Capture The Flag (CTF) using Multi Agent RL Algorithms

Shirish Shekhar Jha, Mohit, Deepesh

Department of EECS and DSE
IISER Bhopal

November 28, 2024

Overview

Team Composition

Team Members are:-

- Shirish Shekhar Jha
- Mohit
- Deepesh Gavit

Problem Statement

- Capture the Flag environment has two teams with two agents in each team.
- Every team has the objective of capturing the opponent's flag, but at the same time defend its own.
- Defending the flag activates when an agent enters a visual depth of 3 near the opponent's flag.
- Obstacles, and flags positions were static, and two agents could occupy same cell.

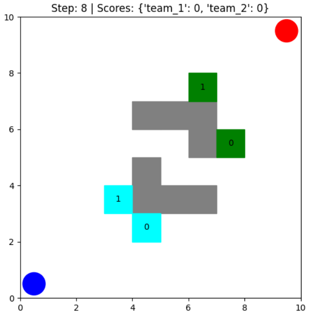


Figure: CTF Environment

Reinforcement Learning

Below are the algorithms used in the study:-

- Independent Q-Learning
- Multi Agent Proximal Policy Optimization (**Prefered Approach**)

MDP: Definition and Components (Part 1)

A Markov Decision Process (MDP) is defined as the tuple (S, A, d_R, d_0, γ), where:

- S denotes the state space. For our environment:

$$S \in \mathbb{R}^{10\times10\times4} \setminus S'$$

  where S' represents the spaces occupied by obstacles or flag positions.

- A defines the action space:

$$A = \{\text{Up } (0,1),\ \text{Down } (0,-1),\ \text{Left } (-1,0),\ \text{Right } (1,0),\ \text{Stay } (0,0)\}$$

- d_R represents the reward distribution. For our problem:
  - $+100$: For capturing the opponent's flag.
  - $+25$: For successfully defending the flag.
  - -25: For getting caught while intruding.
  - -2: For staying in the same position.

MDP: Definition and Components (Part 2)

Continuing the tuple (S, A, d_R, d_0, γ):

- d_0 (Initial State): Defines the starting positions of agents:
  - Team 1:

$$x \sim \text{Uniform}(0, \frac{\text{grid\_size}[n_R]}{2}),\ y \sim \text{Uniform}(0, \frac{\text{grid\_size}[n_C]}{2}),\ (x, y) \notin S'$$

  - Team 2:

$$x \sim \text{Uniform}(\frac{\text{grid\_size}[n_R]}{2}, n_R),\ y \sim \text{Uniform}(\frac{\text{grid\_size}[n_C]}{2}, n_C),$$

$$(x, y) \notin S', n_R, \& n_C$$

  are number of rows and columns in grid.

- γ (Discount Factor): Balances immediate and future rewards:

$$\gamma = 0.99$$

Challenges in the Problem

- Effectively shaping reward for exploration, defending, and capturing
- Achieving team objectives, when to start exploring to capture, when to defend the own territory.
- Delayed rewards for defending made it challenging for agents to learn.

Why Choose MAPPO?

- **Centralized Training with Decentralized Execution:** Facilitates effective coordination among agents.
- **Proven Performance:** Achieves competitive or superior results in cooperative multi-agent scenarios.
- **Stable Learning Dynamics:** On-policy nature ensures stability in complex interactions.

Alignment with Problem Requirements

- **Coordination:** Enables agents to learn joint policies for balanced offensive and defensive strategies.
- **Stability:** Ensures stable learning in environments with complex agent interactions.

Model Parameters

- **Policy Network:**
  - **Input Layer:** 100 neurons (corresponding to the flattened observation space).
  - **Hidden Layers:** Two fully connected layers with 128 neurons each, activated by ReLU functions.
  - **Output Layer:** 5 neurons representing the action logits.
- **Optimizer:** Adam optimizer with a learning rate of 0.0003.
- **Policy Loss:** Clipped surrogate objective to ensure stability during training:

$$L_{\text{policy}} = -\mathbb{E}\left[\min\left(r_t(\theta)A, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)A\right)\right]$$

where $r_t(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}$ is the probability ratio.

Defining Metrics for Evaluation

- **High Win Rate:** The team with a higher win rate indicates better team performance.
- **Low Draw Rate:** Indicates decisive outcomes, less stalemates.
- **Performance Stability:** Variance in scores across episodes to measure consistency.

Why These Metrics?

- Assess overall dominance and effectiveness of each algorithm.
- Lower variance in scores suggests consistent performance.

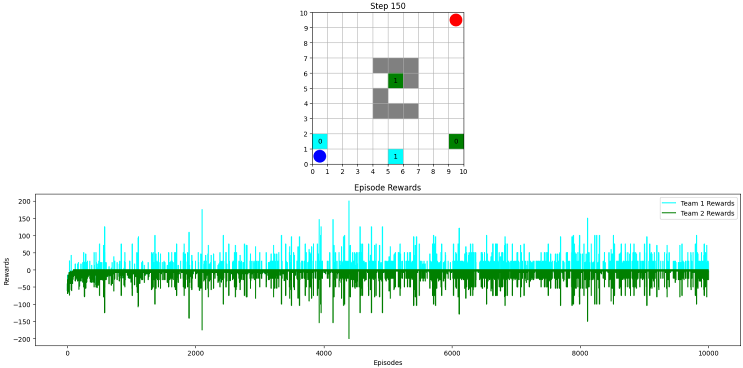


Figure: Loss progression during training (IQL)

Training Metrics: Loss and Rewards (MAPPO)

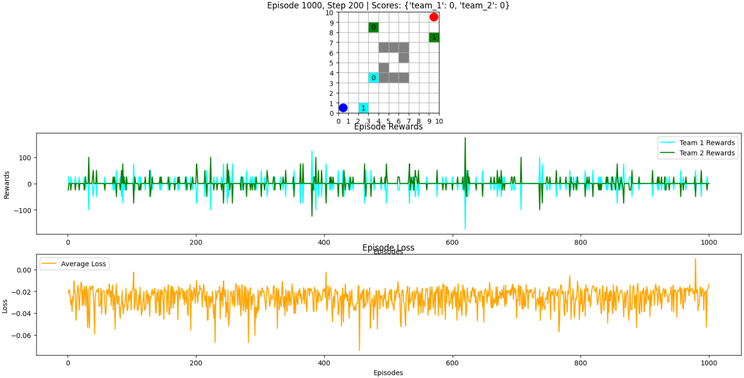
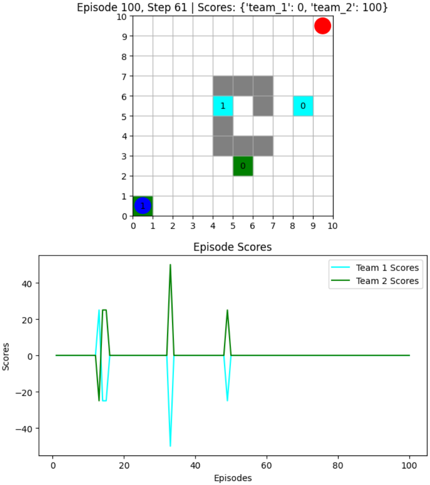
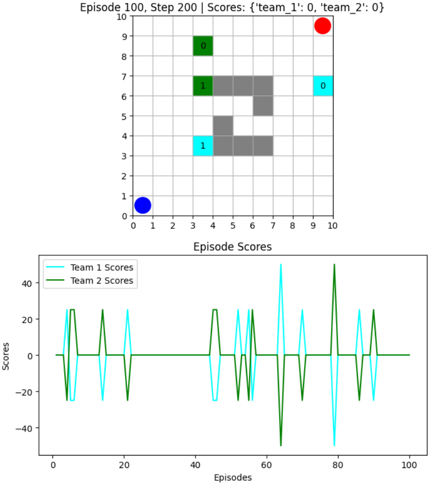


Figure: Loss progression during training (MAPPO)

Testing Metrics: Rewards (IQL)



Testing Metrics: Loss and Rewards (MAPPO)



Observations:

- **Win and Draw Rates:**
  - MAPPO: Team 1 (53.13%), Team 2 (46.88%).
  - IQL: Team 1 (57.14%), Team 2 (42.86%).

Visualized Results:

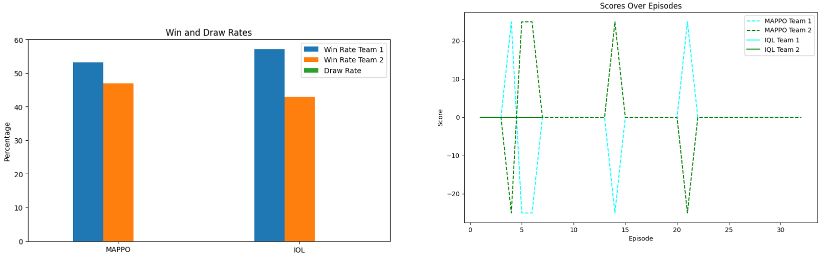


Figure: (Left) Win and Draw Rates; (Right) Scores Over Episodes.

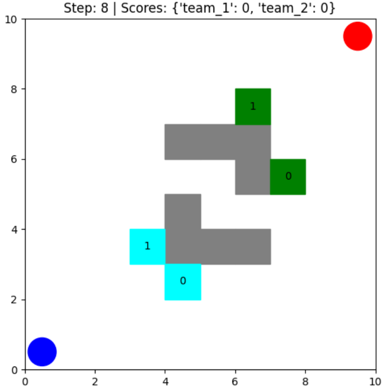


Figure: CTF Environment (Original)

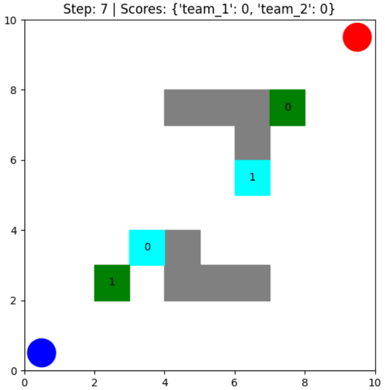


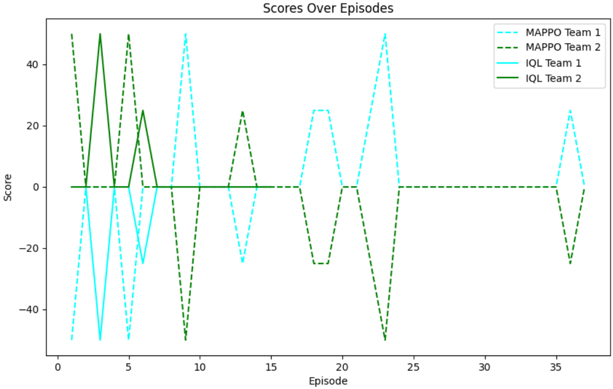
Figure: CTF Environment (Changed Obstacles)

Performance Metrics: Changed Environment

- **Win Rates:** MAPPO shows competitive balance (48.65% Team 1, 51.35% Team 2), while IQL favors Team 1 (53.33%).

Performance Metrics Over Episodes

- **Score Evolution:**
  - MAPPO exhibits stable dynamics with alternating scores over episodes, showing competitive engagement.
  - IQL has steeper score fluctuations, indicating independent decisions often fail to adjust dynamically.



Visualized Results:

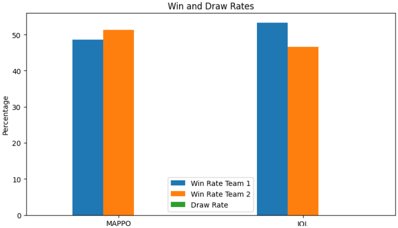
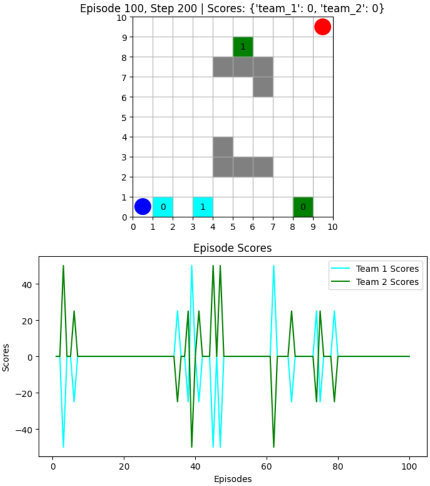


Figure: Win and Draw Rates

Testing Metrics: Rewards (IQL) in Changed Environment



Testing Metrics: Loss and Rewards (MAPPO) in Chaged Environment