# Notes on **Multiagent Learning**

(Chapter 10 of Multiagent Systems by *Gerhard Weiss*)

Rom Parnichkun

August 15, 2022

## 1 Challenges in Multiagent Learning

- Linear increase in the number of agents increases the state space and action space exponentially.
- Credit assignment is difficult for two reasons. One, *delayed feedback* mechanism makes it more difficult to assign credit to a particular action is a sequence; and two, the *structural credit assignment problem* of how to assign credit to a particular agent based on the performance of a set of agents.
- System rewards may not be directly used as an agent's rewards.

## 2 Measures of reward structure

- **Alignment** (factoredness): defines the correlation between an agent's reward and the system's reward. In an environment with high factoredness, an agent that improves their own performance tend to improve system performance.
- **Sensitivity** (learnability): defines how discernible the impact of an action is on an agent's reward function.

### 2.1 Example reward structures

1. **Full system reward**: Each agent receives the full system reward. Resulting in high factoredness but low learnability (because it is not easy to discern which actions led to the system reward).
2. **Local reward**: Each agent receives a portion of the full system reward depending on its state. This results in higher learnability compared to using full system reward but may have a low degree of factoredness.
3. **Difference reward**: Each agent receives a reward based on the difference between the system reward and the system reward that would have resulted had the agent performed some "null" action.

## 3 Reinforcement Learning for Multiagent Systems

### 3.1 Multiagent MDP Formulations

| **State** | **Action** | **Reward** | **MDP** |
|---|---|---|---|
| Full | Joint | Team | $\langle S, A, T, R, \Pi \rangle$ |
| Full | Independent | Team | $\langle S, A, T, R, \Pi_i \rangle$ |
| Full | Independent | Local | $\langle S, A_i, T, R_i, \Pi_i \rangle$ |
| Local | Independent | Team | $\langle S_i, A_i, T, R, \Pi_i \rangle$ |
| Local | Independent | Local | $\langle S_i, A_i, T, R_i, \Pi_i \rangle$ |

Table 1: A non-exhaustive list of potential MDP formulations

### 3.2 Markov Games

Markov games can be thought of extensions of Markov decision processes.

**Definition 3.1** (Markov Games). *The game $G = \langle n, S, A, T, \tau, \pi^1 ... \pi^n \rangle$ is a stochastic game with n players and k states. In each state $s \in S = (s^1, ..., s^k)$ each player $i$ chooses an action $a^i$ from its admissible action set $A^i(s)$ according to its strategy $\pi^i(s)$.*

*The payoff function $\tau(s, a) : \prod_{i=1}^{n} A^i(s) \mapsto R^n$ maps the joint action $a = (a^1, ..., a^n)$ to an immediate payoff value for each player.*

*The transition function $T(s, a) : \prod_{i=1}^{n} A^i(s) \mapsto \Delta^{k-1}$ determines the probabilistic state change, where $\Delta^{k-1}$ is the $(k-1)$-simplex and $T_{s'}(s, a)$ is the transition probability from state $s$ to $s'$ under joint action $a$.*

### 3.3 Example Algorithms

- **Joint Action Learning**: Same as single agent Q-learning, Q-values are computed for each joint action.
- **Nash-Q Learning**: Similar to single agent Q-learning except that the value of a state is estimated as the sum of the current reward with the value of an agent following a policy at Nash equilibrium.

### 3.4 Solution Concepts

**Solution Concept 3.1** (Nash equilibrium). *When two players play the strategy profile $s = (s_i, s_j)$ belonging to the product set $S_1 \times S_2$, then $s$ is a Nash equilibrium if*

$$P_1(s_i, s_j) \geq P_1(s_x, s_j) \ \forall x \in \{1, ..., n\}$$
$$P_2(s_i, s_j) \geq P_2(s_i, s_y) \ \forall y \in \{1, ..., n\}.$$

*In which, $P_1$ and $P_2$ are the payoffs of player 1 and player 2 respectively.*

**Solution Concept 3.2** (Pareto optimal)**.** *A strategy combination $s = (s_1, ..., s_n)$ for $n$ agents in a game is Pareto optimal if there does not exist another strategy combination $s'$ for which each player receives at least the same payoff $P_i$ and at least one player $j$ receives a strictly higher payoff than $P_j$.*

## 4 Swarm Intelligence

- **Ant Colony Optimization**: Based on the concept of pheromones. Each agent independently finds the goal, but in the process leaves a trail of pheromones, which attract other ants.
- **Bee Colony Optimization**: Similar to ant colony optimization. But instead of using pheromones (which can be thought of as an indirect medium of communication between ants), bees are able to signal to other bees the location of the food source; in other words, bee colony optimization employ direct communication.