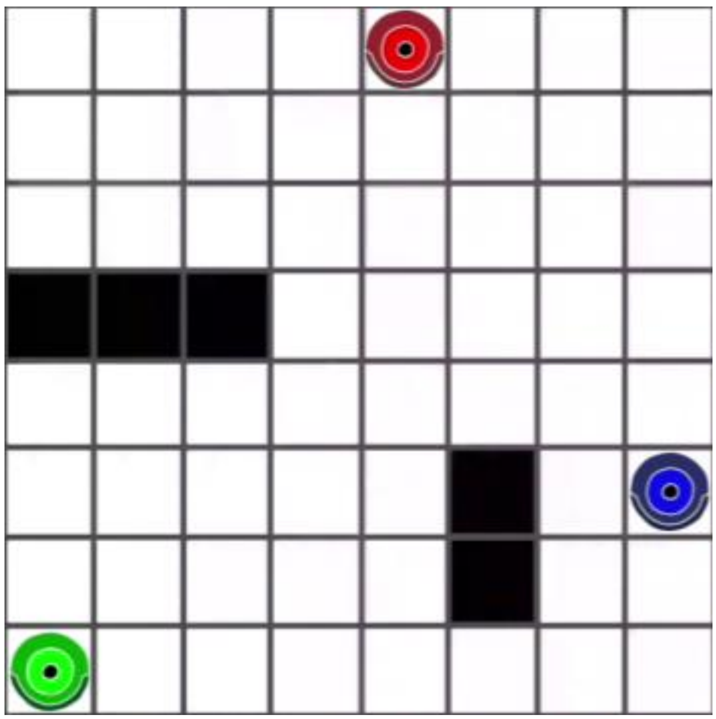

Fig. 1 : An example 8x8 environment with 3 agents and obstacles

2. Approach

2.1 Problem Formulation

The grid coverage problem is framed as a Markov Decision Process (MDP), defined by the following components:

- States (S): Each agent's position on the grid (x, y), combined with the positions of obstacles and visited cells.
- Actions (A): Movement in one of four cardinal directions (up, down, left, right).

- Transition Function (T): Defines the next state based on the agent's action and grid constraints (e.g., walls, obstacles).
- Reward Function (R):
  - New Cell: +10/(distance to start + 1)
  - Revisited Cell: -1
  - Obstacle Collision: -1
  - Completion Bonus: +50

2.2 Q-Learning Algorithm

The agents use Q-Learning, a reinforcement learning algorithm, to iteratively learn an optimal policy for grid coverage. The algorithm updates Q-values for each state-action pair based on the following Temporal Difference (TD) formula:

$$Q(s, a) \leftarrow Q(s, a) + \alpha [R + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

Where:
- $Q(s, a)$: The Q-value of state s and action a.
- α: Learning rate.
- γ: Discount factor.
- R: Immediate reward for taking action a in state s.
- $\max_{a'} Q(s', a')$: Maximum future reward from state s'.

2.3 Epsilon-Greedy Exploration

To balance exploration and exploitation, the agents use an epsilon-greedy policy:
- With probability ε: Select a random action (exploration).
- With probability $1 - \varepsilon$: Select the action with the highest Q-value (exploitation).

3. Experimentation

3.1 Experimental Setup

1. Grid Configuration:
- Size: 10x10 grid.
- Obstacles: 10 randomly placed static obstacles.
- Agents: 3 agents with fixed starting positions: (0,0), (9,0), (9,9).

2. Rewards:
- Positive reward for visiting new cells.
- Negative penalties for revisits and obstacle collisions.
- Bonus reward for complete grid coverage.

3. Steps per Episode:
- Each episode is limited to 100 steps to encourage efficient exploration.

4. Agent Communication:
- No explicit communication. Agents coordinate implicitly by observing the shared environment.

5. Scenarios:
- Without Obstacles: Grid with no obstructions.
- With Obstacles: Grid containing static obstacles.

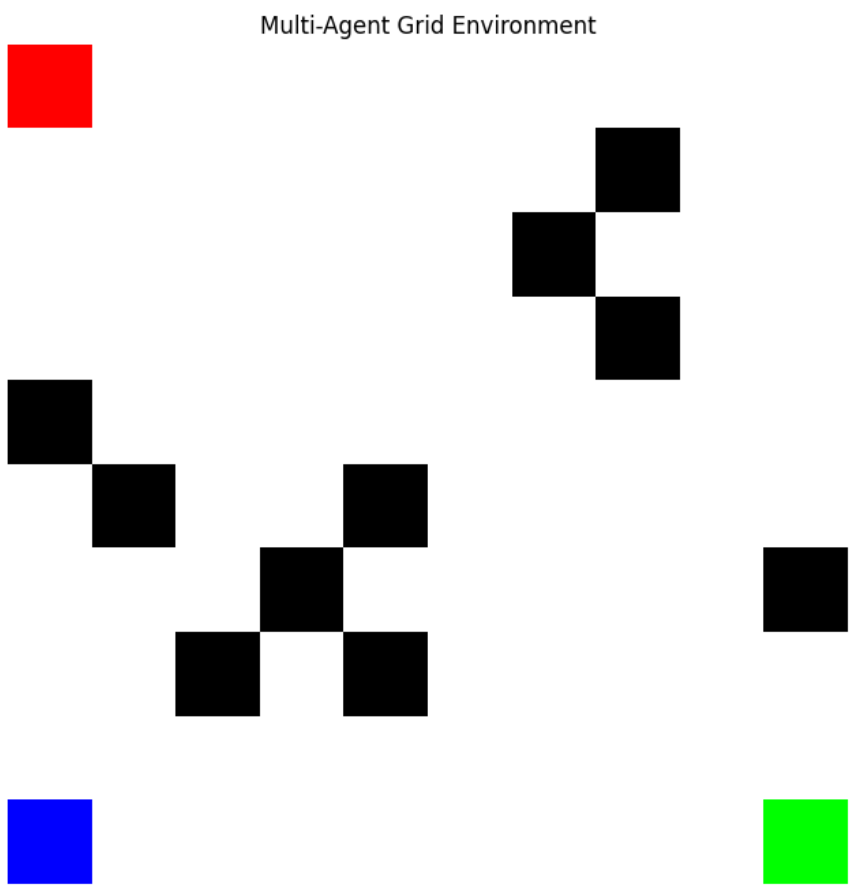

Fig. 2 : The 10x10 environment with 3 agents and 10 obstacles

4. Results

4.1 Learning Trends

1. Cumulative Rewards:
- Agents achieved higher cumulative rewards over episodes, indicating effective learning and policy optimization.

- Without Obstacles: Faster convergence to optimal policies.
- With Obstacles: Slower convergence due to penalty zones around obstacles.

```
Episode 50/2000, Total Reward: 27.464107494743022
Episode 100/2000, Total Reward: -7.621635271632205
Episode 150/2000, Total Reward: 19.928954051501442
Episode 200/2000, Total Reward: -1.3538826600300666
Episode 250/2000, Total Reward: -14.618023951294504
Episode 300/2000, Total Reward: 40.25254033104653
Episode 350/2000, Total Reward: 21.670723254195337
Episode 400/2000, Total Reward: 23.445669649173215
Episode 450/2000, Total Reward: 31.200729629345318
Episode 500/2000, Total Reward: 12.95585931010099
Episode 550/2000, Total Reward: 41.54253347441003
Episode 600/2000, Total Reward: 15.542593554422766
Episode 650/2000, Total Reward: 23.491401336821227
Episode 700/2000, Total Reward: 42.735695009851305
Episode 750/2000, Total Reward: 23.49140133682122
Episode 800/2000, Total Reward: 44.28991649170088
Episode 850/2000, Total Reward: 43.80486483084135
Episode 900/2000, Total Reward: 43.804864830841325
Episode 950/2000, Total Reward: 43.804864830841325
Episode 1000/2000, Total Reward: 43.804864830841325
Episode 1050/2000, Total Reward: 43.804864830841325
Episode 1100/2000, Total Reward: 43.804864830841325
Episode 1150/2000, Total Reward: 43.804864830841325
Episode 1200/2000, Total Reward: 43.804864830841325
Episode 1250/2000, Total Reward: 43.804864830841325
...
Episode 1850/2000, Total Reward: 43.804864830841325
Episode 1900/2000, Total Reward: 43.804864830841325
Episode 1950/2000, Total Reward: 43.804864830841325
Episode 2000/2000, Total Reward: 43.804864830841325
```

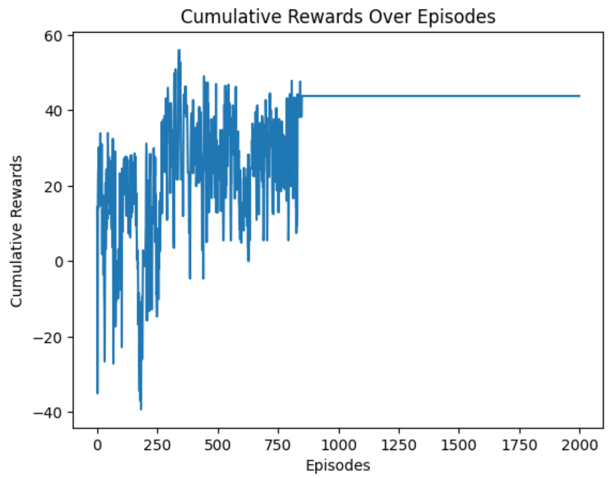

Fig. 3: Rewards Trend for Grid with Obstacles

```
Episode 50/2000 completed with cumulative reward: 1680
Episode 100/2000 completed with cumulative reward: 1320
Episode 150/2000 completed with cumulative reward: 780
Episode 200/2000 completed with cumulative reward: 1620
Episode 250/2000 completed with cumulative reward: 1860
Episode 300/2000 completed with cumulative reward: 1200
Episode 350/2000 completed with cumulative reward: 1860
Episode 400/2000 completed with cumulative reward: 1740
Episode 450/2000 completed with cumulative reward: 1320
Episode 500/2000 completed with cumulative reward: 1980
Episode 550/2000 completed with cumulative reward: 1620
Episode 600/2000 completed with cumulative reward: 1560
Episode 650/2000 completed with cumulative reward: 1560
Episode 700/2000 completed with cumulative reward: 1500
Episode 750/2000 completed with cumulative reward: 1080
Episode 800/2000 completed with cumulative reward: 840
Episode 850/2000 completed with cumulative reward: 960
Episode 900/2000 completed with cumulative reward: 1200
Episode 950/2000 completed with cumulative reward: 1260
Episode 1000/2000 completed with cumulative reward: 1980
Episode 1050/2000 completed with cumulative reward: 2580
Episode 1100/2000 completed with cumulative reward: 1980
Episode 1150/2000 completed with cumulative reward: 2520
Episode 1200/2000 completed with cumulative reward: 2460
Episode 1250/2000 completed with cumulative reward: 2100
```

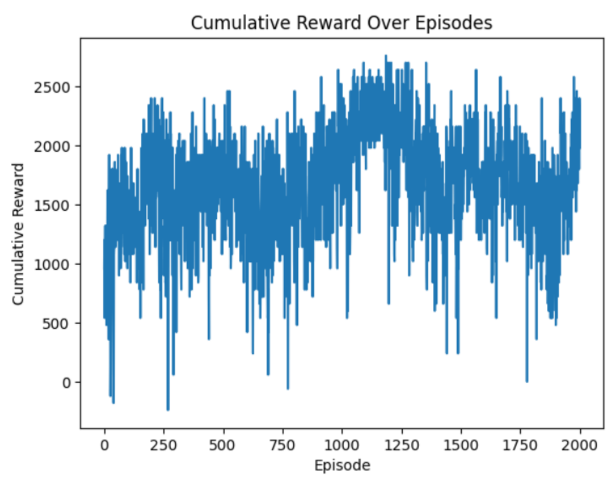

Fig. 4: Rewards Trend for Grid without Obstacles

2. Temporal Difference (TD) Errors:

- TD errors decreased over episodes, reflecting convergence of Q-values as agents learned optimal actions.

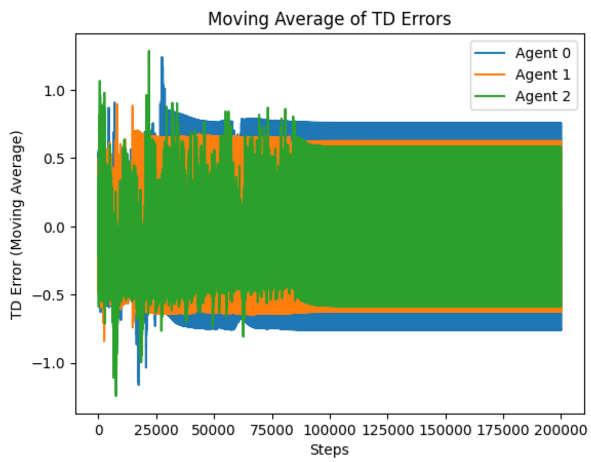

Fig. 5: TD Average with Obstacles

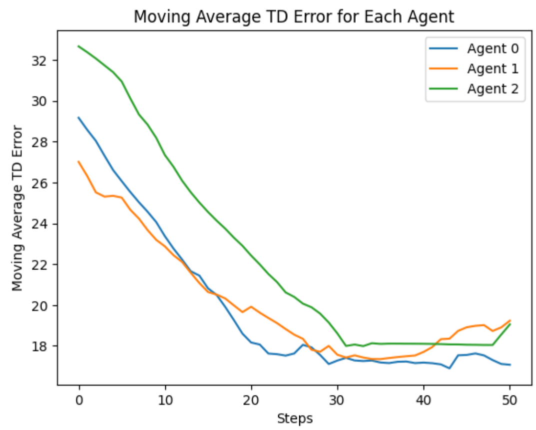

Fig. 6: TD Average without Obstacles

4.2 Grid Coverage Analysis

<u>Scenario 1: Without Obstacles</u>

- Coverage Efficiency: Uniform and efficient coverage due to the absence of obstacles.
- Breakpoint: ~80% of grid coverage before revisits increased.
- Time per Cell: ~1.2 steps per cell on average.

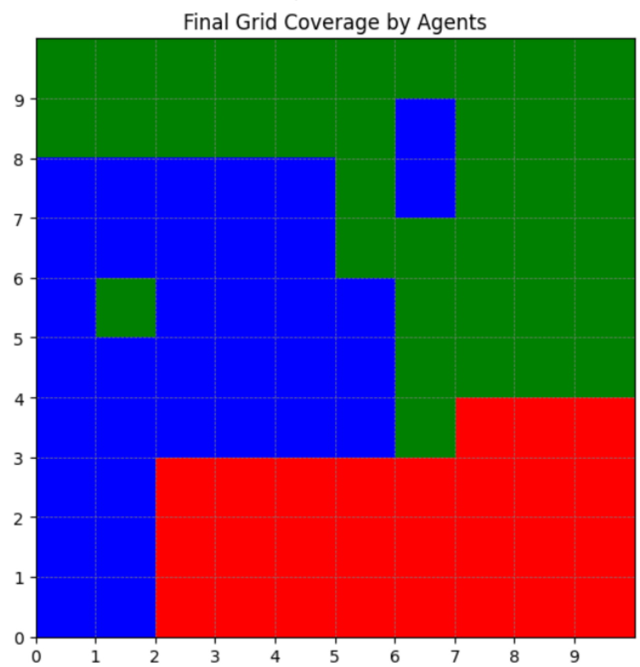

Fig. 7: Grid Coverage without Obstacles

<u>Scenario 2: With Obstacles</u>

- Coverage Efficiency: Slower and uneven coverage due to obstacle navigation challenges.

- Breakpoint: ~65% of grid coverage before revisits and penalties dominated.
- Time per Cell: ~2.5 steps per cell on average.

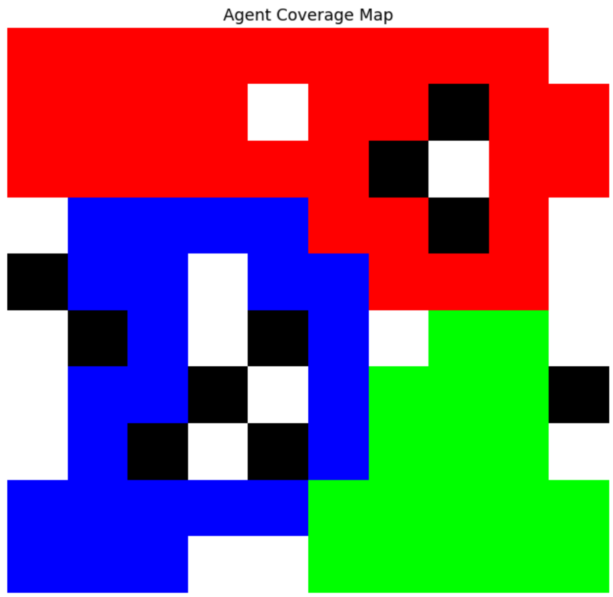

Fig. 8: Grid Coverage with Obstacles

Agent Behavior Near Obstacles:
Agents avoided cells near obstacles that could trap them into penalty zones. This cautious behavior occasionally led to “frozen” agents, particularly in cornered grid sections.

5. Conclusion

This project successfully demonstrated a multi-agent reinforcement learning framework for grid coverage using Q-Learning. Agents learned effective policies for collaborative exploration, adapting their behavior to dynamic environments with and without obstacles. The findings reveal the following key insights:

1. Obstacle-Free Environments: Faster and more uniform grid coverage, with later breakpoints.
2. Obstacle-Laden Environments: Increased complexity led to earlier breakpoints, slower coverage, and strategic avoidance of penalty zones.

Applications:
The methodology and findings from this project have potential applications in:
- Search and Rescue Operations: Coordinating robots to locate survivors in disaster zones.
- Precision Agriculture: Monitoring crops and optimizing resource allocation.
- Surveillance: Efficient area coverage in security and military applications.

Future work can focus on integrating agent communication and dynamic obstacle repositioning to enhance real-world applicability.