|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | game theory | | | reinforcement learning | | | multiagent | |
| coop | randomised policy | observability | sparse reward | randomly generated env. | scalable  env. | agent | agent type |
| Matrix Games | coop or comp or mixed | ✓ | full | x | x | x | 2 | 1-2 |
| LBF | mixed | ✓ | full, partial | varies | ✓ | ✓ | 1-n | 1 |
| MPE | teams | ✓ | full, partial | ✓ | ✓ | ✓ | 1-n | 1-2 |
| SMAC | coop | x | partial | x | x (addressed by SMACv2) | x | 1-n | about 30, but with the same goal |
| DOTA | teams | ✓ | partial | ✓ | x | x | 10 | about 100, but with the same goal |
| RWARE | comp | ✓ | partial | ✓ | x | x | 1-n | 1 |
| GRF | teams | ✓ | full, partial | ✓ | x | x | 2-22 | 1 |
| Hanabi | coop | x | partial | x | x | x | 2-5 | 1 |
| Overcooked | coop | x | full | ✓ | x | ✓ | 1-n | 1 |
| My Game | complex | ✓ | complex | varies | ✓ | ✓ | 2-n | 3 |

A game is defined by a tuple

* : a set consists of the set of agents of each type. They are the seeker, the hider and the defender, respectively. Let denote , respectively. We must have . Let and .
* the “map” on which the game is played. It is a bounded space in the 2-D lattice.
* the set of locations that are watched by the defenders.
* : the game’s state space. A state consists of the positions of all agents on where no position should appear twice (agents cannot share the same space).
* : the set of possible initial states. The game starts from one of these states.
* : the set of terminal states. The game is said to end once the current state is in . It is natural that we assume . A state is a terminal state if all the seekers or all the hiders are eliminated.
* : the action space which is the same for every agent. It is a set of 5 vectors.

The vectors translate to moving directions in the space . They are “pass”, “left”, “right”, “up” and “down”, respectively.

* : the state transition function (game mechanics).

The function computes as follows. At current state , upon receiving the joint action , it returns the probability for the game to transition into state . Apparently, it must satisfy the axiom of probability.

Alternatively, it may be convenient that we define an equivalent version. , where denotes the set of all probability distribution over the set .

To obtain the next state, the position of any agent will update by performing vector addition on the position and the action. The new position must still be in , or the action is undone. If multiple agents end up on the same spot, uniformly randomly undo agents’ actions to resolve conflicts. If a seeker captures (comes within a Manhattan distance of 1) a hider, the hider is eliminated and removed from . Eliminated agents can be represented with a position outside the map. The same thing happens to the seeker when they get caught by a defender. Eliminated agents only perform the “pass” action.

* : a set of reward functions, one for each agent.

For every hider captured, the seekers that capture (is adjacent to) the hider evenly share 1 point, and the hider gets -1. The rewards are assigned similarly when defenders capture a seeker.

* : a set of observation spaces, one for each agent. Each element in contains positions of some agents.
* : a set of observation functions, one for each agent.

It shows the positions of all agents within a certain distance. The defenders and hiders, however, can see anything in the watched area . Depending on the observability of the game, seekers may see each other and so can the defender/hider team. It is assumed that all agents know what and look like.

Starting with a randomly chosen , at each timestep with the current state , each agent receives observation and chooses an action forming a joint action . The game then moves into the next state , randomly sampled from the distribution (, ) and a reward is assigned to each agent. If , the game ends. Otherwise, the game continues in the next timestep.

The goal is to find the optimal strategy (policy) for any agent to pick actions given the observations they receive from the environment, such that their expected reward is maximized.

Currently, many games well studied in RL don’t require a randomised policy to achieve optimality. A general non-cooperative game would not have a pure Nash equilibrium (NE), but a mixed NE (randomized strategy). Traditional RL algorithms where the agent choses the empirically optimal action will fail in these games. Take the following example in my game where a seeker has pushed a hider to a corner.

|  |  |
| --- | --- |
| h |  |
|  | s |

Consider all possible next steps (ignoring the cases where the seeker moves away from the hider). Hider’s choices: right, down, pass. Seeker’s choices: left, up, pass. Reminder: If multiple agents move into the same spot, uniformly randomly undo agents’ actions to resolve conflicts. A seeker captures a hider if they come within a Manhattan distance of one.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| → | h |  | → | h |  | → | h |
| s | ← |  |  | s ↑ |  |  | s - |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ↓ |  |  | ↓ | s |  | ↓ |  |
| h | s ← |  | h | ↑ |  | h | s - |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| h - |  |  | h - | s |  | h - |  |
| s | ← |  |  | ↑ |  |  | s - |

We can translate them into a payoff matrix for the next game tic.

|  |  |  |  |
| --- | --- | --- | --- |
|  | left | up | pass |
| right | 0,0 | -1,1 | -1,1 |
| down | -1,1 | 0,0 | -1,1 |
| pass | -1,1 | -1,1 | 0,0 |

We can see there is no single optimal action for neither the seeker nor the hider if we only consider the next step. No matter what the seeker chooses, there is always a way for the hider to counter its opponent and vice versa. In other words, there is no pure NE, so a mixed strategy needs to be adopted by both players to achieve NE. Traditional RL algorithms will not be able to find an optimal policy in this game because they choose actions deterministically.

Let’s talk about coordination. Below is a classic game called “Stag Hunt”, which belongs to a type of games called coordination games in game theory. Two hunters are going to choose what animals to hunt. They can only successfully hunt a stag if they both commit to it, but they can individually catch hare.

|  |  |  |
| --- | --- | --- |
|  | stag | hare |
| stag | 10,10 | 0,8 |
| hare | 8,0 | 5,5 |

Players must decide between stag (payoff dominant) or hare (risk dominant). You might argue that, well if both players are smart, they will choose to hunt for the stag without any communication, because it is Pareto optimal (best in every way possible). What if they must choose one of the two stags? The “Pure Coordination Game”.

|  |  |  |
| --- | --- | --- |
|  | stag 1 | stag 2 |
| stag 1 | 10,10 | 0,0 |
| stag 2 | 0,0 | 10,10 |

Now there is no risk of payoff options. You either both get it or don’t. Because of the symmetry, no logic can rationalize one option over the other. The agents need to agree on what action to take. Sometimes this can be achieved through a headquarter that manages all the agents, but this is not always possible. Agents might belong to different parties that would like to have private control to satisfy their own interests. A control centre is also far less scalable than decentralized execution.

Without a headquarter, coordination strategies need to emerge among the agents themselves though communication. Communication problems in MARL (Comm-MARL) is a rapidly rising topic. Intuitively, it will only make sense for the agents to communicate if the game is not fully competitive. Even if the game if fully competitive, if it involves more than two players, players can form temporary alliances to gain advantage. In a cooperative or partially cooperative game, agents should communicate for two reasons.

1. Share existing knowledge. To share information in the partially observable environment.
2. Coordinate. To build a consensus; to create a plan.

It might be possible that we further narrow it down to a fully observable setting, which eliminates the first reason to communicate but preserves the second.

In Comm-MARL, the agent communication is usually categorized into two types, explicit communication and non-explicit. Explicit communication means agents send information directly through a dedicated channel. Non-explicit communication means agents send information by letting other agents observe their actions. In nature, creatures will have to use non-explicit communication. e.g. waggle dance of bees. Non-explicit may seem more interesting in terms of sociology. However, if we build a multiagent system, it seems always possible to also build a communication channel for them to share information. Otherwise, they must use actions for show which entangles with the state transition. Current studies mainly focus on the explicit communication.

Some research has been done on Comm-MARL in fully cooperative games. There is currently a gap in the non-cooperative setting. I think this is nice as my game is a complex mixed situation. Maybe I can later reuse some of the findings in the studies of my game. I am also thinking about adding back the overseers (the one that call the guards when they see an intruder) into the game as they are an agent solely dedicated to communication. For example, "can agents learn to lie?" can be an interesting question in the partially cooperative setting. Or "how to prevent agents from lying?". Given the means, would agents learn to form a moral compass to punish agents who lie? Later we will see how these concepts apply to my game.

In the next example, we will see how coordination is required for the seekers to catch a hider. Two seekers have pushed a hider to the boundary of a room.

|  |  |  |
| --- | --- | --- |
|  | h |  |
| s |  | s |

If both seekers ignore each other, and one moves horizontally. The hider can sneak through, regardless of what the other seeker does.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| h | ← |  |  | h | ← | s |
| → | s | s - |  | → | s | ↑ |

If both seekers move up, the hider, as shown below, can sneak through, too.

|  |  |  |
| --- | --- | --- |
| s | ↓ | s |
| ↑ | h | ↑ |

The only way that guarantees that the seekers capture the hider, is to that one seeker stands still and the other moves upward. This way, the seekers form a net such that no matter what the hider does, it will be captured by one of the seekers.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| s | h -/← |  |  | s | → | h |
| ↑ |  | s - |  | ↑ |  | s - |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| s | ↓ |  |  | h | ← |  |
| ↑ | h | s - |  | s ↑ |  | s - |

Even if the seekers are smart enough to know this strategy, they still face the symmetry dilemma. They need to agree on who moves up, or the hider has a chance to live. The chance depends on the hider’s policy which is unknown to the seekers. Let’s denote the hider’s policy with a vector where each entry represents the probability that the hider takes the action left, right, down, and pass, respectively (“up” is ignored). If we assign equal rewards to the seekers when they capture the hider, the payoff matrix with expected return is as the following.

|  |  |  |
| --- | --- | --- |
|  | up | pass |
| up |  | 1 |
| pass | 1 |  |

The scenario becomes a coordination game, if .

What if we split rewards to the seekers that capture the hider? We want to give individual rewards because that a seeker captures a hider light-years away grants reward doesn’t make sense and is difficult for the agent to learn what is going on. The following is the individual reward matrix, with uniformly randomly undone actions taken into consideration. The seeker on the left is the row player.

|  |  |  |
| --- | --- | --- |
|  | up | pass |
| up |  |  |
| pass |  |  |

Combined payoff for both seekers moving up:

Hider lives if it moves down.

Combined payoff for both seekers passing:

Hider lives if it passes.

Combined payoff for both seekers coordinating:

Hider is captured with probability 1.

If we distribute the rewards, the agents now have the motivation to lie in communication. They can benefit from lying if the environment is partially observable, or if the agents are unaware of each other’s action space or reward function. This competitiveness makes coordination exceptionally difficult.