Multi-agent Reinforcement Learning

Maxime Toquebiau

11th of March 2021

Agent
A computational system able to perform actions in his environment based on information received from the environment. (Panait and Luke, 2005)

Multi-agent environment
One in which there are more than one agent. Agents are able to interact with one another and influence different parts of the environment. This may lead to dependency relationships between agents.

Multi-agent learning

The application of machine learning to problems involving multiple agents. These problem are by definition more complex:

- Multiple agents ⇒ larger search space
- Multiple learners ⇒ non-stationarity
- Small changes in learned behaviors ⇒ big changes in the properties of the MAS

Nash equilibrium
In game theory, it is a set of strategies (one for each agent) such that no agent can gain by changing their strategy, as long as all the other agents keep theirs unchanged.
Thus it forms a stable state of the system.

Pareto efficiency
An outcome is Pareto efficient if there is no other outcome that makes at least one person better off without leaving anyone worse off.

Markov Games

Def: Also called **stochastic games**, they're an extension of the MDP to the multi-agent case. Formally, a Markov game is defined as the tuple $\langle N, S, A, R, T \rangle$:

- N the set of n agents,
- S the set of states,
- $A = \{A_1, ..., A_n\}$ the collection of action sets, A_i is the set of actions available to agent i,
- $R = \{R_1, ...R_n\}$ where $R_i : S \times A \to \mathbb{R}$ is the reward function of agent i,
- $T : S \times A^n \times S \to [0, 1]$, the state transition function.

Zero-sum game

A zero-sum game is one where all agents' gains and losses sum to 0. In a 2-player zero-sum game, one agent's gain is always the other agent's loss.
The optimal strategy will be the one which minimizes the opponents' payoff.

General-sum game

The sum of gains and losses isn't restricted to 0. Thus, agents no longer necessarily have opposite interests.
The notion of "optimality" loses its meaning since each agent's payoff depends on other agents' choices. The solution to such a game will be a Nash equilibrium (Hu and Wellman, 1998).

Game theory	Reinforcement learning
Player	Agent
Strategy	Policy
Best response	Greedy policy
Utility/Payoff	Reward
State	State
Move	Action
Coalition	Team
Collective	System

Team learning VS **Concurrent learning**

Centralized VS **Decentralized**

Cooperative VS **Competitive**

Credit Assignment

Def: How to distribute the reward obtained at a team level to the individual learners.

Distribution rules: (Panait and Luke, 2005)

- **Global reward:** Split the team reward equally to each of the learners
- **Local reward:** Each agent gets its own individual reward based on his own individual behavior
- **Observational reinforcement:** Reward obtained by observing other agents and imitating their behavior (Mataric, 1994)
- **Vicarious reinforcement:** Small reward received whenever other agents are rewarded (Mataric, 1994)

Aristocrat Utility (AU)

Def: The difference in world utility between the agent's action and the average action (Wolpert and Tumer, 2002).
The AU of an agent i is defined as:

$$u_i(s, \vec{a}) = U(s, \vec{a}) - \sum_{\vec{a}' \in \vec{A}} Pr[\vec{a}'] U(s, \vec{a}_{-i}, \vec{a}'_i),$$

with $\boldsymbol{u_i(s, \vec{a})}$: agent i's reward in state s,
U: the world utility (i.e. global reward),
$\boldsymbol{\vec{a}}$: joint action of all agents in the MAS,
$\boldsymbol{\vec{a}_{-i}}$: joint action of all agents except i,
$\boldsymbol{Pr[\vec{a}']}$: probability that $\vec{a}'$ happens.

Wonderful Life Utility (WLU)

Def: The change in world utility that would have arisen if the agent "had never existed" (Wolpert and Tumer, 1999; 2002).
The WLU of an agent i is given by:

$$u_i(s, \vec{a}) = U(s, \vec{a}) - U(s, \vec{a}_{-i}, CL(\vec{a}_i)),$$

with $\boldsymbol{u_i(s, \vec{a})}$: agent i's reward in state s,
U: the world utility (i.e. global reward),
$\boldsymbol{\vec{a}}$: joint action of all agents in the MAS,
$\boldsymbol{\vec{a}_{-i}}$: joint action of all agents except i,
$\boldsymbol{CL(\vec{a}_i)}$: agent i's action clamped (e.g. replaced by $\vec{0}$, or averaged).

Shapley Value

Def: The average of the marginal contributions of an agent i in all the possible different coalitions (Shapley, 1953).
In a coalition game defined by a set N of n agents and $v(S)$ giving the worth of any coalition S, the Shapley value of agent i is:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

$$\Leftrightarrow \varphi_i(v) = \frac{1}{n} \sum_{S \subseteq N \setminus \{i\}} \binom{n-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S))$$

$$\Leftrightarrow \varphi_i(v) = \frac{1}{\text{number of players}} \sum_{\text{coalitions excluding i}} \frac{\text{marginal contribution of } i \text{ to coalition}}{\text{number of coalitions excluding } i \text{ of this size}}$$

Advantages:

Desirable properties:

- **Efficiency:** $\sum_{i \in N} \varphi_i(v) = v(N)$
- **Symmetry:** if $v(S \cup \{i\}) = v(S \cup \{j\})$ then $\varphi_i(v) = \varphi_j(v)$
- **Linearity:** $\varphi_i(v + w) = \varphi_i(v) + \varphi_i(w)$
- **Fairness:** if a player i is *null* $(v(S \cup \{i\}) = v(S))$ then $\varphi_i(v) = 0$

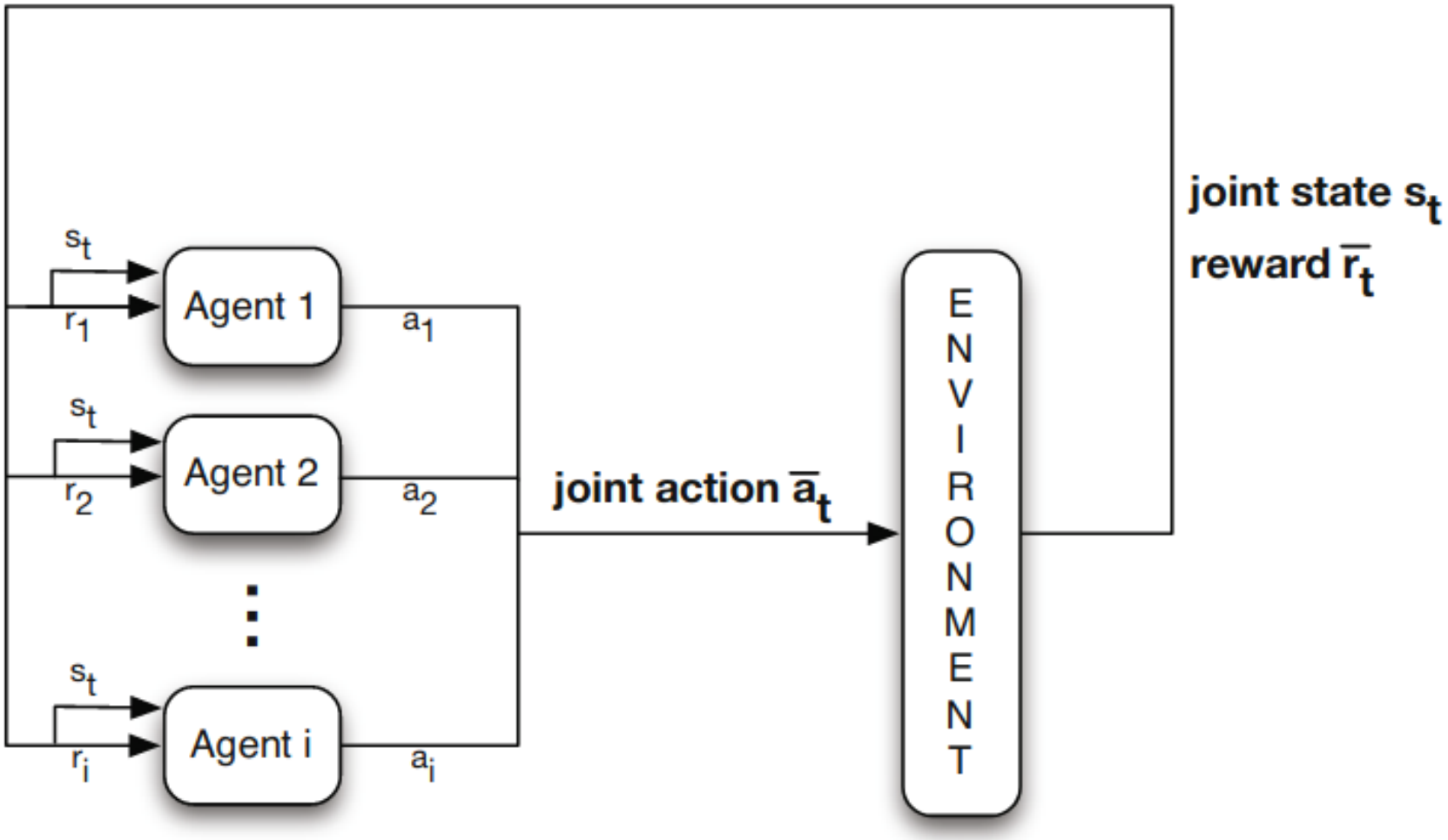
Disadvantage:

Exponential computation time

Application examples:

Network design (Anshelevich et al., 2008; Michalak et al., 2014), MADRL (Shapley Q-value DDPG by Wang et al., 2020)

Multi-agent RL



Nowé, Vrancx and De Hauwere, 2012

Sharing information

- *Multi-agent Reinforcement Learning: Independent vs Cooperative Agents*, Ming Tan, 1993.
  Three cases of study on cooperation in MAS:
  - Sharing sensation,
  - Sharing policies and episodes,
  - Joint tasks.

Opponent modelling

- *Joint Action Learner* (JAL; Claus and Boutillier, 1998): keeping track of the empirical frequency of play for the possible joint actions of the other agents in order to calculate the expected Q-values of a state.

Assuming the other agents' behavior

- *Minimax Q-learning* (Minimax-Q; Littman, 1994): in a 2-player zero-sum game, assuming the opponent will take the action minimizing the agent's payoff.
- *Friend-or-foe Q-learning* (FF-Q; Littman, 2001): opponents marked either as **friends**, who maximize the agent's payoff, or **foes**, who minimize it.
- *Nash Q-learning* (Nash-Q; Hu and Wellman, 1998; 2003): at each stage, constructing a payoff matrix with the estimated Q-values of all agents and assuming agents will play according to a **Nash equilibrium**.
- *Correlated Q-learning* (CE-Q; Greenwald et al., 2003): generalizing Friend-or-foe Q-learning and Nash Q-learning.

Learning to coordinate

Coordination isn't always needed, why not learn when and how to coordinate.

- *Coordinated RL* (Guestrin et al., 2002),
- *Sparse cooperative Q-learning* (Kok and Vlassis, 2004; 2006): representing dependencies between agents in Coordination graphs, thus using the joint actions only when agents explicitly need to coordinate.

Adaptation

- *Win or Learn Fast* (WoLF; Bowling and Veloso, 2002): take larger gradient steps when losing than when winning in order to escape quickly losing policies.
- *AWESOME* (Conitzer and Sandholm, 2005): try to adapt to the opponents' strategies when they appear stationary, otherwise use a precomputed equilibrium strategy.

No-regret

Def: Regret measures how much worse an algorithm performs compared to the best static strategy. Thus, **no-regret** aims at minimizing expected regret.

- *Generalized Infinitesimal Gradient Ascent* (GIGA; Zinkevich, 2003),
- *GIGA-WoLF* (Bowling, 2005).

- Previously presented algorithms have strict limitations:
  - stateless tasks (JAL, GIGA, GIGA-WoLF, AWESOME),
  - limited to zero-sum games (Minimax-Q),
  - convergence is limited to the existence of some equilibrium (Nash-Q, CE-Q),
  - limited to fully observable environments,
  - designed for only two agents;
- Difficulties for generalizing to different tasks,
- Complex architectures/algorithms,
- Low scalability.

- **Stability** (i.e. convergence to a stationary policy):
  $\Rightarrow$ reduce non-stationarity
- **Adaptation**
- **Robustness**
- Communicated sensations must be selective as they increase complexity and come with a communication cost (Tan, 1993)

Thank you!