University of Stuttgart

Reinforcement Learning

Multi-Agent Reinforcement Learning

Daniel Hennes

03.07.2017

University Stuttgart - IPVS - Machine Learning & Robotics

Multi-agent reinforcement learning

Multi-agent systems have many applications, e.g.

- teamwork (sports, search & rescue robots, sensor nets)
- scheduling (job shops)
- trading (auctions)
- simulations (military, economical)

Why many agents?

- Independent from the application:
  - many applications require a decentralized and adaptive (learning) system
  - a centralized solution is computationally hard
  - environments are dynamic / unpredictable
- Dependent on the application:
  - agents may need to cooperate on a common task, or coordinate their actions for the best result, or they may be (partially) competitive
  - what is it that individual agents need to do?

Definition of multi-agent learning

Peter Stone and Karl Tuyls:

> *The study of multiagent systems in which one or more of the autonomous entities improves automatically through experience*

- Common interest:
  - e.g. robot teams
- Conflicting interest:
  - e.g. board games

Several multi-agent learning paradigms

- **MARL towards individual utility**
- **MARL towards social welfare**
- Co-evolutionary learning
- Swarm Intelligence
- Adaptive mechanism design

Key issues

- Multiple agents need to learn to find optimal solutions by acting autonomously
  - what are optimal solutions in this case?
  - how do we learn those?
- Multi-agent systems do not have the Markov property, because information concerning other agents is generally missing (unobservable)
- Even in fully cooperative systems, communication is expensive & unreliable

Hernandez-Leal et al. (2017):

> *The key challenge in multi-agent learning is learning a best response to the behaviour of other agents, which may be non-stationary: if the other agents adapt their strategy as well, the learning target moves.*

The multi-agent learning problem

- A precise problem formulation is (*was?*) still lacking:
  If multi-agent learning is the answer, what is the question?, Shoham et al. (2006)
- Some MAL objectives:
  - learning should converge to a stationary strategy
  - in "self-play" learning (all agents use same learning algorithm), learners should jointly converge to an equilibrium strategy
  - learning should achieve payoffs as good as a best–response to other agents' strategies (zero regret)
  - learning should guarantee a minimum payoff (worst case bound)

Game theory

- **Strategic interactions** between agents can be modelled using game theory
- Game theory formalizes these interactions
  - each agent has a set of actions
  - a strategy gives a probability to each action
  - joint actions lead to a payoff to each agent
- Given *fully rational* players with *full information* on the game:
  game theory can predict the strategies each agent will use
- Central solution concept: **Nash equilibrium**

Game theory

- A game is specified by: **players** $(1 \dots n)$, **actions**, and (expected) **payoff matrices** (function of joint–actions)

		Player 2		
		Rock	Paper	Scissor
Player 1	Rock	0, 0	-1, 1	1, -1
	Paper	1, -1	0, 0	-1, 1
	Scissor	-1, 1	1, -1	0, 0

- If payoff matrices are *"identical"*, players are **cooperative**, else **non-cooperative** (*zero-sum* = purely competitive)

Game theory: basic terminology

- Games with no states: *matrix games*
- Games with states: stochastic games, *Markov games* (state transitions are functions of joint–actions)
- Games with simultaneous moves: *normal form*
- Games with alternating turns (tree): *extensive form*
- No. of rounds $= 1$: *one-shot game*
- No. of rounds > 1: *repeated game*
- deterministic policy: *pure strategy*
- non-deterministic policy *mixed strategy* e.g. $\Pr(R, P, S) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

States and agents

	single state	many states
single agent	**multi-armed bandits**	**markov decision process**
many agents	**normal-form game**	**markov game**

Stochastic vs. matrix games

- A **Markov game** generalizes MDPs to multiple agents
  - set of n agents / players
  - finite state space $s \in \mathcal{S}$
  - joint–action $\boldsymbol{a} = \langle a_1(s), \ldots, a_n(s) \rangle$
  - stationary reward distribution $r_1(s, \boldsymbol{a}, s')$
  - stationary transition probabilities $\Pr\{s' \mid s, \boldsymbol{a}\}$
- A **matrix game** has no state information, only actions and payoffs ($|\mathcal{S}| = 1$)

Basic analysis

- Agent i's mixed strategy $\boldsymbol{\pi}_i$ is a *best–response* to others' $\boldsymbol{\pi}_{-i}$ if it maximizes payoff given $\boldsymbol{\pi}_{-i}$
- $\boldsymbol{\pi}_{-i}$ is a *dominant strategy* if it maximizes payoff regardless of what others do
- A joint strategy $\boldsymbol{\pi}$ is an *equilibrium* if each agent's strategy is simultaneously a best–response to everyone else's strategy, i.e., no incentive to deviate
- *Nash equilibrium* is the main one, but there may be exponentially many of them, and very hard to compute
- *Equilibrium selection* is a big problem
  (players need to agree on which equilibrium to choose)

Assumptions in normal form games

- Game specification is fully known:
  - actions and payoffs are fully observable by all players
  - (this is not scalable!)
- Players act "simultaneously", i.e. without observing actions of others
- Assume no communication between players, or it doesn't affect play (communication is "cheap talk")
- Basic analysis assumes the game is only played once (called *one-shot*)

Example: battle of the sexes

		Bob: Football	Bob: Ballet
Alice	Football	2, 1	0, 0
	Ballet	0, 0	1, 2

- Dominant strategy?
- Nash equlibrium/equilibria?

Example: battle of the sexes

		Bob: Football	Bob: Ballet
Alice	Football	2, 1	0, 0
	Ballet	0, 0	1, 2

- No (iterated) dominant strategy equilibrium
- Two Nash equilibria (Football, Football) and (Ballet, Ballet)
- Third equilibrium is *mixed*

Example: battle of the sexes

		Bob: Football	Bob: Ballet
Alice	Football	2, 1	0, 0
	Ballet	0, 0	1, 2

- Players will only mix when indifferent w.r.t. payoff, given the other player's strategy
- What are the payoffs for the *pure* and *mixed* equilibria?

Fictitious play

- **Fictitious play:** agent observes time–average frequency of other players' action choices, and models:

$$\hat{\pi}_{-i}(s, a) = \frac{\#\text{ times } a \text{ observed in } s}{\#\text{ visits to } s}$$

  agent plays best-response to this model

$$Q(S_t, A_t \mid \hat{\pi}_{-i})$$

- Ignores *non–stationarity* of other agents' strategies
- Variants of fictious play:
  - exponential recency weighting
  - "smoothed" best response ($\sim$ softmax)
  - small adjustment towards best response
  - ...
- One of the simplest *opponent models*

What if all agents use fictitious play?

- Strict Nash equilibrium are absorbing points of fictitious play
- Typical result is limit–cycle behavior of strategies
- In certain cases, product of empirical distributions converge to Nash even though actual play cycles

Learning in Markov games

- Learning is especially important in Markov games, here NE are hard to compute
- What do we know?
  - our own payoff matrix?
  - others' rewards?
  - transition probabilities?
  - others' strategies?
- All of the above might be unknown ...

Learning in Markov games

- "Classic" approaches, adapted from single-agent RL
- **Independent learners** (ignoring others)
  - Q-learning
- **Joint–action learners** (model others)
  - Minimax Q-learning
  - Nash Q-learning
- State-of-the-art: usually in between the two extremes

Independent learners

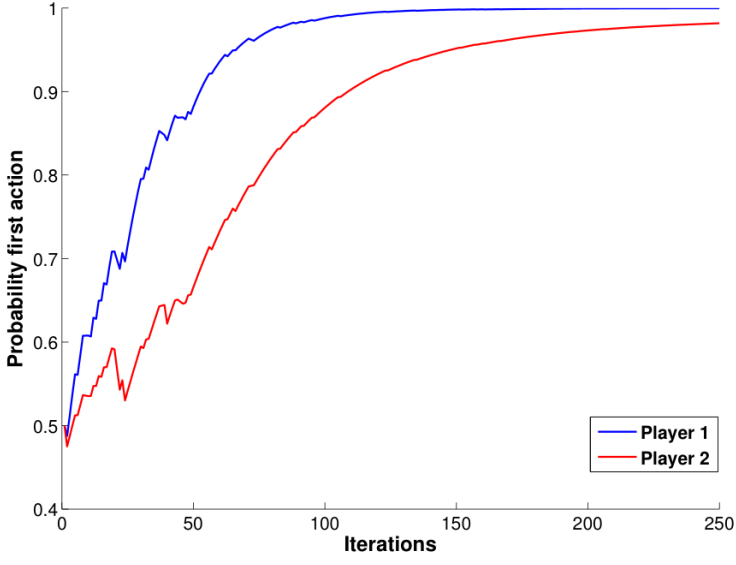
- **Independent learners** mutually ignore each other
- Implicitly perceive interaction with other agents as noise in a stochastic environment
- **Advantages:**
  - straightforward application of single-agent techniques
  - scales easily with number of agents
- **Disadvantages:**
  - convergence guarantees from single–agent setting are lost
  - no explicit means of coordination

Independent learners in normal-form games

- Two *Q-learners* interact in the battle of the sexes
  - learning rate $\alpha = 0.01$
  - softmax with $\tau = 0.2$
- Both players only observe their immediate reward
- Policy is gradually improved

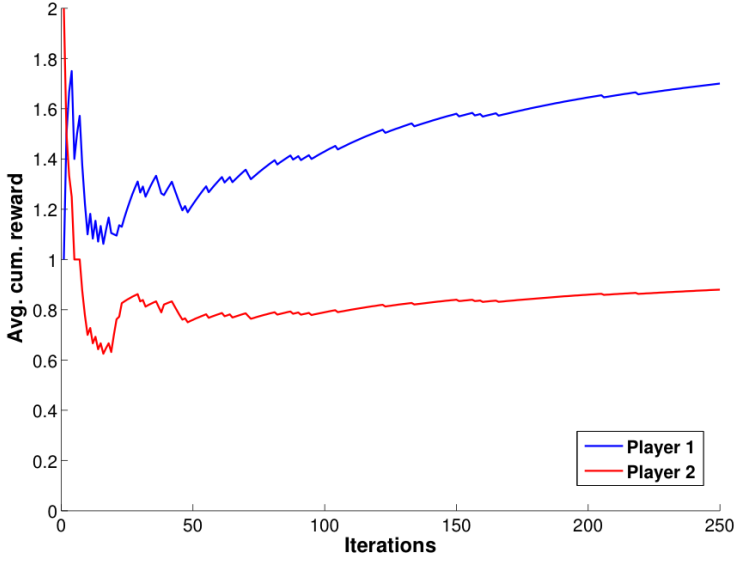
	B	S
B	2, 1	0, 0
S	0, 0	1, 2

Independent learners in normal-form games



	B	S
B	2, 1	0, 0
S	0, 0	1, 2

Independent learners in normal-form games



	B	S
B	2, 1	0, 0
S	0, 0	1, 2

Joint–action learners

- **Joint–action learners** observe the actions of other agents
- Similar to fictitious play, they assume a stationary policy
  - policy is estimated
  - best–response against estimate
- **Advantage:**
  - better means of coordination
  - explicitly taking other agents into account
- **Disadvantages:**
  - need to observe other agents' actions
  - complexity grows exponentially with number of agents

Minimax Q-learning

- **Zero sum:** payoffs balance out, only need to observe own payoff
- Update rule based on joint–action $\langle a_1, a_2 \rangle$:

$$Q(s, a_1, a_2) \leftarrow Q(s, a_1, a_2) + \alpha\big[r + \gamma V(s') - Q(s, a_1, a_2)\big]$$

$$V(s') = \max_{\pi} \min_{a'_2} \sum_{a'_1} Q(s', a'_1, a'_2)\pi(a'_1 \mid s')$$

Minimax Q-learning

- Performs better than naive independent Q–learning
- Converges to Nash
  - under similar conditions as single–agent Q–learning
- Limited to zero–sum games . . .

Q-learning in general–sum games

- Can we extend the algorithm to general–sum stochastic games?
- **Nash Q-learning** is such an extension
  - much worse computational and theoretical properties
  - must learn $Q_i(s, a_1, \ldots, a_n)$ for all states s, joint–actions $\langle a_1, \ldots, a_n \rangle$, and every agent i

$$Q_i(s, a_1, \ldots, a_n) \leftarrow Q_i(s, a_1, \ldots, a_n) + \alpha \big[r + \gamma V_i(s') - Q_i(s, a_1, \ldots, a_n) \big]$$

- $V_i(s')$ is the payoff for player i in the Nash equilibrium
  - *needs to be computed!*
- Assumes all players play the same Nash equilibrium
  - *selection problem!*

Convergence of Nash Q-learning

- **Theoretical guarantees:** Nash-Q converges, ...
  - if every stage game encountered during learning has a global optimum
  - if every stage game encountered during learning has a saddle point
- Both are *very* strong assumptions
- However,
  - can converge in practice without
  - performs better than independent Q-learning

Gradient ascent based approaches

- **Gradient ascent:** update policy directly in the direction of the gradient of the value function
- Examples:
  - Infinitesimal Gradient Ascent (IGA)
  - Generalised IGA (GIGA)
- Main ideas:

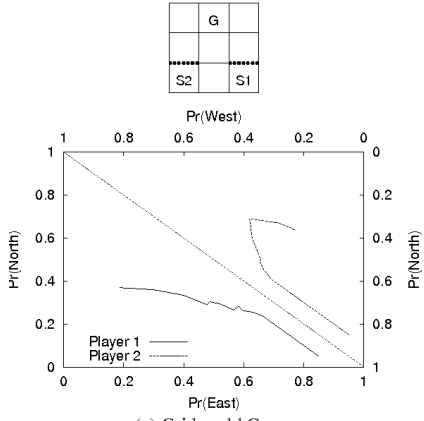
$$\Delta \boldsymbol{\pi}_i \leftarrow \alpha \frac{\partial V(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}_i}$$

$$\boldsymbol{\pi} \leftarrow \mathit{projection}(\boldsymbol{\pi} + \Delta \boldsymbol{\pi})$$

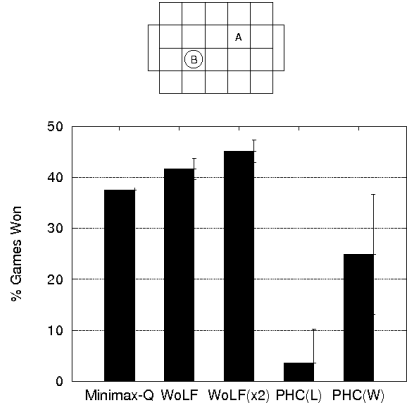
- Only suited for (2-player) matrix games

Gradient ascent based approaches

- Can improve convergence using a *variable learning rate*
- **WoLF:** "win or learn fast"
  - decrease learning rate when performing well,
    increases learning rate when doing badly
  - improves convergence of (G)IGA and Policy Hill-Climbing (variant of Q-learning)
- **Weighted policy learner:**
  - decrease learning rate unless gradient direction changes
  - i.e., learn fast when something unexpected happens



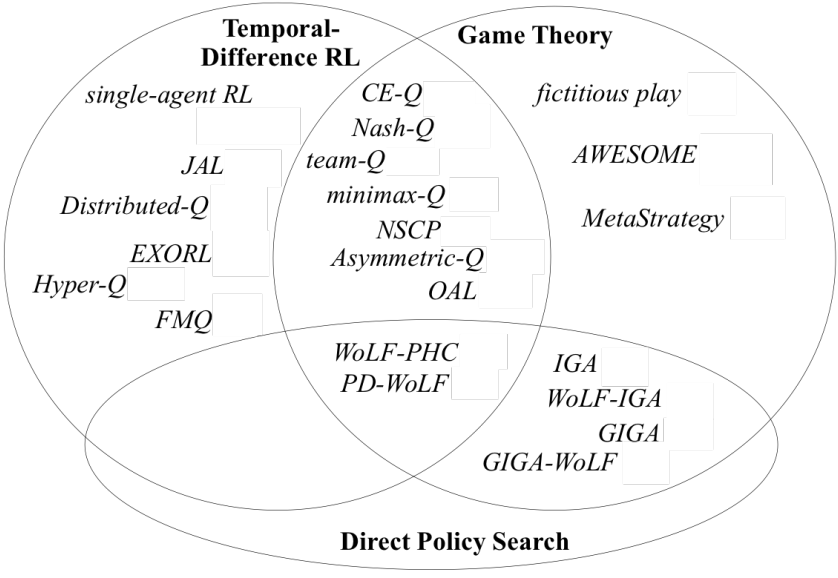
(a) Gridworld Game



(b) Soccer Game

Bowling et al.

Taxonomy of multi-agent learning algorithms



See survey paper by Busoniu et al.

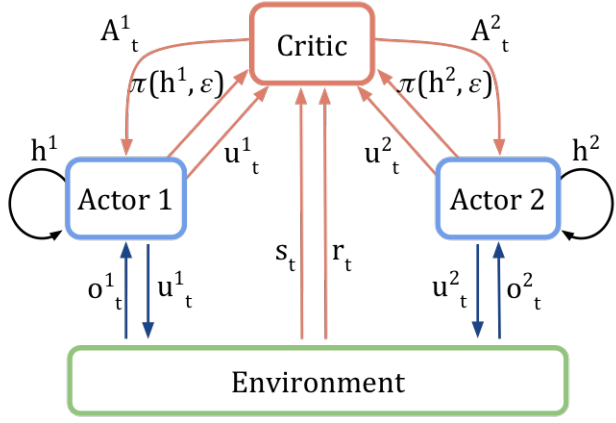
Real-world opportunities

Multi-agent systems where it's hard to do game theory:

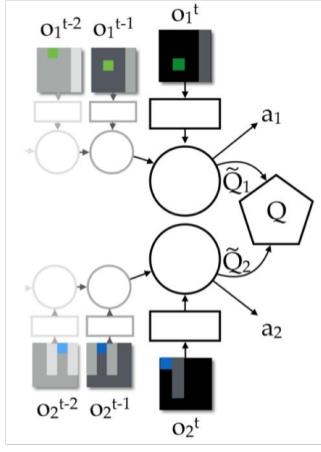
- Electronic marketplaces
- Mobile networks
- Counter-terrorism applications
- Teams of robots
- Video games

Recent advances

- Multi-agent *deep* learning
- Exploiting *centralized* learning
- Learning *opponent models*



Counterfactual multi-agent policy gradients, Jakob Foerster et al. (2017)



Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward, Peter Sunehag et al. (2018)