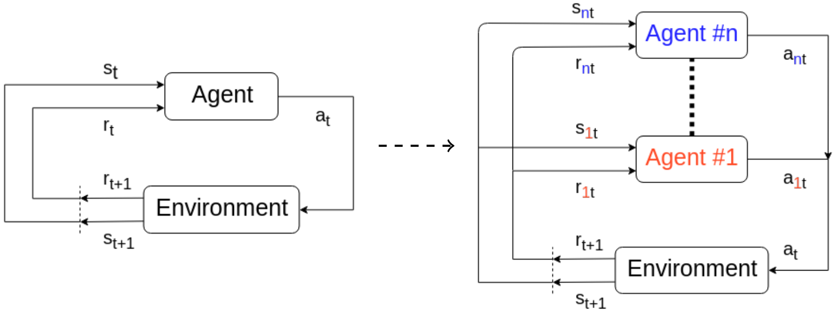
Multi-agent reinforcement learning

Grishin A., Fritsler A.

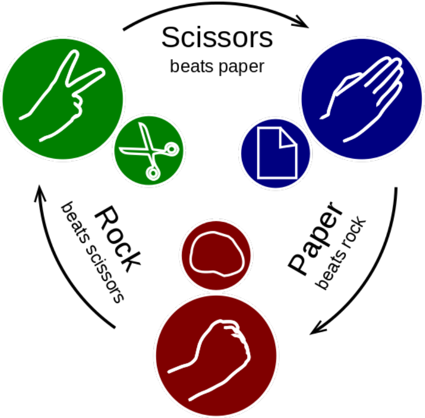
May 14, 2018

Informal definition

Group of autonomous, interacting entities sharing a common environment



Informal definition



[Guild] [Krieg]: any guild ones?
[Eliana] whispers: i havent found this raid =(
[Giolas] whispers: I'm doing my 1st 5 talent points in holy
Chaoswarrior has joined the raid group
Speedo brushes up against Pitbull and farts loudly.
[Mako] yells: too much lag
[Lighthawk] yells: DAMN LAG!!!
Palie has gone offline.

Real-world analogies

- Traffic
- Economy
- Markets
- Workplace
- Sports
- Family

Task-specific skills

- Compete
- Cooperate
- Coordinate
- Communicate
- Predict actions
- Negotiate

Practical tasks

- Distributed control
- Robotic teams
- Automated trading
- Resource Management
- The discovery of communication and language
- Multi-player games
- ...

Formal definition

Stochastic (Markov) game

$$\langle \mathcal{S}, \mathcal{A}^1, \dots, \mathcal{A}^N, r^1, \dots, r^N, p, \gamma \rangle$$

- $\mathcal{S}$ - state space
- $\mathcal{A}^j$ - action space of agent j
- $r^j : \mathcal{S} \times \mathcal{A}^1 \times \dots \times \mathcal{A}^N \to \mathbb{R}$ - reward function of agent j
- $p : \mathcal{S} \times \mathcal{A}^1 \times \dots \times \mathcal{A}^N \to \Omega(\mathcal{S})$ - transition probability map of env.
- γ - discount factor

Value function

$$v^j_{\boldsymbol{\pi}}(s) = v^j(s; \boldsymbol{\pi}) = \sum_0^{\infty} \gamma^t \mathbb{E}_{\boldsymbol{\pi}, p}[r^j_t | s_0 = s, \boldsymbol{\pi}]$$

$$v^j_{\boldsymbol{\pi}}(s) = \mathbb{E}_{\mathbf{a} \sim \boldsymbol{\pi}}[Q^j_{\boldsymbol{\pi}}(s, \mathbf{a})] \qquad Q^j_{\boldsymbol{\pi}}(s, \mathbf{a}) = r^j(s, \mathbf{a}) + \gamma \mathbb{E}_{s' \sim p}[v^j_{\boldsymbol{\pi}}(s')]$$

Types of games

Fully-cooperative game

$$r^1 \equiv r^2 \equiv \ldots \equiv r^N$$

e.g. football players (from one team)

Fully-competitive game (Zero-sum stochastic game)

$$r^1 + r^2 + \ldots + r^N \equiv 0$$

e.g. chess, go

Mixed game (General-sum stochastic game)

otherwise

e.g. free-for-all type of games

Types of games

	one agent	many agents
one state	Multi-armed bandit	Static (matrix) game
many states	Single-agent RL	Multi-agent RL

Standard methods appropriateness

Q-learning

Given a finite MDP, the Q-learning algorithm, given by the update rule

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(s_t, a_t)[r_t + \gamma \max_{b \in \mathcal{A}} Q_t(s_{t+1}, b) - Q_t(s_t, a_t)]$$

converges w.p. 1 to the optimal Q-function as long as

$$\sum_t \alpha_t(x, a) = \infty \qquad \sum_t \alpha_t^2(x, a) < \infty$$

Policy gradient

Intuition: unpredictable influence of agent's parameters on opponent → his actions → your rewards

Challenges

- All of problems from single agent
- Other agents unpredictable or non-stationary
- Various settings

Optimal policy: two-player zero-sum stochastic game

Find the strategy for the agent that has the best "worst-case" scenario

For the strategy $\boldsymbol{\pi}^*$ to be **optimal** it needs to satisfy:

$$\pi_1^* = \arg\max_{\pi_1 \in \Omega(\mathcal{A}^1)} \min_{a \in \mathcal{A}^2} \mathbb{E}_{a \sim \pi_1} Q(s, a, o)$$

Similarly:

$$V_1(s) = \max_{\pi_1 \in \Omega(\mathcal{A}^1)} \min_{a \in \mathcal{A}^2} \mathbb{E}_{a \sim \pi_1} Q(s, a, o)$$

$$Q_1(s, a, o) = R_1(s, a, o) + \gamma \sum_{s'} p(s, a, o, s') V(s')$$

Minimax Q-learning

Algorithm 1: Mini-max Q

```
Initialize $Q_1(s, a_1, a_2)$, $V_1(s)$ and $\pi_1$;
for Each iteration do
    Choose action $a_1$ from current state $s$ based on strategy;
    Observe $r_1, r_2, a_2$ and $s'$;
    Update $Q_1(s, a_1, a_2)$ ;
    Use linear programming to solve max-min ;
end
```

$$Q_1(s,a,o) \leftarrow (1-\alpha)Q_1(s,a,o) + \alpha(r + \gamma V(s'))$$

$$\pi_1^*(s,a) \leftarrow \underset{\pi_1 \in \Omega(\mathcal{A}^1)}{\arg\max} \min_{a \in \mathcal{A}^2} \mathbb{E}_{a \sim \pi_1} Q(s,a,o)$$

Minimax Q-learning

Properties

- Need to observe other agent's action
- Slow learning
- Opponent-independent algorithm
- Converging, but not rational

Single-stage Nash Equilibrium

Battle of sexes

		Pat	
		Opera	Fight
Chris	Opera	1, 2	0, 0
	Fight	0,0	1,2

No person can single-handedly change his/her action to increase their respective payoffs

NE: (Opera, Opera), (Fight, Fight)

Nash equilibrium

Simple games

Nash equilibrium is represented as a set of N policies $\boldsymbol{\pi}_* = \{\pi^1_*, \ldots, \pi^N_*\}$ such that $\forall \pi^j \in \Omega(\mathcal{A}_j)$, it satisfies:

$$v^j(s;\boldsymbol{\pi}_*) = v^j(s;\pi^j_*;\boldsymbol{\pi_*}^{-j}) \geq v^j(s;\pi^j,\boldsymbol{\pi}_*^{-j}),$$

where $\boldsymbol{\pi}_*^{-j} = \{\pi^1_*, \ldots, \pi^{j-1}_*, \pi^{j+1}_*, \ldots, \pi^N_*\}$ - joint policy of all players except j.

Theorem [Fin64]

Every general-sum discounted stochastic game possesses at least one equilibrium point in stationary strategies.

Nash Q-learning

Algorithm 2: Nash Q-learning

```
Initialize $Q_i(s, a_1, \dots, a_n)$;
for Each iteration do
    Choose action $a_t^i$ from current state $s$ based on strategy;
    Observe $r_t^1, \dots, r_t^n; a_t^1, \dots, a_t^n$ and $s'$;
    for $j = 1, \dots, n$ do
        $Q_{t+1}^j(s, a^1, \dots, a^n) = (1 - \alpha_t)Q_t^j(s, a^1, \dots, a^n) + \alpha_t(r_t^j)$
    end
    Use linear programming to find Nash Equilibrium ;
end
```

Properties

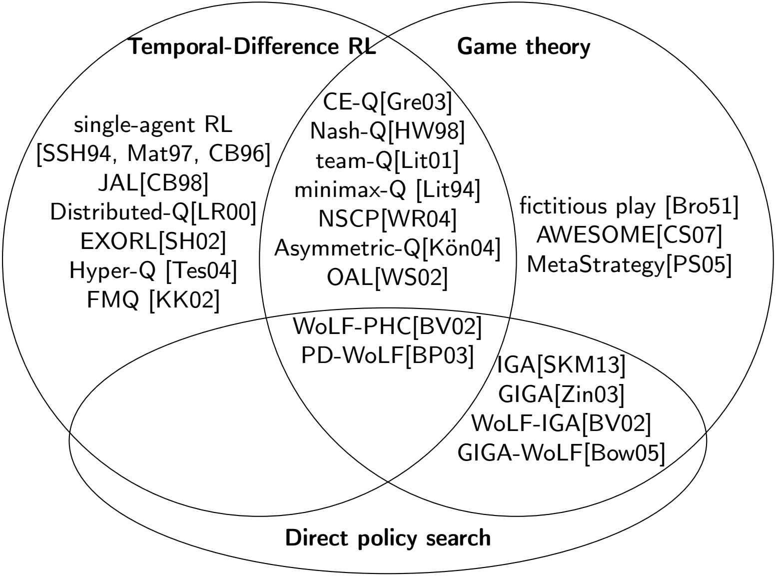
- For 1-player game (MDP), Nash-Q is simple maximization – Q-learning
- For zero-sum games, Nash-Q is Minimax-Q – guaranteed convergence
- Nash equilibrium is not unique → convergence is not guaranteed

Outline

- Theory of MDP and Markov Games are strongly correlated
- Minimax-Q learning is a Q-learning scheme proposed for two-player ZS games
- Minimax-Q is very conservative in its action, since it chooses a strategy that maximizes the worst-case performance of the agent
- Nash-Q is developed for multi-player, general-sum games but converges only under strict restrictions
- FFQ relaxes the restrictions (uniqueness) a bit, but not much

Taxonomy of classic approaches

- Reward function
    - Cooperative
    - Competitive
    - Mixed
- Degree of agent awareness
    - Independent
    - Aware
    - Tracking
- Homogeneity
    - Homogeneous
    - Heterogeneous
- Prior knowledge
    - Model-free
    - Model-based
- Agent's input



Temporal-Difference RL

Game theory

single-agent RL
[SSH94, Mat97, CB96]
JAL[CB98]
Distributed-Q[LR00]
EXORL[SH02]
Hyper-Q [Tes04]
FMQ [KK02]

CE-Q[Gre03]
Nash-Q[HW98]
team-Q[Lit01]
minimax-Q [Lit94]
NSCP[WR04]
Asymmetric-Q[Kön04]
OAL[WS02]

fictitious play [Bro51]
AWESOME[CS07]
MetaStrategy[PS05]

WoLF-PHC[BV02]
PD-WoLF[BP03]

IGA[SKM13]
GIGA[Zin03]
WoLF-IGA[BV02]
GIGA-WoLF[Bow05]

Direct policy search

References I

[Bow05] Michael Bowling, *Convergence and no-regret in multiagent learning*, Proceedings of NIPS-2005 (2005), 209–216.

[BP03] Bikramjit Banerjee and Jing Peng, *Adaptive policy gradient in multiagent learning*, Proceedings of the second international joint conference on Autonomous agents and multiagent systems - AAMAS '03 (2003), 686.

[Bro51] G W Brown, *Iterative Solutions of Games by Fictitious Play, m in Activity Analysis of Production and Allocation, ed, by T. Koop# mans*, New York: Wiley **374** (1951), 376.

[BV02] Michael Bowling and Manuela Veloso, *Multiagent learning using a variable learning rate*, Artificial Intelligence **136** (2002), no. 2, 215–250.

References II

[CB96] Robert Crites and Andrew Barto, *Improving Elevator Performance Using Reinforcement Learning*, Advances in Neural Information Processing Systems 8 **8** (1996), 1017–1023.

[CB98] Caroline Claus and Craig Boutilier, *The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems*, AAAI/IAAI (1998), no. 746, 752.

[CS07] Vincent Conitzer and Tuomas Sandholm, *AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents*, Machine Learning **67** (2007), no. 1-2, 23–43.

[Fin64] A. M. Fink, *Equilibrium in a stochastic n-person game*, Hiroshima Mathematical Journal **28** (1964), no. 1, 89–93.

References III

[Gre03] Amy Greenwald, *Correlated- Q Learning*, Proc. of the 20th International Conference on Machine Learning **1** (2003), 1–30.

[HW98] Junling Hu and Michael P Wellman, *Multiagent reinforcement learning: Theoretical framework and an algorithm*, Proceedings of the fifteenth international conference on machine learning **242** (1998), 250.

[KK02] Spiros Kapetanakis and Daniel Kudenko, *Reinforcement learning of coordination in cooperative multi-agent systems*, Proceedings of the 18th National Conference on Artificial Intelligence (2002), 326–331.

[Kön04] V Könönen, *Asymmetric multiagent reinforcement learning*, Web Intelligence and Agent Systems **2** (2004), no. 2, 105–121.

References IV

[Lit94] Michael L Littman, *Markov Games As a Framework for Multi-agent Reinforcement Learning*, Proceedings of the Eleventh International Conference on International Conference on Machine Learning (San Francisco, CA, USA), ICML'94, Morgan Kaufmann Publishers Inc., 1994, pp. 157–163.

[Lit01] Michael L. Littman, *Value-function reinforcement learning in Markov games*, Cognitive Systems Research **2** (2001), no. 1, 55–66.

[LR00] Martin Lauer and Martin Riedmiller, *An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-Agent Systems (2000)*, Seventeenth International Conference on Machine Learning (2000), 535–542.

References V

[Mat97] Maja J Mataric, *Reinforcement Learning in the Multi-Robot Domain*, Robot Colonies, vol. 4, 1997, pp. 73–83.

[PS05] Rob Powers and Yoav Shoham, *New Criteria and a New Algorithm for Learning in Multi-Agent Systems*, Advances in Neural Information Processing Systems 17, 2005, pp. 1089–1096.

[SH02] Nobuo Suematsu and Akira Hayashi, *A multiagent reinforcement learning algorithm using extended optimal response*, Proceedings of the first international joint conference on Autonomous agents and multiagent systems part 1 - AAMAS '02 (New York, New York, USA), ACM Press, 2002, p. 370.

References VI

[SKM13] Satinder Singh, Michael Kearns, and Yishay Mansour, *Nash convergence of gradient dynamics in general-sum games*, Journal of Chemical Information and Modeling **53** (2013), no. 9, 1689–1699.

[SSH94] Sandip Sen, Mahendra Sekaran, and John Hale, *Learning to coordinate without sharing information*, Proceedings of the National Conference on Artificial Intelligence (1994), 426–431.

[Tes04] Gerald Tesauro, *Extending Q-Learning to General Adaptive Multi-Agent Systems*, Advances in neural information processing systems **16** (2004), 871–878.

References VII

[WR04] Michael Weinberg and J.S. Rosenschein, *Best-response multiagent learning in non-stationary environments*, Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004. (2004), 506–513.

[WS02] Xiaofeng Wang and Tuomas Sandholm, *Reinforcement learning to play an optimal Nash equilibrium in team Markov games*, Advances in Neural Information Processing Systems **15** (2002), 1571–1578.

[Zin03] Martin Zinkevich, *Online Convex Programming and Generalized Infinitesimal Gradient Ascent*, Machine Learning **20** (2003), no. February, 421–422.