zalando

REINFORCEMENT LEARNING IN MULTI-AGENT SYSTEMS

MACHINE LEARNING MEETUP

DR. ANA PELETEIRO RAMALLO

29-08-2016

TABLE OF CONTENTS

zalando

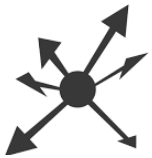
ZALANDO

Zalando is **the largest e-commerce** platform in Europe.

Zalando Tech employs **1000+** people in tech.

Our purpose: **to deliver** award-winning, best-in-class **shopping**

Experiences to our **+15 million** customers.



Radical agility:
- Purpose, autonomy and mastery

zalando

FASHION INSIGHTS CENTER

- Zalando Fashion Insights Centre was founded with the aim of understanding fashion through technology.
- R&D work to organise the world's fashion knowledge.
- We work with one of the richest datasets in eCommerce; products, profiles, customers, purchasing and returns history, online behaviour, Web information and social media data.
- Three main teams:
  - Smart Product Platform
  - Customer Data Science
  - Fashion Content Platform

zalando

MULTI-AGENT SYSTEMS

- Multi-agent Systems (MAS) is the emerging subfield of AI that aims to provide both principles for construction of complex systems involving multiple agents and mechanisms for coordination of independent agents' behaviors.
- Agent: autonomy, social ability, reactivity, pro-activeness
- Increasingly relevant within artificial intelligence.
- Technological challenges require decentralised solutions
    - Robotic soccer, disaster mitigation and rescue, automated driving.
- Dynamic and non-deterministic environments, they need to learn

zalando

COORDINATION IN MULTI-AGENT SYSTEMS

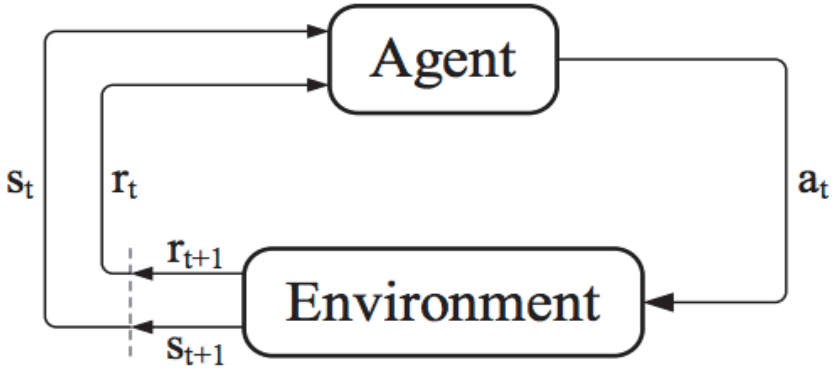
- Improve coordination and cooperation.
- Achieving cooperation and/or in multi-agents systems (MAS) is a challenging issue, particularly when agents are self-interested.
- Tasks that are too complex to solve individually or also when groups perform more efficiently than individuals.
- Designing mechanisms that promote the emergence and maintenance of cooperation for self-interested agents has become a major area of interest in MAS.
- Cooperation and teamwork, including: distributed problem solving; human-robot/agent interaction; multi-user/multi-virtual-agent interaction; coalition formation; coordination
- Several game theory approaches have been used to provide a framework to study cooperation in those cases.

GAME THEORY

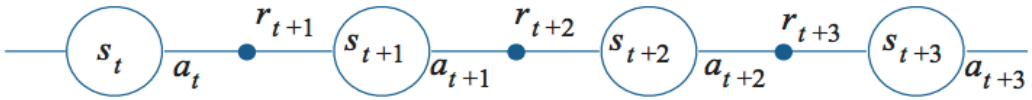
- Discipline that studies the interactions between self-interested agent to model strategic interactions as games.
- How interaction strategies can be designed that will maximise the welfare of an agent in a multi-agent encounter.
- Applications of game theory in agent systems have been to analyse multi-agent interactions, particularly those involving negotiation and coordination.
- Non cooperative games
  - Non-cooperative game is one in which players make decisions independently
  - Thus, while players could cooperate, any cooperation must be self-enforcing.
  - Self-interested agents.
- Stochastic games are defined as non-cooperative games where agents pursue their self-interests and choose their actions independently.

REINFORCEMENT LEARNING (II)

Learning by interacting with the environment: trial and error.
Environment may be unknown, non linear, stochastic and complex



Fundamentals of Multi-Agent Reinforcement Learning. Daan Bloembergen, Daniel Hennes



zalando

REINFORCEMENT LEARNING (II)

- Agent aims to learn a policy to map states to actions

$$\pi_t(s, a) = P(a_t = a | s_t = s)$$

- RL specifies how to change the policy as a result of experience
- Goal: maximize cumulative reward long term (E(Rt))

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

- Exploration (unknown territory) vs. exploitation (known territory)

MARKOV DECISION PROCESS (MDP)

- A Markov decision process is defined by:
  - Set of actions
  - Set of states
  - State transition probabilities (Eq. 1)
  - Reward probabilities (Eq. 2)
  - Discount factor
- If space and actions are finite, then it is a finite MDP.
- If a reinforcement learning task that satisfies the Markov property (Eq. 3), then it called is called a MDP.
- The conditional distribution of the future states of the process only depend only upon the present state.

$$\mathcal{P}^{a}_{ss'} = P(s_{t+1} = s' | s_t = s, a_t = a)$$

Eq. 1

$$\mathcal{R}^{a}_{ss'} = E(r_{t+1} | s_t = s, a_t = a, s_{t+1} = s')$$

Eq. 2

$$P(s_{t+1}, r_{t+1} | s_t, a_t) = P(s_{t+1}, r_{t+1} | s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots, r_1, s_0, a_0)$$

Eq. 3

(MDP II)

- When following a fixed policy π we can define the value of a state s under that policy as in Eq. 1
- Similarly we can define the value of taking action a in state s as in Eq. 2.
- Most of RL are based on estimating the value functions.
- We want to find the policy that maximizes long term reward, which equates to finding the optimal value function (Eq. 3)
- The value of a state under an optimal policy must equal the expected return for the best action from that state (Eq. 4).
- Every MDP has at least one optimal policy.

$$V^{\pi}(s) = E_{\pi}(R_t|s_t = s) = E_{\pi}(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s)$$

Eq. 1

$$Q^{\pi}(s, a) = E_{\pi}(R_t|s_t = s, a_t = a)$$

Eq. 2

$$V^*(s) = \max_{\pi} V^{\pi}(s) \qquad \forall s \in S$$

$$Q^*(s, a) = \max_{\pi} Q^{\pi}(s, a) \qquad \forall s \in S, a \in A(s)$$

Eq. 3

$$V^*(s) = \max_{a \in A(s)} Q^{\pi *}(s, a)$$

Eq. 4

AGENT LEARNING FRAMEWORKS

- There are different theoretical frameworks for the different learning problems.
  - Single-agent: Markov decision processes (MDP)
  - Multi-agent, static (stateless): normal form games
  - Multi-agent, dynamic (multi-state): Markov games

SINGLE AGENT LEARNING

- Can be modeled as a MDP.
- Convergence guarantees.
- E.g., a robot that has to search for cans.
  - Actions: wait, search, recharge
  - States: low, high
- At each such time the robot decides whether it should (1) actively search for a can, (2) remain stationary and wait for someone to bring it a can, or (3) go back to home base to recharge its battery.

Table 3.1: Transition probabilities and expected rewards for the finite MDP of the recycling robot example. There is a row for each possible combination of current state, s, next state, s', and action possible in the current state, $a \in \mathcal{A}(s)$.

$s = s_t$	$s' = s_{t+1}$	$a = a_t$	$\mathcal{P}^a_{ss'}$	$\mathcal{R}^a_{ss'}$
high	high	search	α	$\mathcal{R}^{\text{search}}$
high	low	search	$1-\alpha$	$\mathcal{R}^{\text{search}}$
low	high	search	$1-\beta$	-3
low	low	search	β	$\mathcal{R}^{\text{search}}$
high	high	wait	1	$\mathcal{R}^{\text{wait}}$
high	low	wait	0	$\mathcal{R}^{\text{wait}}$
low	high	wait	0	$\mathcal{R}^{\text{wait}}$
low	low	wait	1	$\mathcal{R}^{\text{wait}}$
low	high	recharge	1	0
low	low	recharge	0	0.

Reinforcement Learning: An Introduction Richard S. Sutton and Andrew G. Barto

MULTI-AGENT LEARNING: MARKOV GAMES

- Agents interact both with the environment and with each other.
- Learning is simultaneous.
- Stochastic n-player games.
- Each state in a stochastic game can be considered as a matrix game with payoff for player i of joint action a in state s determined by Ri (s, <a1, a2, …an>).
- After playing the matrix game and receiving the payoffs, the players are transitioned to another state (or matrix game) determined by their joint action.
- The transition and payoff functions depend on the joint action a=<a1, a2, …,an>

$$\mathcal{R}^i : S \times A^1 \times \cdots \times A^n \mapsto \mathbb{R}$$

$$\mathcal{P} : S \times A^1 \times \cdots \times A^n \mapsto \triangle(S)$$

- In this type of games, performance depends critically on the choice of the other agent.

MULTI-AGENT LEARNING (II)

INDEPENDENT LEARNERS

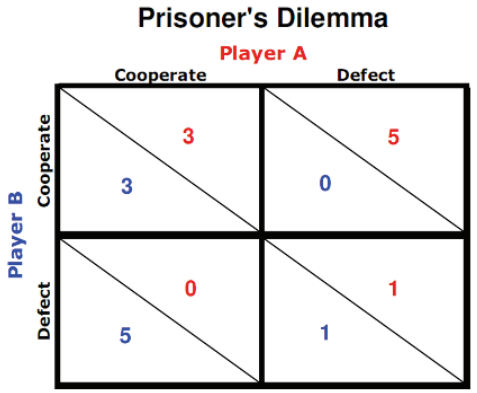
- Ignore other agents.
- Perceive the other agents interactions as noise.
- Adv:
  - Easy to scale
  - Application of single-agent techniques
- Dis:
  - No convergence guarantees
  - Less coordination
- Algorithms:
  - Q-learning
  - Learning Automata

JOINT LEARNERS

- Observe the actions of other agents
- A joint action learner is an agent that learns Q-values Q(s,<a1,a2,…,an>) for joint actions as opposed to individual actions.
- Adv:
  - Better coordination
- Dis:
  - Need to observe other agents behaviour
  - Exponential complexity growth
- Algorithms:
  - Minimax-Q

STATELESS MULTI-AGENTS

- A Markov game where agents are stateless can be reduced to a normal form game.
- All players simultaneously select an action, and their joint action determines their individual payoff
  - One shot interaction
  - Represented as a n-dimensional matrix for n-players
- Player's strategy is defined as a probability distribution over his possible actions
- In this games we have
  - Competitive or zero sum (Matching Pennies)
  - Symmetric games (Prisoner's Dilemma)
  - Asymmetric games (Battle of Sexes)



http://blankonthemap.blogspot.ie/2012/09/optimal-strategies-in-iterated.html

zalando

Q-LEARNING

- Temporal difference (TD) method:
  - Learn directly from experience
  - Agents do not need to know the model of the environment
- Each state-action pair has a corresponding Q-value: represents expected cumulative payoff from performing action in the given state.
- Q-learning updates state-action values based on the immediate reward and the optimal expected return.
- Off-policy: directly learns the optimal value function independent of the policy being followed.
- Exploration vs. exploitation: $\mathcal{E}$-greedy action selection
  - Optimal action a* with probability 1- $\mathcal{E}$
  - Random with $\mathcal{E}$
  - Decrease $\mathcal{E}$ during each episode

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right]$$

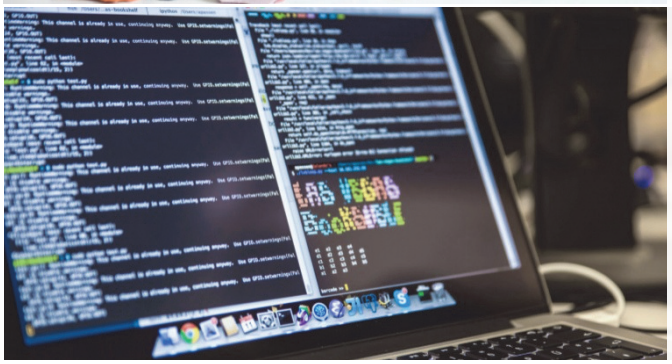
Q-LEARNING

Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
  Initialize s
  Repeat (for each step of episode):
    Choose a from s using policy derived from Q (e.g., ε-greedy)
    Take action a, observe r, s'
    $Q(s, a) \leftarrow Q(s, a) + \alpha\left[r + \gamma \max_{a'} Q(s', a') - Q(s, a)\right]$
    $s \leftarrow s'$;
  until s is terminal

Reinforcement Learning: An Introduction Richard S. Sutton and Andrew G. Barto

RESOURCES

- *Reinforcement Learning: An Introduction. Richard S. Sutton and Andrew G. Barto*
- *T2: Multiagent Reinforcement Learning (MARL). Daan Bloembergen, Tim Brys, Daniel Hennes, Michael Kaisers, Mike Mihaylov, Karl Tuyls*
- *Multi-Agent Reinforcement Learning ALA tutorial. Daan Bloembergen*
- *Reinforcement Learning, Hierarchical Learning, Joint-Action Learners. Alexander Kleiner, Bernhard Nebel*
- *L. Bus¸oniu, R. Babuska, and B. De Schutter, "Multi-agent reinforcement learning: ˇ An overview," Chapter 7 in Innovations in Multi-Agent Systems and Applications – 1 (D. Srinivasan and L.C. Jain, eds.), vol. 310 of Studies in Computational Intelligence, Berlin, Germany: Springer, pp. 183–221, 2010.*
- *GAME THEORY. Thomas S. Ferguson*
- *Game Theory and Decision Theory in Multi-Agent Systems. Simon Parsons, Michael Wooldridge*
- *MULTIAGENT SYSTEMS Algorithmic, Game-Theoretic, and Logical Foundations. Yoav Shoham, Kevin Layton-Brown*
- *Multi-agent Systems: A Survey from a Machine Learning Perspective Peter Stone Manuela Veloso*



zalando

DR ANA PELETEIRO RAMALLO

DATA SCIENTIST

ana.peleteiro@zalando.ie

@PeleteiroAna

29-08-2016