Cooperative and Adversarial Multi-Agent Reinforcement Learning

Sean Osier, Emily Ye, Danny Zhang, Yutong Zhou

Project Motivation

- The algorithms we studied in class were all **single-agent focused**

- We were curious about reinforcement learning in **multi-agent settings**, particularly ones were there are mixes of agent groups (e.g., a team and an opposing team) that required aspects of both **cooperation and competition**

Background: Multi-Agent Reinforcement Learning (MARL)

- Type of MARL environments:
  - Agent Relationship: **Cooperative vs. Adversarial vs. Mixed**
  - Learning Style: **Decentralized vs. Centralized**
  - Degree of **Shared Knowledge / Communication**

- While there does **not appear to be a singular standard for best MARL algorithms** or approaches, popular ones include:
  - MAPPO (Multiagent-PPO)
  - Self-play
  - Curriculum learning

Background: DEC-POMDPs

- **DEC-POMDPs** = Decentralized Partially Observable Markov Decision Process

- Formally defined by: $\langle S, A, O, R, P, n, \gamma \rangle$
  - *S* = **Global** state space
  - *A* = Shared action space
  - $O(s, i) = o_i$ = **Local observation for agent *i***, at global state *s*
  - *R* = Reward function
  - *P* = Transition probabilities given state and **joint action** $(a_1, \ldots, a_n)$
  - *n* = Number of agents
  - γ = Discount factor

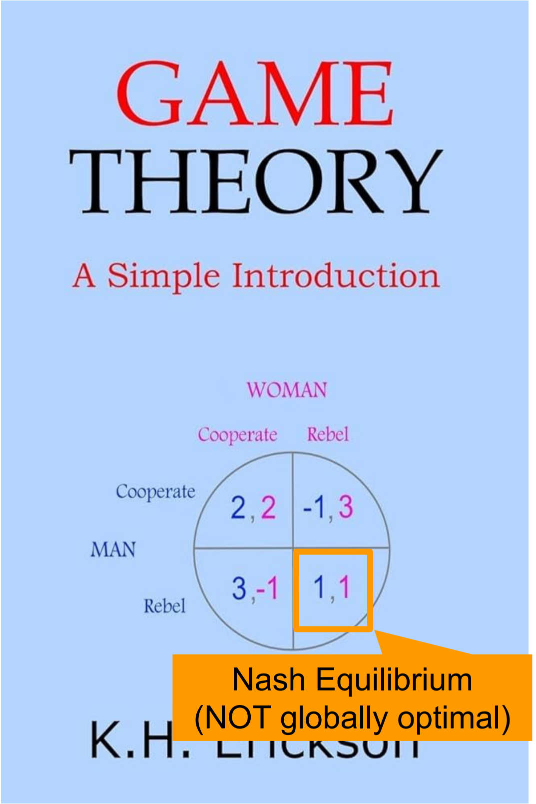
- **Decentralized** = Each agent keeps its own experience buffer (as opposed to have a shared buffer)
  - **Major drawback:** Markov process is **NOT stationary**, so **no guarantee of convergence**

Background: Game Theory 101 / Nash Equilibrium

- **Game Theory**: The study of mathematical models of strategic interactions among rational agent***s***
  - As such, ***very relevant*** to multiagent reinforcement learning

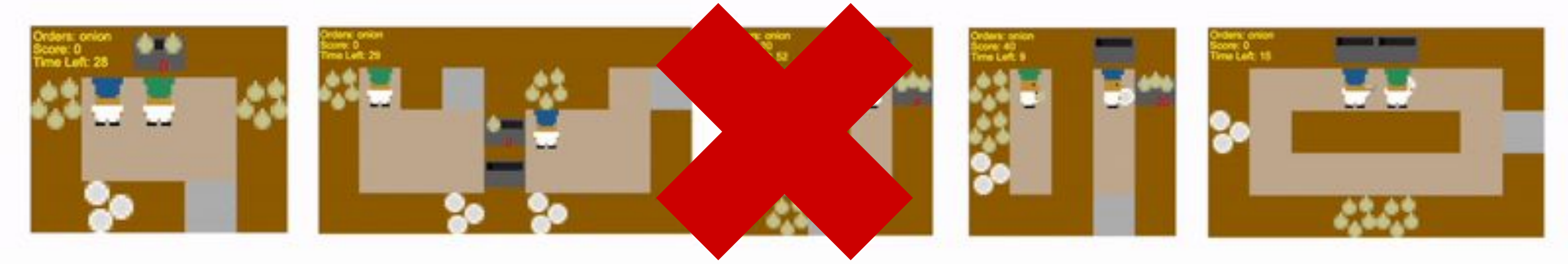
- **Nash Equilibrium:** A situation where **no agent can gain by changing their own strategy** (holding all other agents' strategies fixed)
  - The **"solution" to a game theoretic problem**
  - Stable solution, but importantly **may NOT be globally optimal** for the individual agents
  - Nash proved a Nash equilibrium **exists for every finite game**
  - As such, we **should expect our experiments to converge to Nash equilibria** (if they converge)

- **Prisoner's Dilemma:** The most famous, simple example



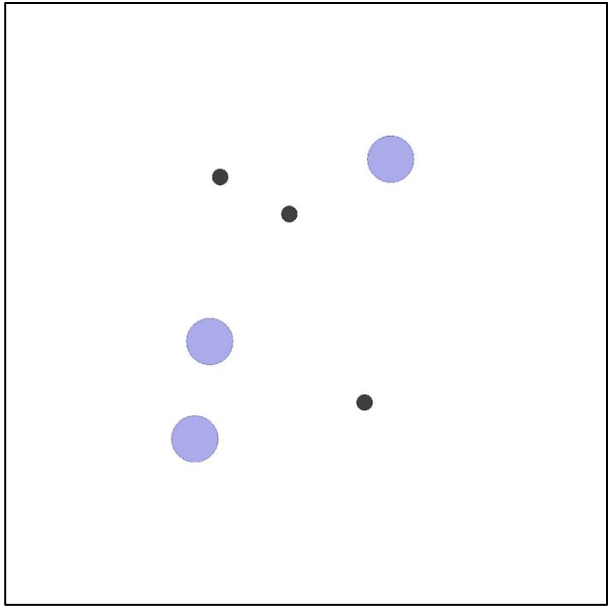
Our Environment: Overcooked → Simple Spread

- The **Overcooked** environment we intended to use was **not maintained**
  - **Caused large range of dependency errors** between the original Overcooked library, derivative libraries, and other utility libraries
  - Debugging and fixing the issues would have occupied a significant amount of our time
- Since **our focus was on experimenting with multi-agent reinforcement learning and not on building environments** for multiple agents, we decided it was **best to pivot** to a widely used and frequently updated/maintained library such as **pettingzoo,** specifically the **Simple Spread** environment



Our Environment: About Simple Spread

- This environment has **N agents, N landmarks** (default N=3)
- Agents are globally **rewarded based on how far the closest agent is to each landmark** (sum of the minimum distances)
- Locally, the agents are **penalized if they collide with other agents** (-1 for each collision)
- Agent **observes its own position / velocity**, as well as the **relative position of the other agents and landmarks**
- Possible actions are **move up, left, right, down, and no action**



Experimental Approach and Setup

- **Goal:** To study multi-agent behavior in cooperative, adversarial, and mixed-play settings, we aimed to examine agent performance under the following lenses: number of agents, amount of communication, and information transparency between adversarial and cooperative agents

- **DQN (for cooperative, adversarial, mixed environments):** Same as studied in class, but one DQN per agent
  - (Really 2 per agent if you count the target network)

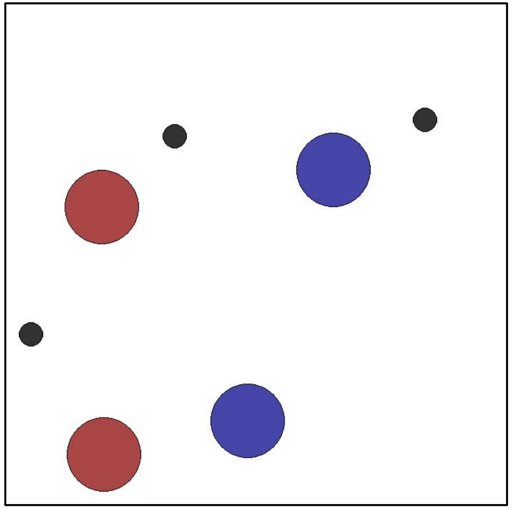
- **MA-PPO (for adversarial environments):** Adapted an MA-PPO library to work with our modified Simple Spread environment.
  - PPO (Schulman et al., 2017) is an on-policy algorithm that restricts magnitude of policy updates
  - We use a “clipped”, decentralized variant of PPO

Environment: Modifying to Support Adversaries

- Simple Spread by default is only cooperative

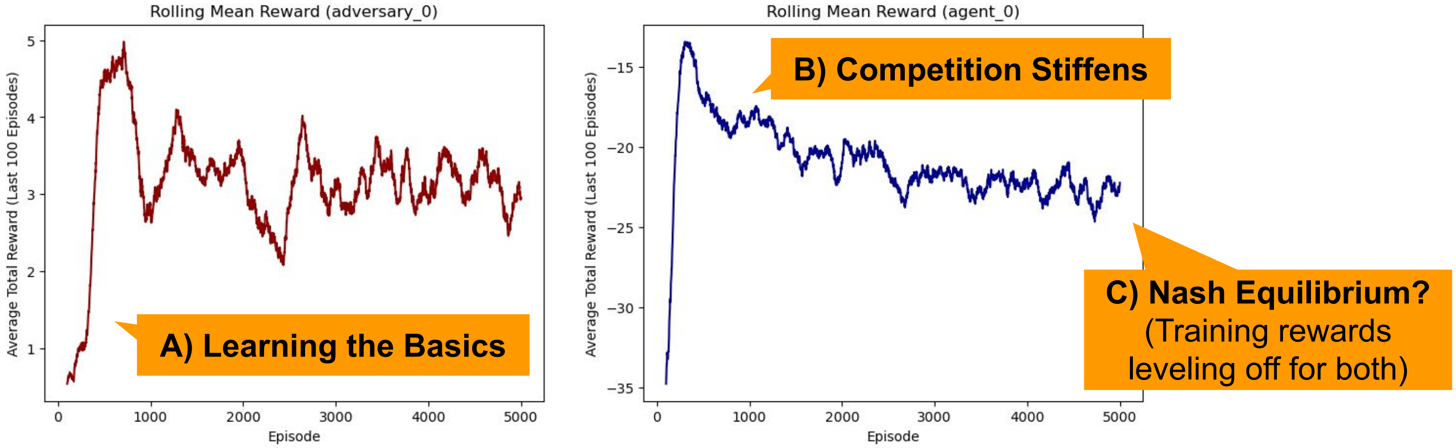
- While other adversarial environments were available, we thought it was important to be able to **compare agents across environments with with the same reward structure**

- Key changes:
  - Adding **adversaries**
  - Supporting **different number of landmarks and agents**
  - **Adversarial agent scoring:**
    - **+1 for collisions with "good agents"**
    - -1 for collisions with other adversaries



Adversarial Training: Example w/ 1 adv., 1 agent, 1 landmark

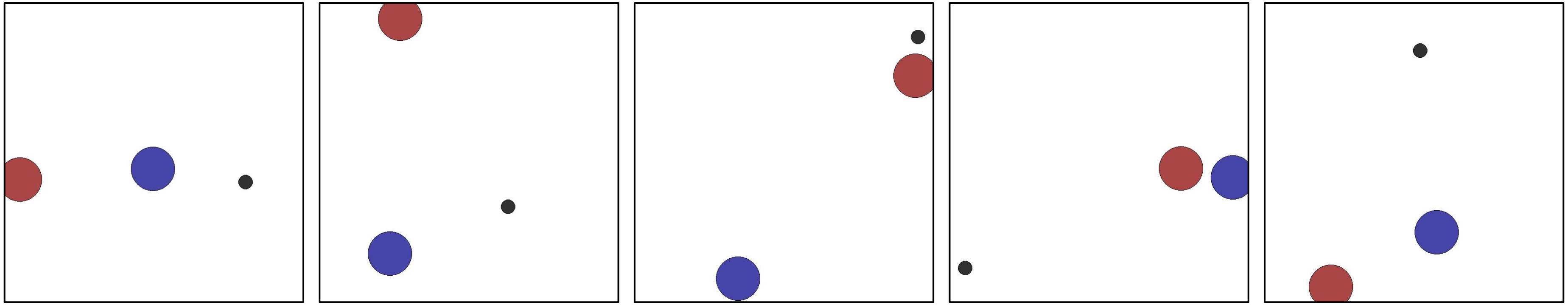
- In adversarial environments, it's **hard to know if your agents have "solved" the environment:**
    a. During training, we **should expect all agents get better at the start** as they learn the basics of the environment
    b. BUT, after that we **should expect to see performance worsen** because their competition is getting stronger also
    c. Ultimately we should expect the agents to converge to a **Nash equilibrium** (if they converge)



Example shown above is of the DQN agents during training. PPO follows a similar trend.

Adversarial Test Results: Example 1 adv., 1 agent, 1 landmark

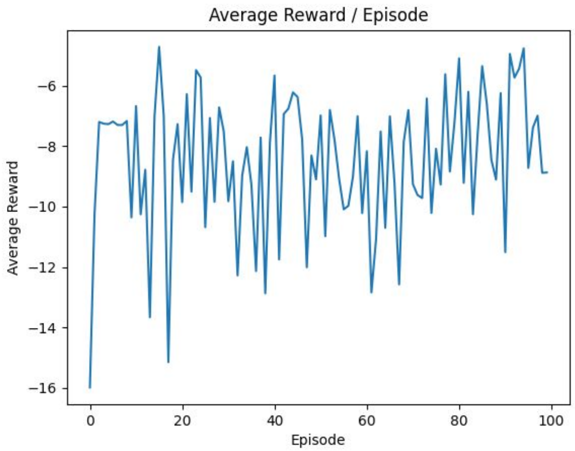
- Qualitatively, we observe that our agents have learned **sophisticated behaviors:**
  - **Precise movement skills**, including varying speeds where appropriate
  - First and even second order **anticipation** (what you might call planning)
  - An ability to **leverage the physics of the environment to their advantage**



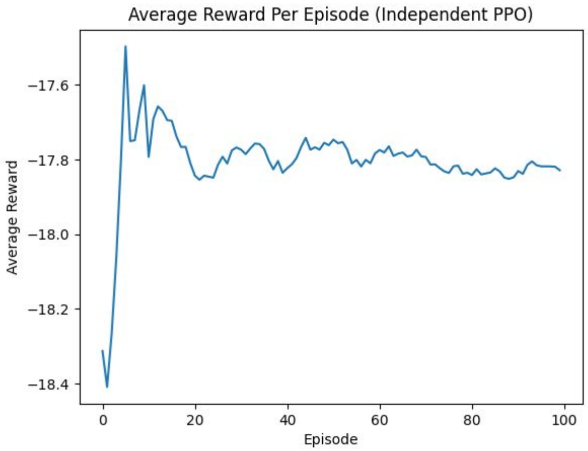
Adversarial Test Results: Example 1 adv., 1 agent, 1 landmark

- This is **what we observe in our DQN and PPO gameplay after training:**
  - Reward levels are roughly comparable (depending on the run, PPO slightly outperforms DQN or vice-versa. Below is an example where DQN slightly outperforms PPO)
  - PPO always appears to be more stable.

DQN



PPO



Scaling Beyond 1v1 Adversarial Gameplay

- We tested various kinds of scenarios with larger numbers of agents, adversaries, and landmarks:
  - E.g., adversaries are “outnumbered” and/or at disadvantage (2 adv, 4 agents, 4 landmarks, 2 adv., 4 agents, 8 landmarks)

- Across all runs for both DQN and PPO, average **adversary rewards** tended to hover in the **low positive/negative single-digits** while **agent rewards** were **more strongly negative, particularly** (and understandably) **as the number of agents and adversaries increased**

- Depending on the run, as well as the scenario, **DQN sometimes performs better than PPO and vice-versa** (though **PPO was always more “stable”** with its average rewards).

Environment: Modifying for Communication Modes

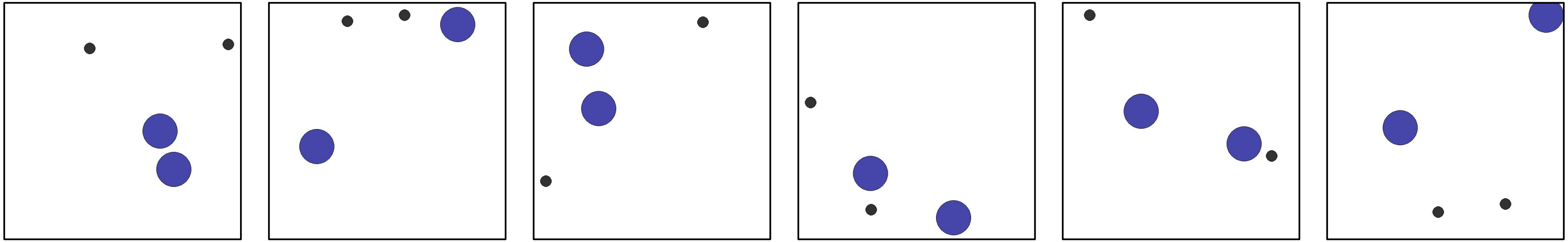
- In addition to modifying the source code to add adversaries, we also modified the amount of information between agents by adding a variety of communication modes

- Key changes:
  - Added following communication modes for what information individual agents can receive:
    - Mode 0: Information on other agent positions is masked
    - Mode 1 (Baseline for information): Information on the location of other agents and landmarks is provided
    - Mode 2: Additional information on other agent velocity
    - Mode 3: Additional information on other agent’s Euclidian distance to all landmarks
    - Mode 4: Similar to Mode 3, but now each other agent provides a binary variable if they are within .5 distance to any landmark
    - Mode 5: Appends the information from mode 2 and 3 to the baseline

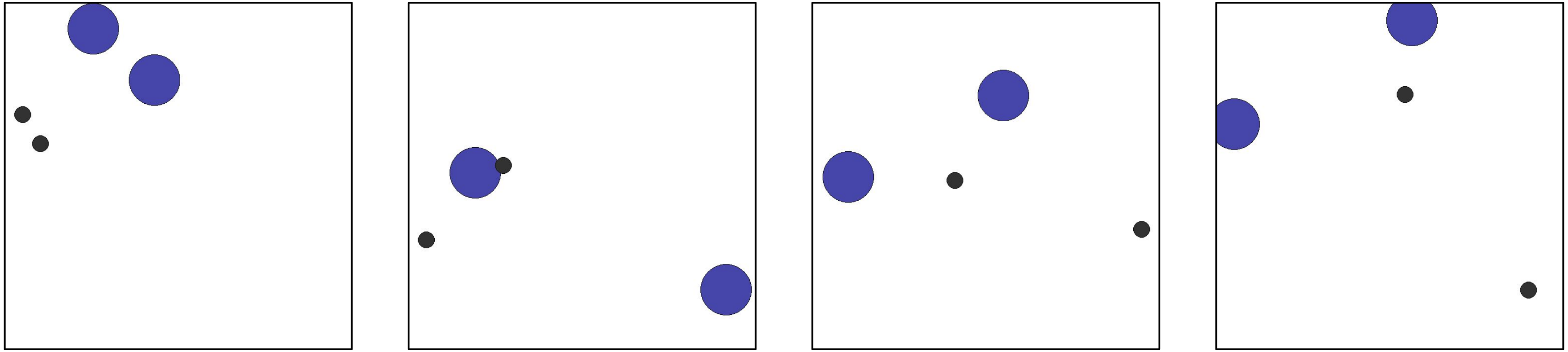
Communication Results: 2 agents, 2 landmarks(1/2)

- **Qualitative results**
  - Counterintuitively, passing more information about other agents in the environment does not always yield improved results
  - Information regarding other agents’ positions/velocities appears to yield best mean rolling rewards. These rewards are decreased when other agents’ Euclidean distances are also included in communication

Mode 0 | Mode 1 | Mode 2 | Mode 3 | Mode 4 | Mode 5



Communication Results: 2 agents, 2 landmarks(2/2)



- **Gameplay patterns**
  - With no communication between agents, agents will often reach landmarks quickly but can sometimes incur a large penalty for repeated collisions
  - With communication added, agents are much more likely to be able to adjust the travel trajectory to avoid collisions
  - In rare cases, agents use bumping as a method to increase velocity towards landmarks

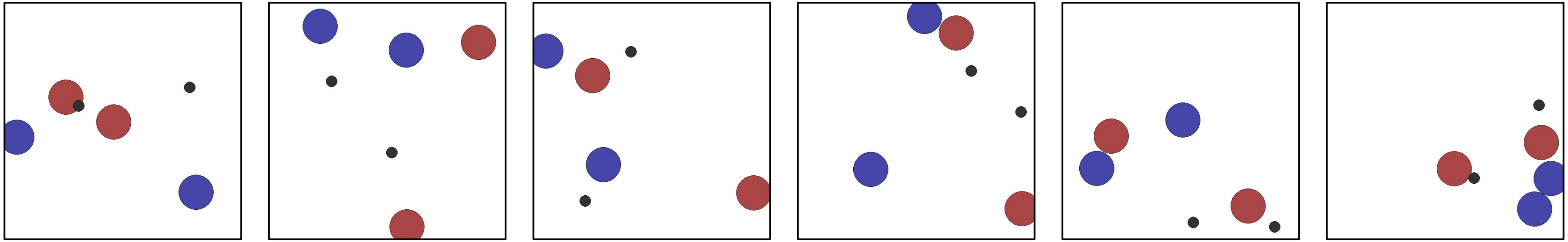
Combining Adversarial and Communication Experiments

- **We incorporate the communication parameter we use for our cooperative communication experiments into our adversarial agents and non-adversarial agents**
  - We would like to observe the effects of different levels of communication in a competitive environment
- **We begin by examining results under a transparent information passing setting**
  - The adversary receives a communication vector with the same parameters from cooperative agents, and vice versa

Adversarial Communication Results: 2 adversaries, 2 agents, 2 landmarks, information transparency

- **Qualitative results**
  - Lack of success for non-adversarial agents in reaching goal landmarks, which can likely be attributed to the adversaries receiving communication regarding their positions
  - Significant number of instances where the adversaries will “sit” on top of the landmarks to prevent non-adversaries from reaching their goals

Mode 0 | Mode 1 | Mode 2 | Mode 3 | Mode 4 | Mode 5



Adversarial and Communication Experiments: No Information Transparency

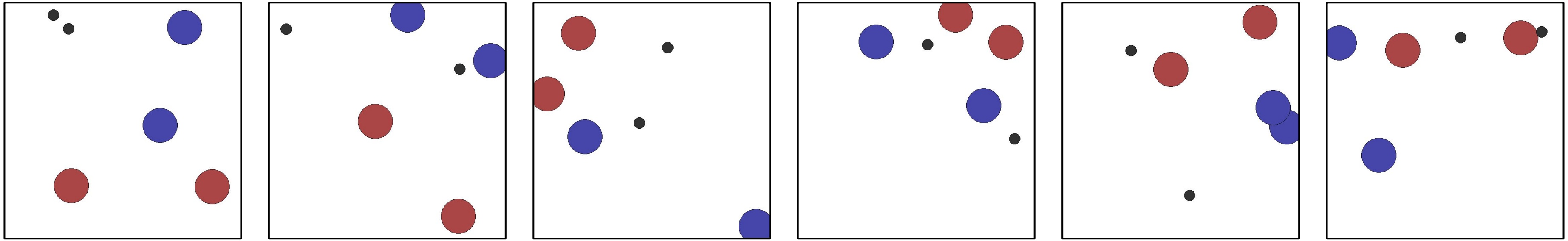
- **Logically, adversaries would not communicate to non-adversaries, and vice versa**
  - We devise the adversarial communication experiments under an information non-transparent setting, where adversaries communicate with adversaries and "regular" agents communicate amongst themselves

- **When compared to the information transparent setting, we observe the following:**
  - Decreased rolling mean rewards for adversaries
  - Increased rolling mean rewards for non-adversaries

Adversarial Communication Results: 2 adversaries, 2 agents, 2 landmarks, no information transparency

- **Qualitative results**
  - More obvious signs of cooperation between agents amongst themselves, and vice versa for adversaries
  - Adversaries seem to underperform compared to the cooperative agents. The cooperative agents are able to reach their goal landmarks most of the time without incurring significant penalties from adversaries

Mode 0 **Mode 1** **Mode 2** **Mode 3** **Mode 4** **Mode 5**



Conclusion: Limitations / Next Steps

- As initially expected, communication helps group performance while the introduction of adversarial elements hurt
- Difficult to both push towards and quantitatively measure the optimality of gameplay in mixed/adversarial environments
- Areas for potential future exploration:
  - More complex multiagent environments or tasks (e.g., tasks with extremely long step-sequences)
  - Different algorithms for multiagent cooperative and/or adversarial gameplay
  - Quantitative methods for measuring optimality of mixed/adversarial gameplay

References:

1. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games: https://arxiv.org/pdf/2103.01955
2. Reinforcement Learning to Play an Optimal Nash Equilibrium in Team Markov Games: https://proceedings.neurips.cc/paper_files/paper/2002/file/f8e59f4b2fe7c5705bf878bbd494ccdf-Paper.pdf
3. Proximal Policy Optimization Algorithms (Schulman et al., 2017) https://arxiv.org/abs/1707.06347
4. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments: https://arxiv.org/pdf/1706.02275

Our GitHub Page:

- https://github.com/sosier/Multi_Agent_Reinforcement_Learning

Appendix

Additional Experiments

Contributions

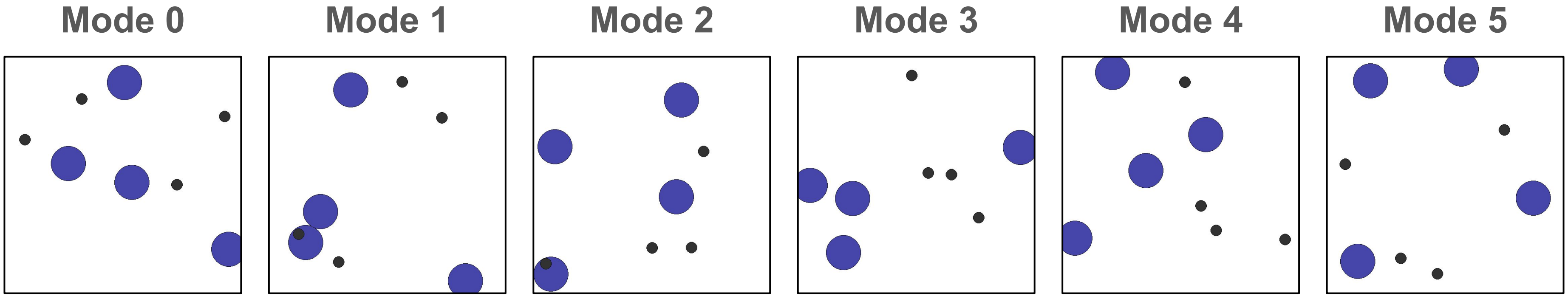
- **Sean Osier:** Built the adversarial simple spread environment; implemented / adapted DQN algorithm for multi-agent setting; performed adversarial experiments
- **Emily Ye:** Adapted PPO algorithm for multi-agent mixed adversarial gameplay; ran/experimented with different versions of MA-PPO on different simple spread settings
- **Danny Zhang:** Modified DQN algorithm and agent class for communication setting; ran communication experiments across all modes for different numbers of agents
- **Yutong Zhou:** Ran communication experiments across all modes for different numbers of agents; modified agent code for communication in adversarial setting under information transparent/non-transparent settings
- **All:** Worked on presentation

Communication Experiments

- **We are examining the effects of different communication modes within the 2 agent, 2 landmark configuration**
  - Other configurations such as 3 agents/3 landmarks, 2 agents/3 landmarks, 2 agents/4 landmarks, and 4 agents/4 landmarks have been omitted from the results due to a high degree of similarity to the 2 agent/2 landmark game
  -
- **The rewards between agents are homogeneous for our cooperative experiments; the charts for only one agent from each experiment are shown**

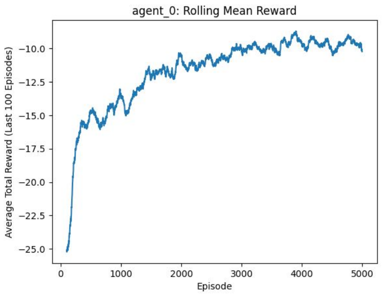
Communication Results: 4 agents, 4 landmarks

- **Qualitative results**
    - Using 4 agents and 4 landmarks as the configuration yields relatively poor performance with no communication enabled
    - Unlike in 2 agent 2 landmark configuration, adding additional communication to this configuration does not meaningfully increase quality of gameplay

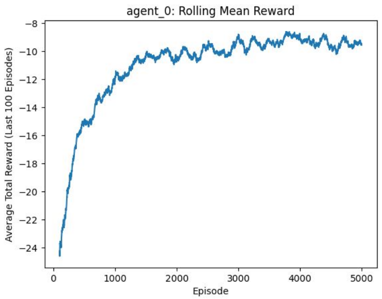


Communication Results: 2 agents, 2 landmarks(2/3)

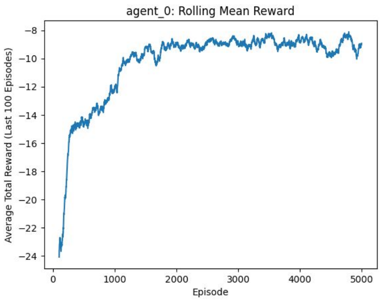
Mode 0



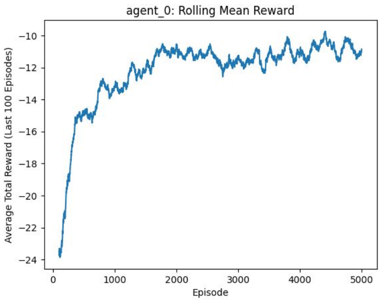
Mode 1



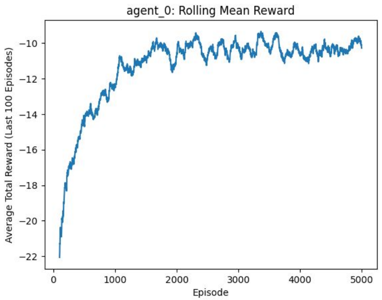
Mode 2



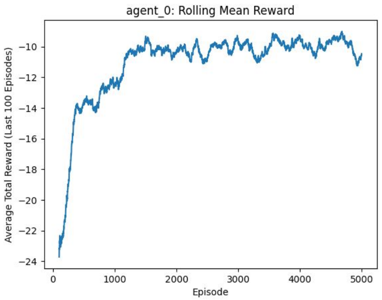
Mode 3



Mode 4

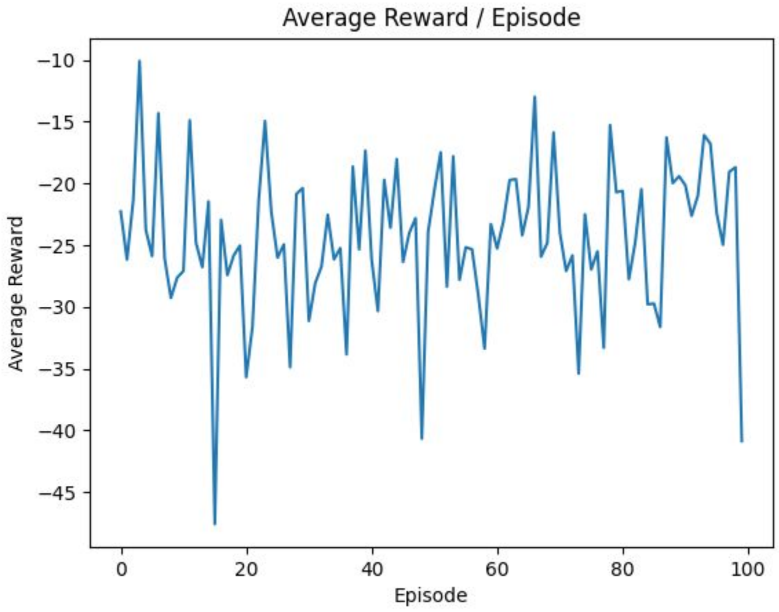


Mode 5

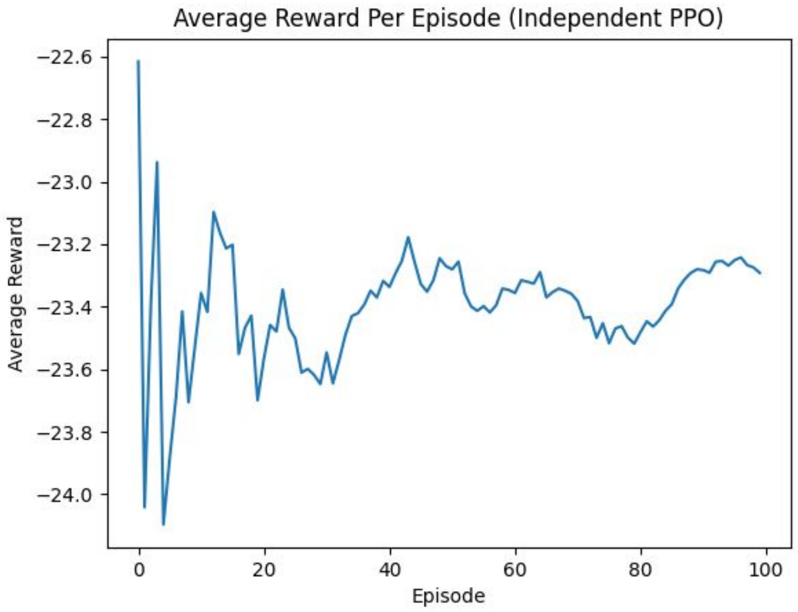


Example: 2 adv., 4 agents, 4 landmarks

DQN



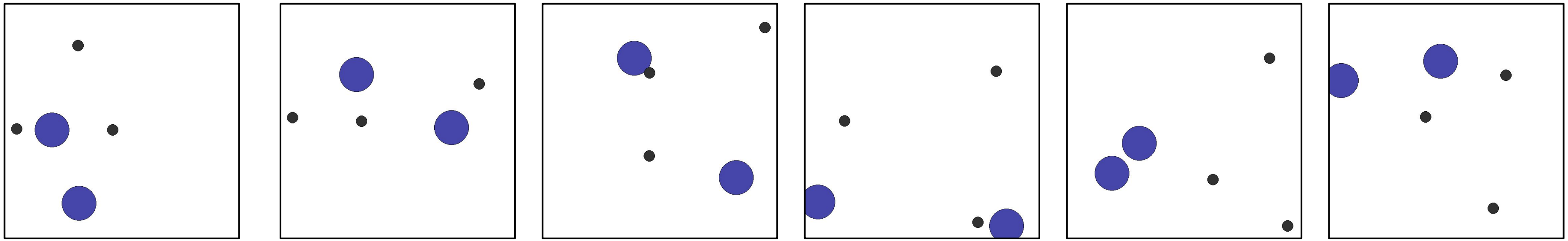
PPO



Communication Results: 2 agents, 3 landmarks

- **Qualitative results**
  - Mode 1, where the agents receive communication on other agents' positions, allows the agents to coordinate their respective goal landmarks, so they do not seek the same ones
  - Mode 3 (Euclidean distance from landmarks) yielded results where one agents was unable to reach the goal landmark
  - Mode 5 (velocities and Euclidean distances) yielded results where both agents failed to reach goal landmarks

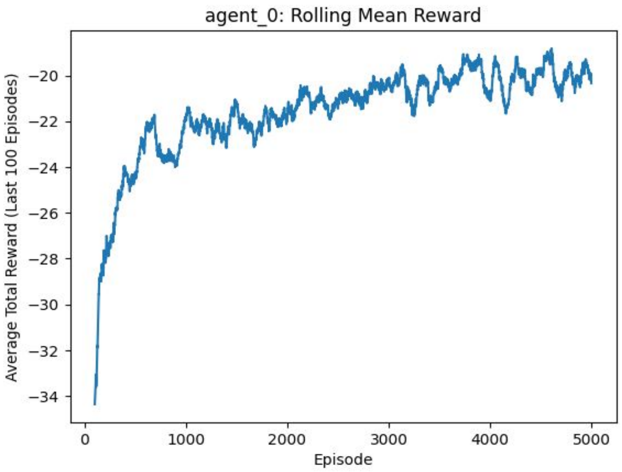
Mode 0 **Mode 1** **Mode 2** **Mode 3** **Mode 4** **Mode 5**



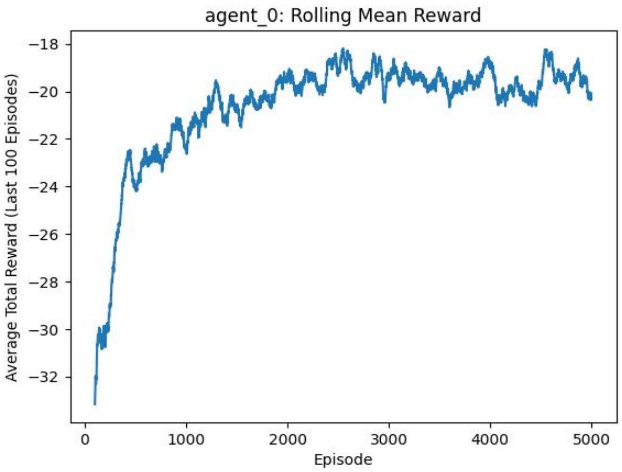
Communication Results: 2 agents, 3 landmarks

- **Decreased rolling mean rewards when compared to 2 agents, 2 landmarks:**
  - Similar rolling mean reward changes between communication modes

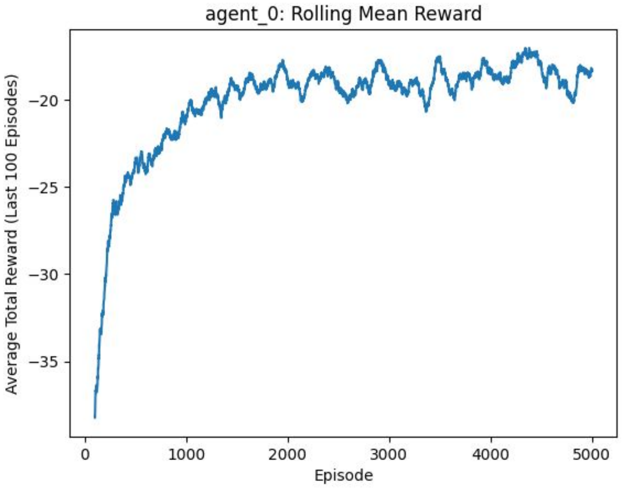
3 landmarks: Mode 0



3 landmarks: Mode 1



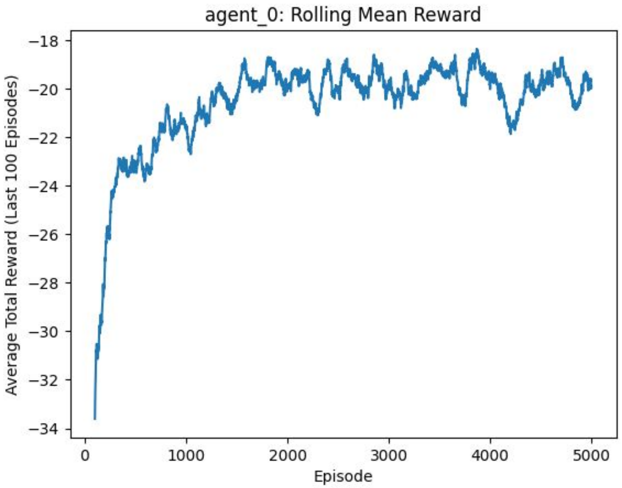
3 landmarks: Mode 2



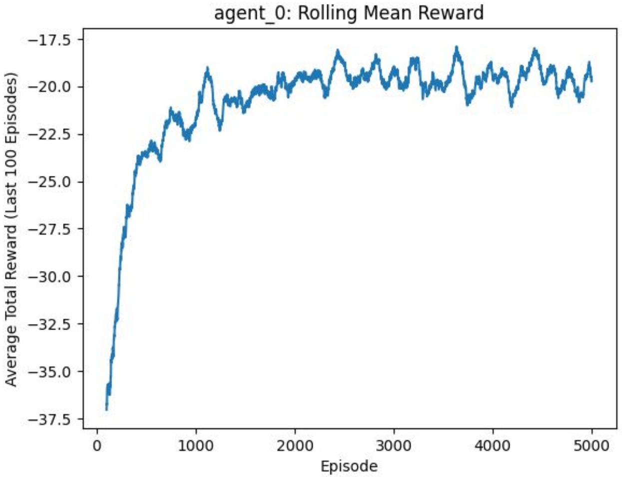
Communication Results: 2 agents, 3 landmarks

- **Benefits of increased communication more evident with increased number of landmarks**

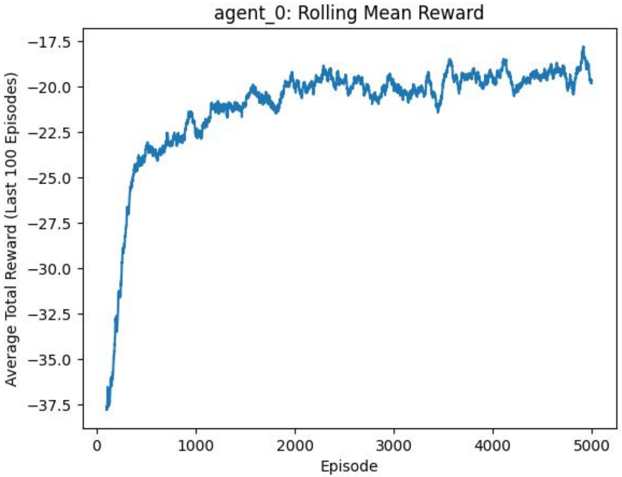
3 landmarks: Mode 3



3 landmarks: Mode 4

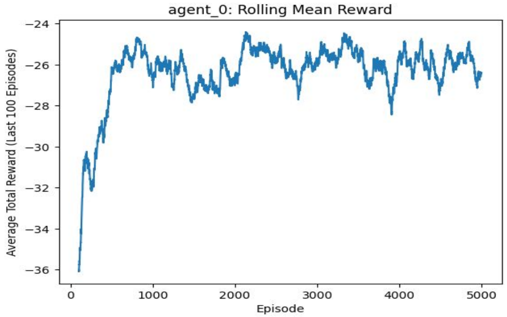


3 landmarks: Mode 5

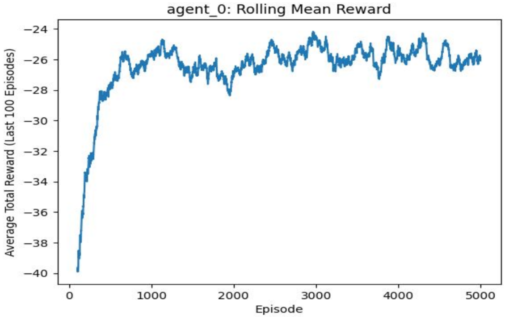


Communication Results: 4 agents, 4 landmarks

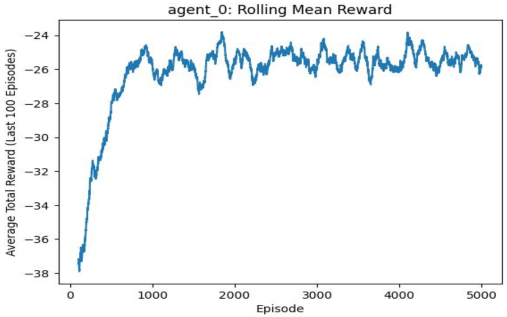
Mode 0



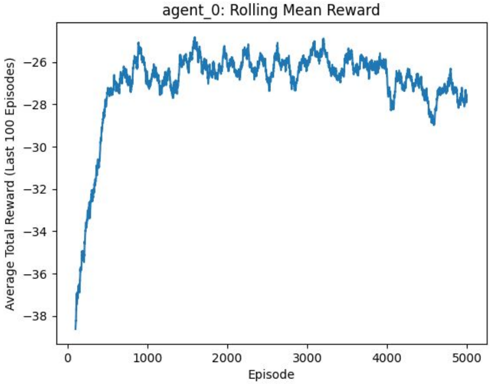
Mode 1



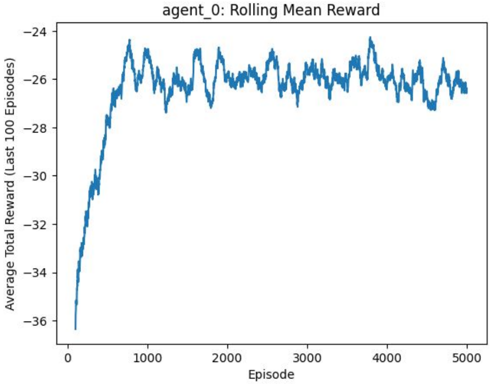
Mode 2



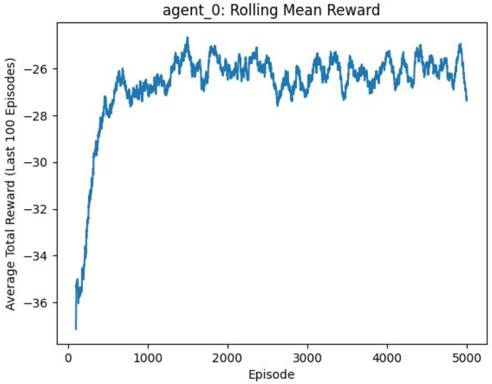
Mode 3



Mode 4

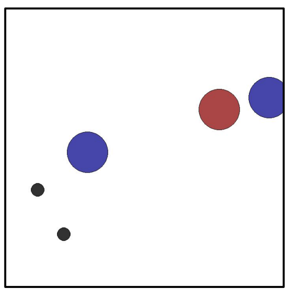


Mode 5

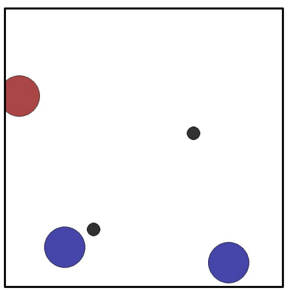


Adversarial Communication Results: 1 adversary, 2 agents, 2 landmarks, information transparency

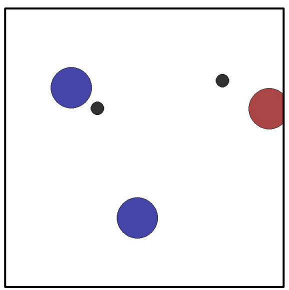
Mode 0



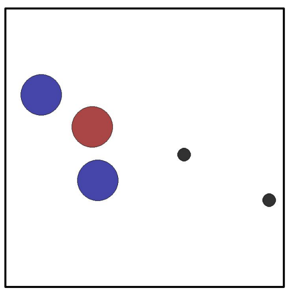
Mode 1



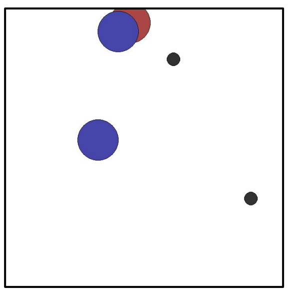
Mode 2



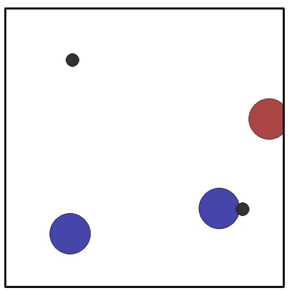
Mode 3



Mode 4

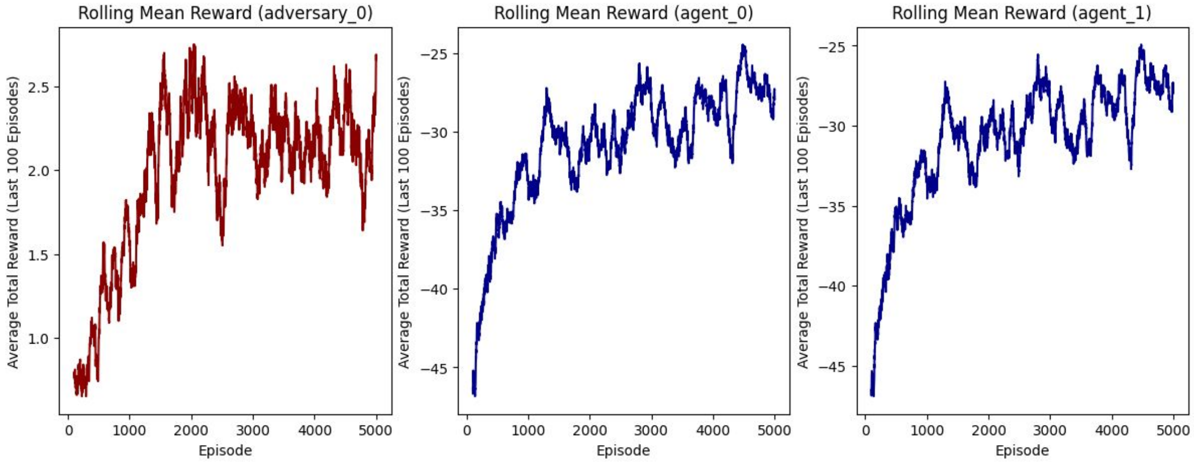


Mode 5

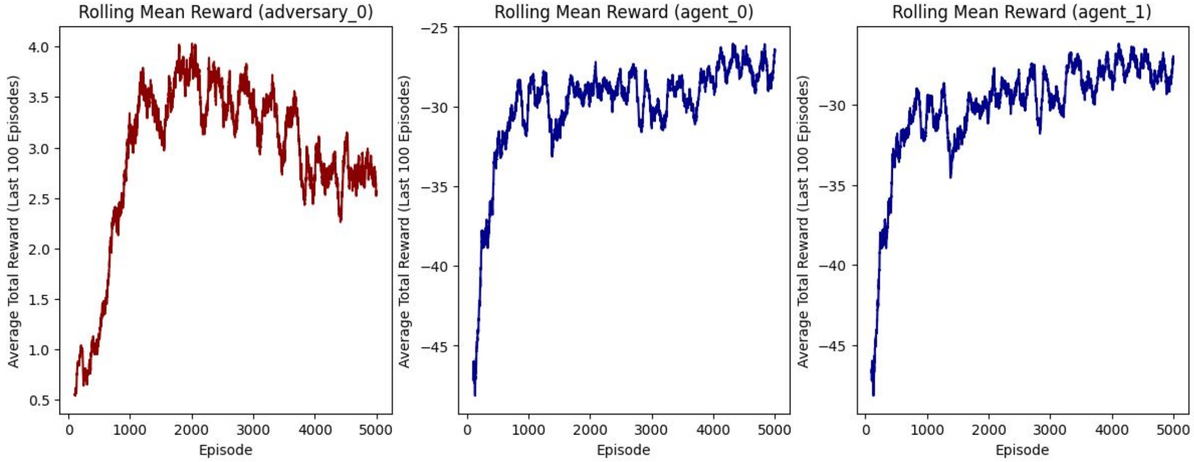


Communication Results: 1 adversary, 2 agents, 2 landmarks, information transparent

Mode 0

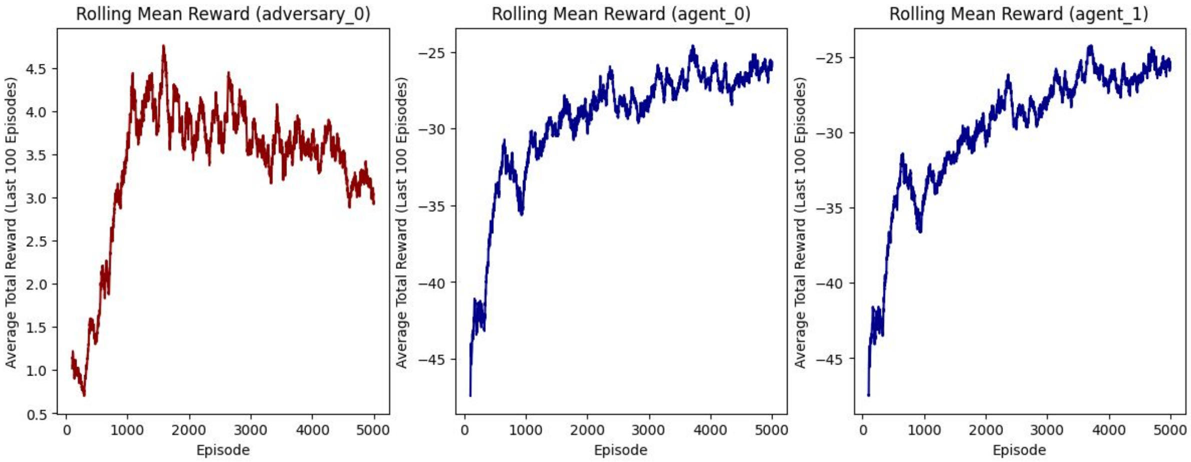


Mode 1

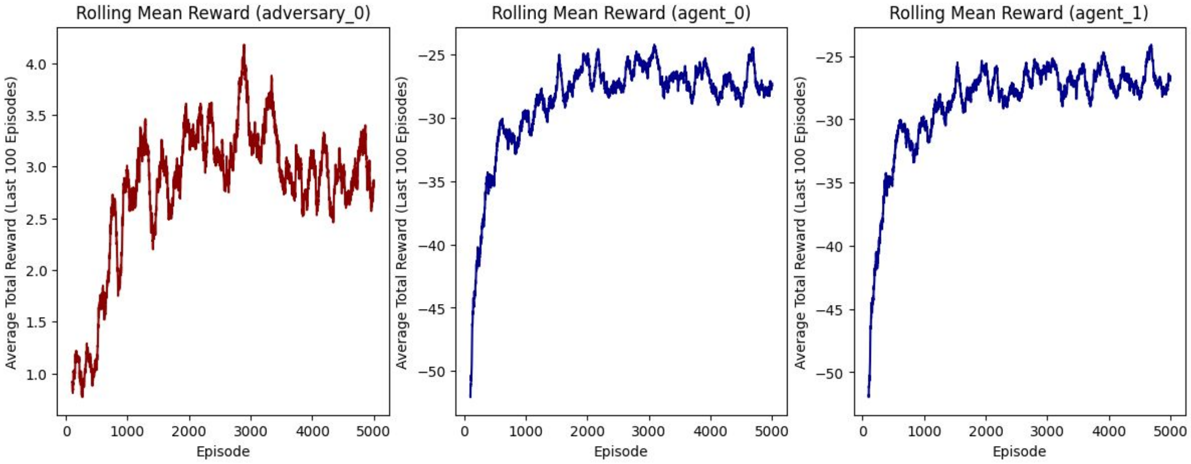


Communication Results: 1 adversary, 2 agents, 2 landmarks, information transparent

Mode 2

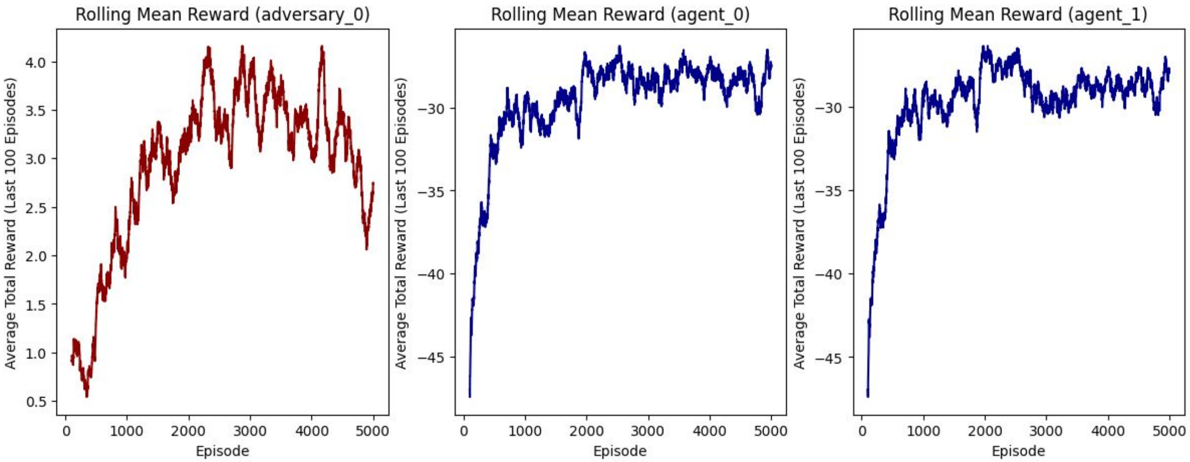


Mode 3

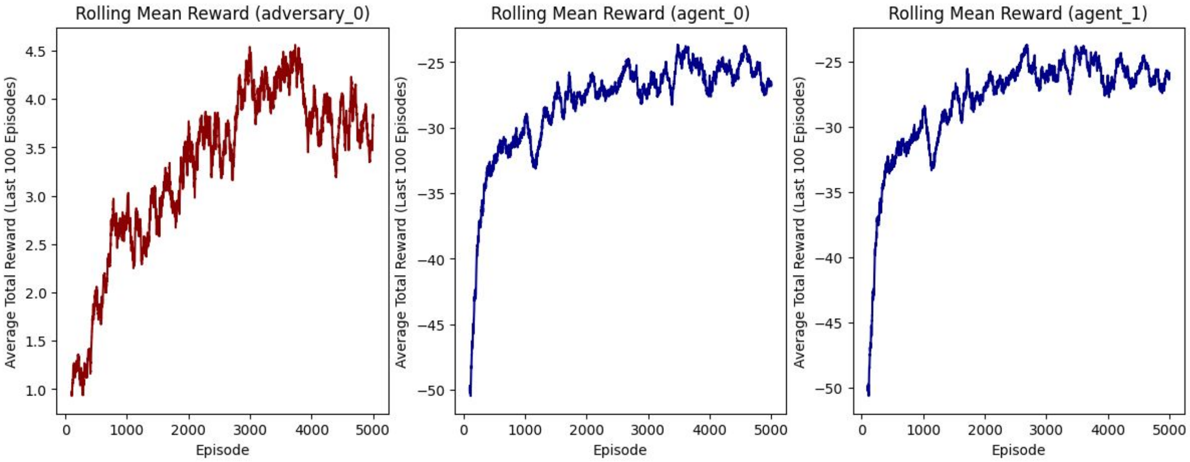


Communication Results: 1 adversary, 2 agents, 2 landmarks, information transparent

Mode 4

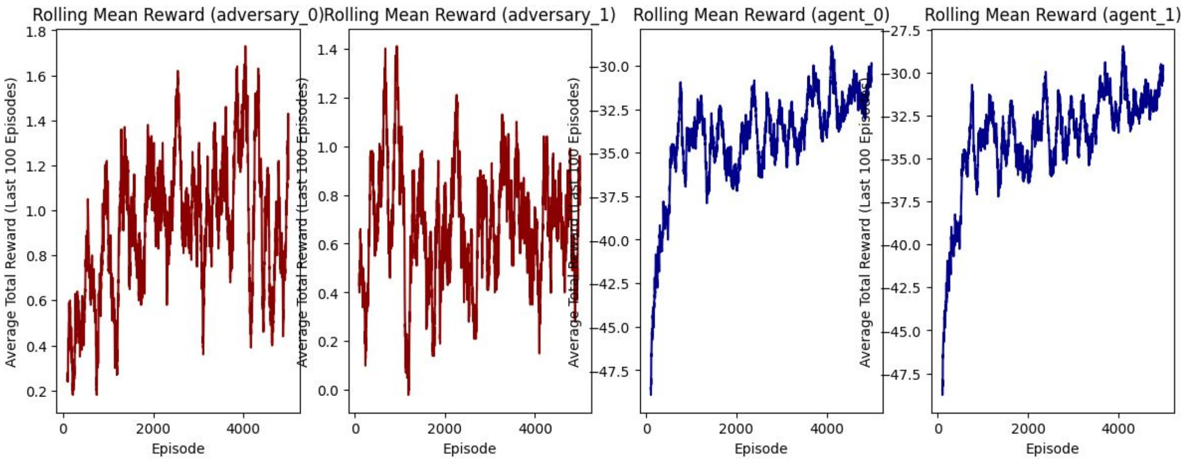


Mode 5

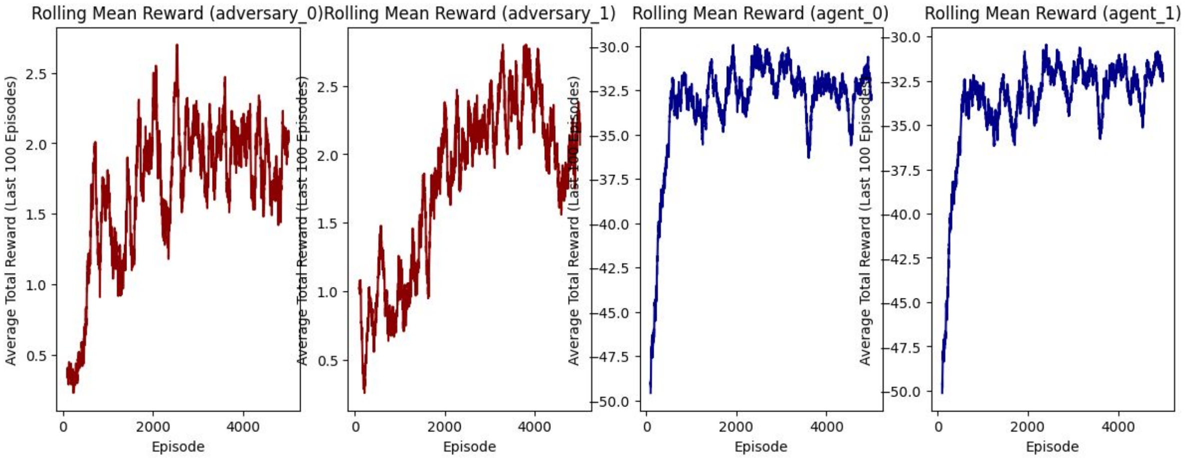


Communication Results: 2 adversaries, 2 agents, 2 landmarks, information transparency

Mode 0

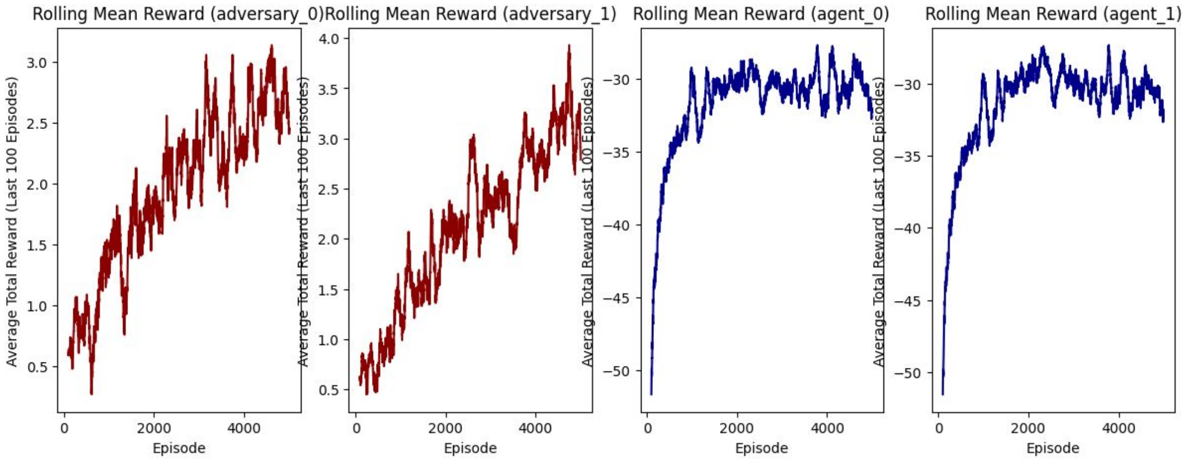


Mode 1

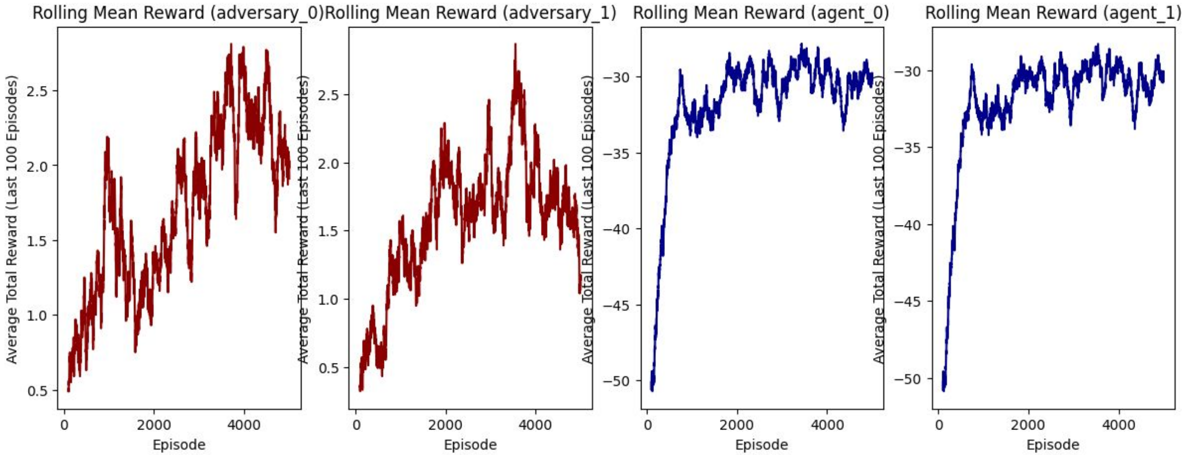


Communication Results: 2 adversaries, 2 agents, 2 landmarks, information transparency

Mode 2

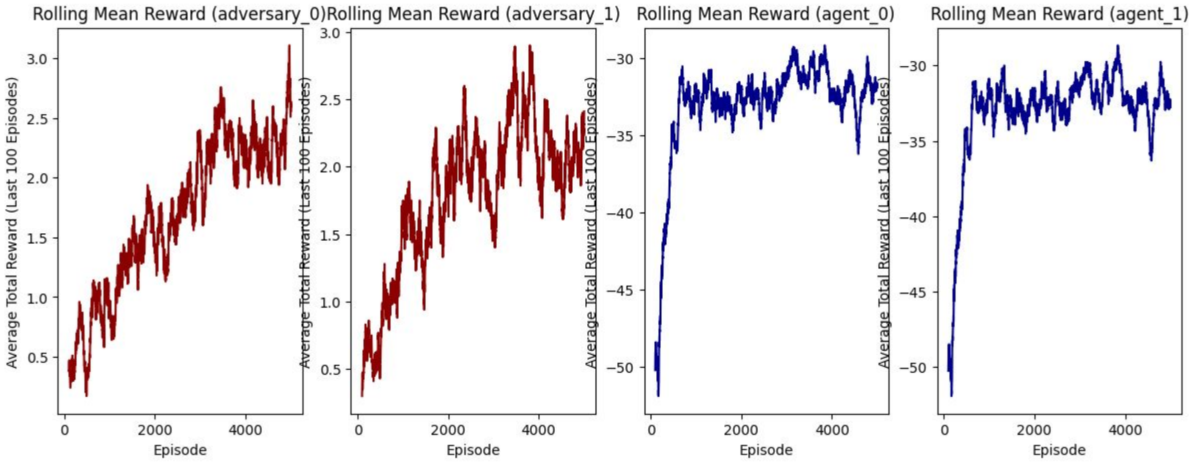


Mode 3

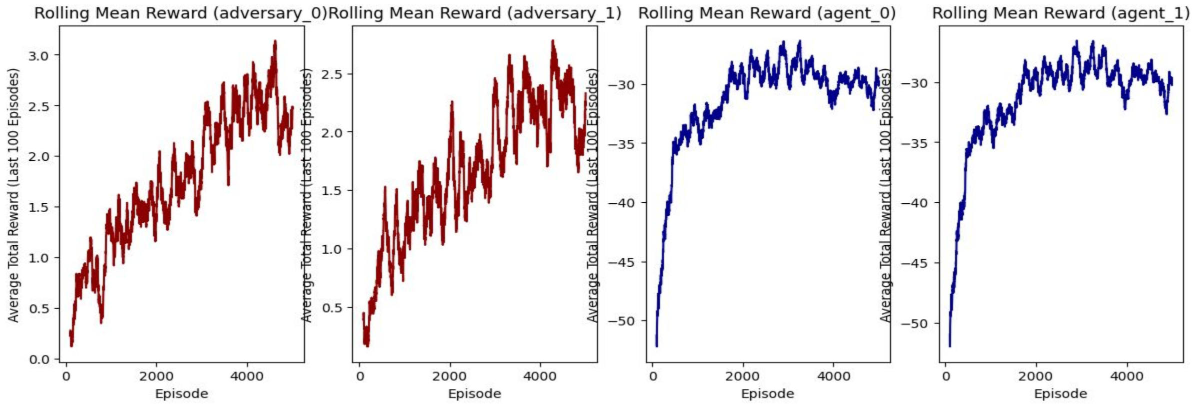


Communication Results: 2 adversaries, 2 agents, 2 landmarks, information transparency

Mode 4



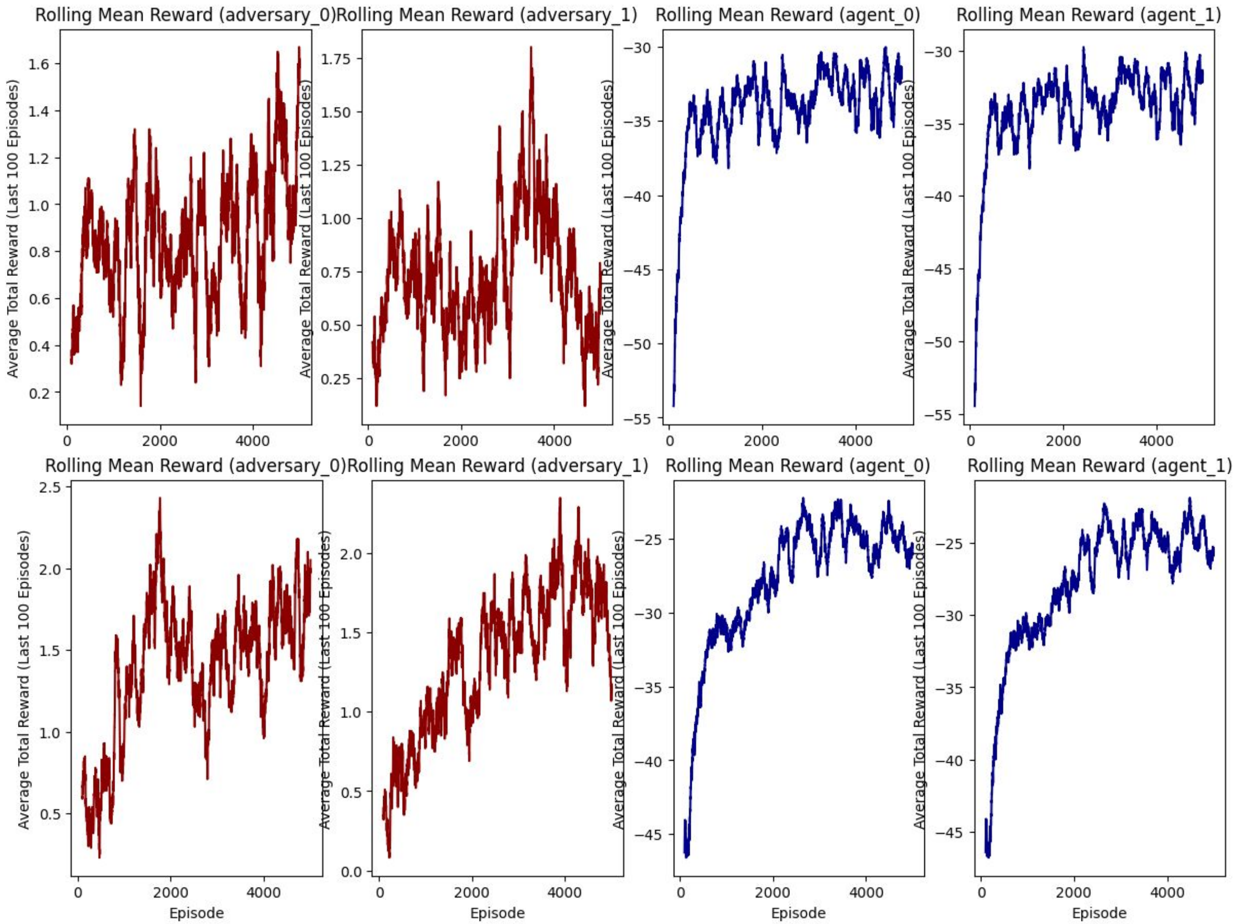
Mode 5



Communication Results: 2 adversaries, 2 agents, 2 landmarks, no information transparency

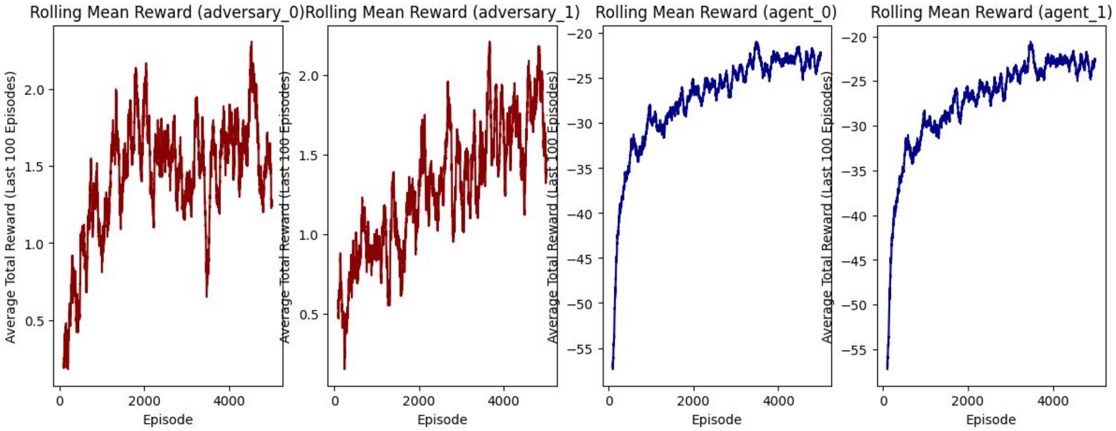
Mode 0

Mode 1

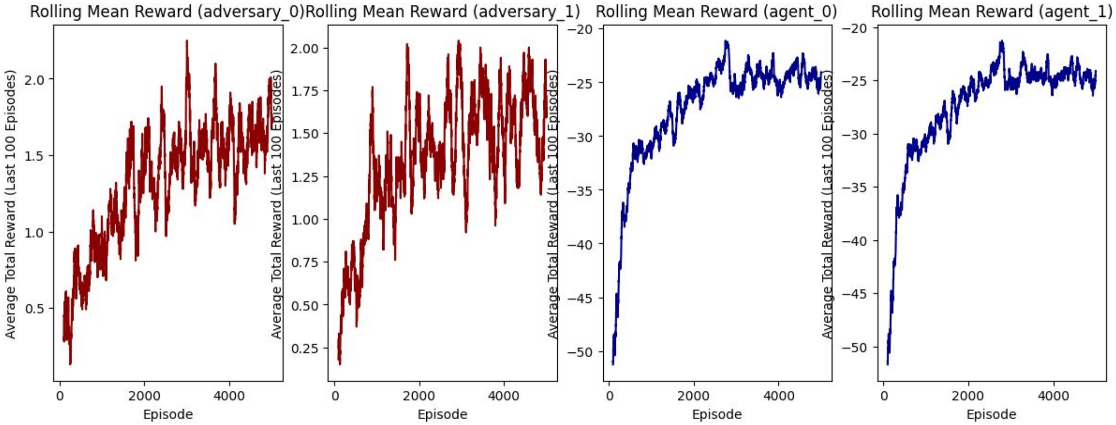


Communication Results: 2 adversaries, 2 agents, 2 landmarks, no information transparency

Mode 2

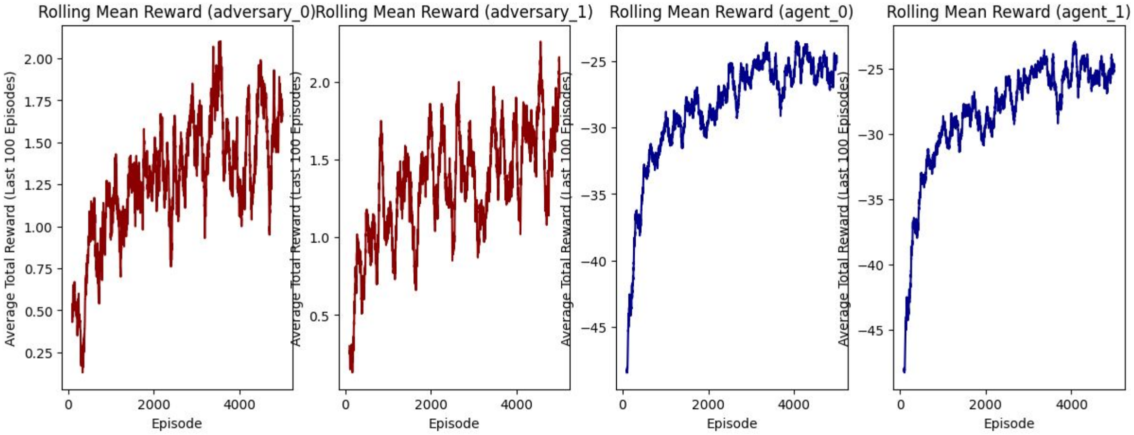


Mode 3



Communication Results: 2 adversaries, 2 agents, 2 landmarks, no information transparency

Mode 4



Mode 5