RL 17: Multi-Agent Reinforcement Learning

Michael Herrmann

University of Edinburgh, School of Informatics

17/03/2015

Overview

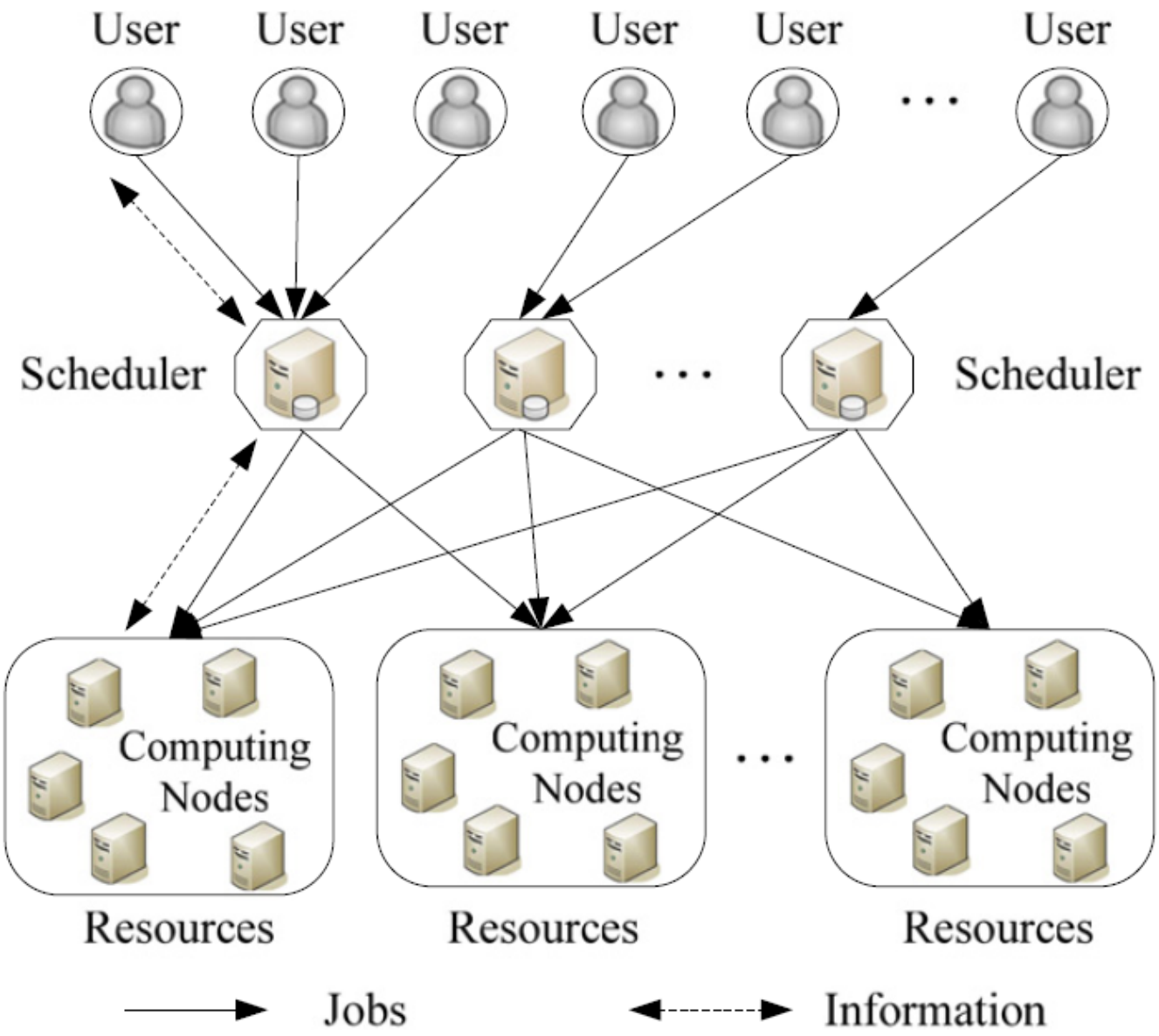
- MARL
- Stateless games
- Markov games
- Decentralised RL

Multi-Agent Reinforcement Learning

- Multi-agent learning is about cooperation or competition: The MDP model may not longer apply
  - if goals are compatible, some degree of coordination may be required
  - if goals are opposed, an optimal solution may no longer exist
- Problems: limited information (including on the presence of other agents), complexity, non-stationarity
- Applications: multi-robot systems, decentralised network routing, distributed load-balancing, traffic, finance, psychology and biology

Ann Nowé, Peter Vrancx, and Yann-Michaël de Hauwere (2012) Game theory and multi-agent reinforcement learning. In: M. Wiering and M. van Otterlo (Eds.): Reinforcement Learning, Springer.

Example: Job Scheduling



MARL vs. distributed RL

- Parallel reinforcement learning: agent learn a single objective collaboratively, e.g.
  - MORL,
  - (hierarchically) divided state space
  - distributed exploration by an RL swarm

  Problem: some agent may have outdated values
  Solution: Use max in the learning rule assuming outdated values are smaller
  This is essentially standard RL with non-standard exploration
- MARL: individual goals and independent decision making capabilities
  - Nash equilibrium: no agent can improve its reward when the other agents retain a fixed policy

Benefits and/or Challenges in MARL

- Speed-up possible by parallelisation
- Experience sharing between agents by communication, mutual teaching or imitation learning
- Scalability: insertion of new agents, robustness vs. failure of some agents
- Exponential complexity in the number of agents
  - sparse interactions
  - experience sharing
- Exploration is as essential as ever, and/but may confuse other agents

Busoniu, L., Babuska, R., & De Schutter, B. (2008)

Stateless Games ("MAB for MA")

- Reward depends on joint actions: $r_k : \mathcal{A} \to \mathbb{R}$
- Zero-sum games or (e.g.) the prisoner's dilemma

$$\begin{array}{ccc} & a_1 & a_2 \\ a_1 & (5,5) & (0,10) \\ a_2 & (10,0) & (1,1) \end{array}$$

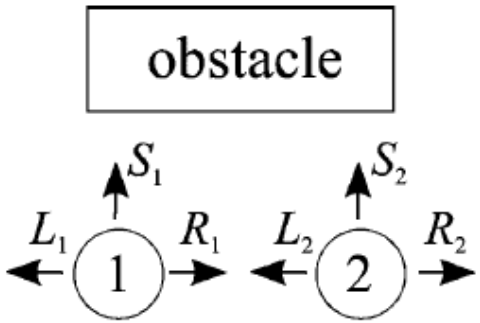
  (a_2, a_2) is a Nash equilibrium (which is not optimal)
- Best response for agent k if
  $r_k(\mathbf{a}_{-k}, a_k) \geq r_k(\mathbf{a}_{-k}, b_k) \;\; \forall b \in \mathcal{A}_k$ or more generally
  $r_k(\pi_{-k}, \pi_k^*) \geq r_k(\pi_{-k}, \pi_k) \;\; \forall \pi_k \in \Pi_k$
- In a Nash equilibrium all agents play their best response
- Stochastic policies can be optimal in MARL, e.g. "matching pennies"

$$\begin{array}{ccc} & a_1 & a_2 \\ a_1 & (1,-1) & (-1,1) \\ a_2 & (-1,1) & (1,-1) \end{array}$$

  best strategy is to choose any action with probability $\frac{1}{2}$

Example: Cooperation

Two agents, three actions each: avoid obstacle, but do not disrupt the formation



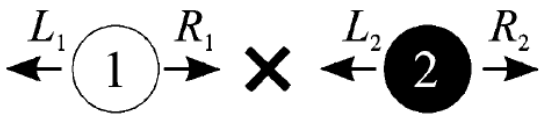
Q	L_2	S_2	R_2
L_1	10	-5	0
S_1	-5	-10	-5
R_1	-10	-5	10

Tie between L-L and R-R. Coordination necessary:

- "social conventions": Agent 1 determines first, agent 2 observes and follows
- communication: agent arriving first tells the other agent (again tie may occur)

Busoniu, L., Babuska, R., & De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 38(2), 156-172.

Example: Competition



Q	L_2	R_2
L_1	0	1
R_1	-10	10

Agent 1 is heading for the goal ($\times$) while avoiding capture by its opponent, agent 2.

Agent 2 aims at preventing agent 1 from reaching the goal, preferably by capturing it.

The agents can only move to the left or to the right.

Mini-max solution for agent one is to move left (or needs to find out what agent 2 is going to do)

Busoniu, L., Babuska, R., & De Schutter, B. (2008)

Reinforcement Learning in Repeated Games

- In MARL, agents do not have full access to the (stochastic) pay-off matrix, i.e. the game is unknown
- Actions of other agents are usually observable
- Goal of learning
  - Nash equilibria
  - joint optimality
  - evolutionary stable strategies

Linear Reward-Inaction

- Adjust action probabilties according to

$$\begin{aligned} p_i(t+1) &= p_i(t) + \eta r(t)(1 - p_i(t)) \quad \text{if } a(t) = a_i \\ p_j(t+1) &= p_j(t) - \eta r(t) p_i(t) \qquad\qquad \text{if } a(t) \neq a_i \end{aligned}$$

  p_i is the probability of playing action a_i, η is a learning rate
- Special case of REINFORCE (Williams, 1992)
- Properties (Sastry et al., 1994)
  - All Nash equilibria are stationary points.
  - All strict Nash equilibria are asymptotically stable.
  - All stationary points that are not Nash equilibria are unstable.

Markov Games

- Markov Games are for MAS like MPDs for single agents
- A Markov game is a tuple $(n, S, A_{1,\ldots,n}, R_{1,\ldots,n}, T)$ where
  - n is the number of agents
  - $\mathcal{S} = \{s_1, ..., s_N\}$ are the states
  - $\mathcal{A}_k$ are the actions of agent k
  - $R_k : \mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_n \times \mathcal{S} \to \mathbb{R}$ are the rewards for agent k
  - $\mathcal{T}$: $\mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_n \times \mathcal{S} \to P(\mathcal{S})$ the action-dependent state transition function.
- Note that rewards and transitions depend on the other agents
- Task: Every agents maximises

$$V_k^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r_k(t+1) \,|\, s(0) = s \right]$$

where r is a realisation of R and $\pi = (\pi_1, \ldots, \pi_n)$ are the policies of all agents

See M. L. Littman, 1994

Markov Games: Value Iteration

- Learning with state transitions
- Combination of repeated games and MDPs
- Agent k need to estimate $\mathcal{Q}(s, \mathbf{a})$ for joint actions $\mathbf{a} = (a_1, \ldots, a_n)$, not only $\mathcal{Q}(s, a_k)$
- Action a_k by agent k is selected based on the observation of $\mathbf{a}_{-k} = (a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_n)$
- Stationary solutions may not exist: Mixed strategies, average reward schemes

MA $\mathcal{Q}$-learning

- $t = 0$
- $\mathcal{Q}_k(s, \mathbf{a}) = 0 \; \forall s, \mathbf{a}, k$
- **repeat**
  - **for all** agents k **do**
    - select action $a_k(t)$
  - execute joint action $\mathbf{a} = (a_1, ..., a_n)$
  - observe new state s' and rewards r_k
  - **for all** agents k **do**
    - $\mathcal{Q}_k(s, \mathbf{a}) = \mathcal{Q}_k(s, \mathbf{a}) + \eta \left(R_k(s, \mathbf{a}) + \gamma V_k(s') - \mathcal{Q}_k(s, \mathbf{a}) \right)$
- **until** Termination Condition

How is the value $V_k(s')$ determined in MA $\mathcal{Q}$-learning?

Various options

- Opponent modelling (Joint Action Learner by Claus & Boutilier, 1998)
  - count how often $(s, \mathbf{a}_{-k})$ is played by other agents in state s
  - calculate frequency

$$F(s, \mathbf{a}_{-k}) = \frac{\#(s, \mathbf{a}_{-k})}{\sum_{\mathbf{a}_{-\ell}} \#(s, \mathbf{a}_{-\ell})}$$

  - $V_k(s) = \max_{a_k} \mathcal{Q}(s, a_k) = \max_{a_k} \sum_{\mathbf{a}_{-k}} F(s, \mathbf{a}_{-k}) \mathcal{Q}(s, \mathbf{a})$
- Assume other agents trying to minimise your return
  - $V_k(s) = \min_{\mathbf{a}_{-k}} \max_{\pi_k} \sum \pi_k \mathcal{Q}(s, \mathbf{a})$
- Assume other agent will follow an equilibrium strategy
  - $V_k = \mathrm{Nash}_k(s, \mathcal{Q}_1, \ldots, \mathcal{Q}_n)$, i.e. return at a Nash equilibrium
  - converges for self-play learning (convergence is often a problem in MAMG)
  - Nash equilibrium is not unique in general

Markov Games: Policy Iteration

- Interconnected Learning Automata for Markov Games (MG-ILA) by Vrancx et al. (2008)
- Policy stored by learning automata: For each agent and state $LA(s, k)$
- Update LAs by expected average reward
- Algorithm converges to Nash equlibrium (if exists)

Interconnected Learning Automata for MG (MG-ILA)

initialise $r_{prev}(s,k)$,$t_{prev}(s)$,$a_{prev}(s,k)$,t,$r_{tot}(k)$,$\rho_k(s,a)$,$\tau_k(s,a) \rightarrow 0$,$\forall s,k,a$,$s \leftarrow s\,(0)$

loop

for all Agents k **do**

if s was visited before **then**

Calculate received reward and time passed since last visit to state s:

$$\Delta r_k = r_{tot}(k) - r_{prev}(s,k),\ \Delta t = t - t_{prev}(s)$$

Update estimates for action $a_{prev}(s,k)$ taken on last visit to s:

$$\rho_k(s, a_{prev}(s,k)) = \rho_k(s, a_{prev}(s,k)) + \Delta r_k$$
$$\tau_k(s, a_{prev}(s,k)) = \tau_k(s, a_{prev}(s,k)) + \Delta t$$

$LA(s,k)$ uses L_{R-I} update with $a(t) = a_{prev}(s,k)$ and av. reward

$$\beta_k(t) = \rho_k(s, aprev(s,k))/\tau_k(s, a_{prev}(s,k)).$$

$LA(s,k)$ selects action a_k.

For current state store: $t_{prev}(s) \leftarrow t,\ r_{prev}(s,k) \leftarrow r_{tot}(k),\ a_{prev}(s,k) \leftarrow a_k$

Execute action $a = (a_1, ..., a_n)$, observe rewards r_k and new state s'

$s \leftarrow s',\ r_{tot}(k) \leftarrow r_{tot}(k) + r_k,\ t \leftarrow t + 1$

Sparse interactions

- Reduce the joint-action space
- If there is no interaction, act indepedently
- Learn at what states where coordination is necessary, e.g Sparse Tabular Multiagent $\mathcal{Q}$-learning (Guestrin et al., 2002).
- Represent dependencies that are limited to a few agents, e.g. Sparse Cooperative $\mathcal{Q}$-learning. (Kok and Vlassis, 2004, 2006)
- Similar to SC$\mathcal{Q}$ with learning of coordination graphs (Utile Coordination by Kok et al. (2005)

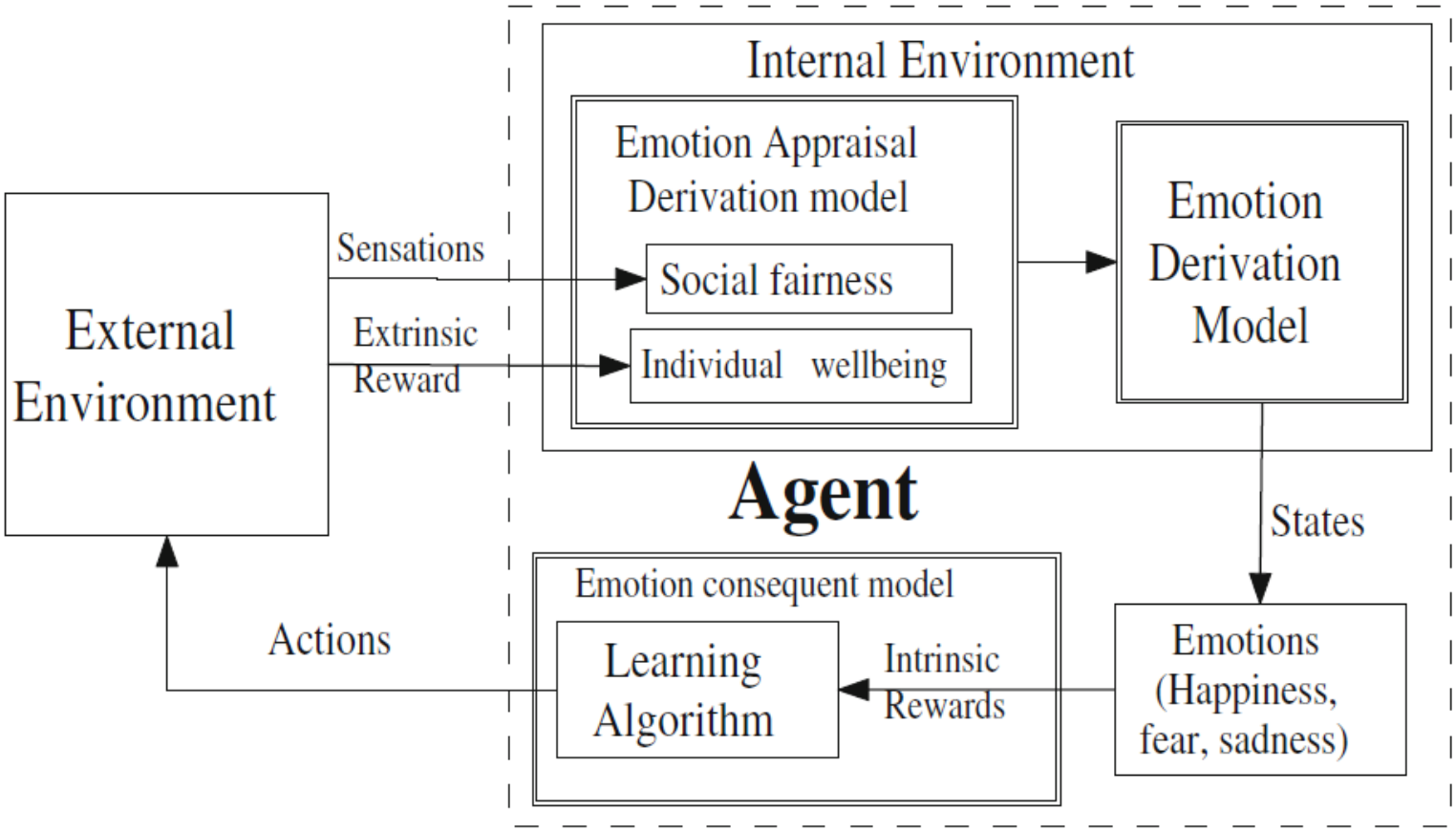
Coordinating $\mathcal{Q}$ (C$\mathcal{Q}$) Learning, (De Hauwere et al, 2010)

Initialise Q_k through single agent learning and Q_k^j;
while true do
  if state s_k of Agent k is unmarked **then**
    Select a_k for Agent k from Q_k
  else
    if the joint state information js is safe **then**
      Select a_k for Agent k from Q_k
    else
      Select a_k for Agent k from Q_k^j based on the joint state information js
  Sample $\langle s_k, a_k, r_k \rangle$
  if t-test detects difference in observed rewards vs expected rewards for $\langle s_k, a_k \rangle$ **then**
    mark s_k
    for $\forall$ other state information present in the joint state js **do**
      if t-test detects difference between independent state s_k and joint state js **then**
        add js to Q_k^j
        mark js as dangerous
      else
        mark js as safe
  if s_k is unmarked for Agent k or js is safe **then**
    No need to update $Q_k(s_k)$.
  else
    Update $Q_k^j(js, a_k) \leftarrow (1-\alpha_t) Q_k^j(js, a_k) + \alpha_t [r(js, a_k) + \gamma \, max_a Q(s'_k, a)]$

Discussion

- MARL algorithms currently limited to simple problems
- Relations to game theory, evolution theory
- Main point: Uncertainty, learning
- Problems: Uncertainty about actions or rewards, delayed rewards, instabilities through careless exploration
- DecPOMDP (Bernstein et al., 2002): Planning takes place in an off-line phase, after which the plans are executed in an on-line phase. This on-line phase is completely **dec**entralized.

Emotional MARL



Yu, C., Zhang, M., & Ren, F. (2013). Emotional Multiagent Reinforcement Learning in Social Dilemmas. In PRIMA 2013: Principles and Practice of Multi-Agent Systems (pp. 372-387). Springer Berlin Heidelberg.

Acknowledgements & References

Marco Wiering and Martijn van Otterlo (Eds.) Reinforcement Learning: State-of-the-Art Springer 2012.

Busoniu, L., Babuska, R., & De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 38(2), 156-172.

See also:

See also: http://umichrl.pbworks.com/w/page/7597585/Myths of Reinforcement Learning

Quotations

Good and evil, reward and punishment, are the only motives to a rational creature: these are the spur and reins whereby all mankind are set on work, and guided.

Locke

The human organism is inherently active, and there is perhaps no place where this is more evident than in little children. They pick things up, shake them, smell them, taste them, throw them across the room, and keep asking, "What is this?" They are unendingly curious, and they want to see the effects of their actions. Children are intrinsically motivated to learn, to undertake challenges, and to solve problems. Adults are also intrinsically motivated to do a variety of things. They spend large amounts of time painting pictures, building furniture, playing sports, whittling wood, climbing mountains, and doing countless other things for which there are not obvious or appreciable external rewards. The rewards are inherent in the activity, and even though there may be secondary gains, the primary motivators are the spontaneous, internal experiences that accompany the behavior.

Deci and Ryan, 1985 (cf. A. Barto, 2013)