UNIVERSITY OF LIVERPOOL

COMP 532
Machine Learning and BioInspired Optimization

Lecture 20: Multi-Agent Learning

Dr. Shan Luo

Department of Computer Science

shan.luo@liverpool.ac.uk

Admin: the Exam

What to expect:

- Five questions, answer four
  - Topics: RL, MAL, SI, DL, AIS/DNA
  - 25 points each

- Each question/topic:
  - **Technical part**
    e.g. compute something, explain an algorithm
  - **Theoretical part**
    e.g. explain intuitions behind certain mechanisms

Admin: Task 2

- **First**: make groups of 3/4 (preferably 3)
- **Each group is assigned one paper** from the list on VITAL, on one or several topics of the lectures
- **Study the paper**, and **prepare a presentation**
  - You will present your paper in class (10 minutes)
  - Deadline to submit your slides: **27 April**
  - Presentations will be scheduled after that date
  - Points++: try the code in GitHub, demonstrate some results (if possible…)

Outline (3-ish lectures)

- Introduction to Evolutionary Game Theory
  - Replicator Dynamics
  - Evolutionarily Stable Strategies
  - Example games
- Formal link between RL and EGT
  - Deriving the dynamics of Cross' Learning
  - Extension to other RL algorithms
- Applications of this model
  - Parameter tuning
  - Analyzing complex strategic interactions

Formal link between RL and EGT

- Main reading material:

Bloembergen, *et al.*., 2015. Evolutionary Dynamics of Multi-Agent Learning: A Survey. *JAIR*, *53*, pp.659-697.

Available in Vital.

Recap: Evolutionary Game Theory

- Studies a **population**..

  ..of individuals of different **types**..

  ..who are randomly paired in interaction..

  ..and whose relative **fitness** determines their reproductive success

- Evolutionary operators:
  - **Selection** ↔ exploitation
  - **Mutation** ↔ exploration

Recap: Replicator Dynamics

- Standard **Replicator Dynamics**:

$$\dot{x}_i = x_i \left[f_i(x) - \sum_j x_j f_j(x) \right]$$

- When interactions are modelled as a **game**:

$$\dot{x}_i = x_i [(Ax)_i - x^\intercal Ax]$$

- When **two populations** co-evolve:

$$\dot{x}_i = x_i [(Ay)_i - x^\intercal Ay]$$

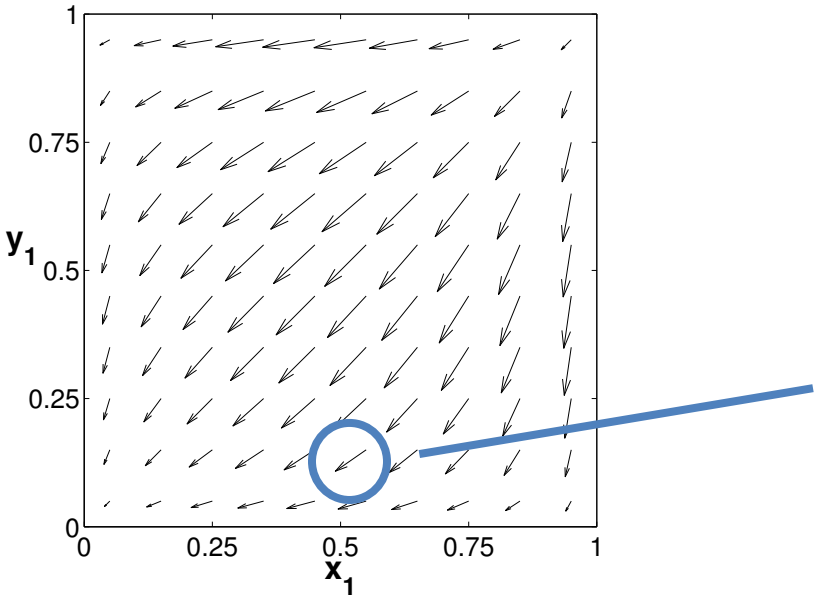
$$\dot{y}_i = y_i [(x^\intercal B)_i - x^\intercal By]$$

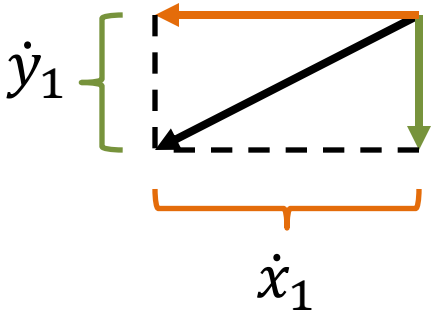
Recap: Replicator Dynamics

$$\dot{x}_i = x_i[(Ay)_i - x^\intercal Ay]$$

$$\dot{y}_i = y_i[(x^\intercal B)_i - x^\intercal By]$$

	C	D	
C	3, 3	0, 5	x_1
D	5, 0	1, 1	x_2
	y_1	y_2	





The Link to Multi-Agent Learning

What is the interpretation of the Replicator Dynamics?

- The evolution of a population of individuals under natural selection
  (Evolutionary Game Theory)
- A player gradually adapting her strategy
  (Classical Game Theory)
- The policy change of a learning agent!
  (Multi-Agent Learning)

Dictionary

Reinforcement Learning	Classical Game Theory	Evolutionary Game Theory
environment	game	game
agent	player	population
action	action	type
policy	strategy	distribution over types
reward	payoff	fitness

The Link to Multi-Agent Learning

A simple example game: **Matching Pennies**

- Two players choose heads or tails
- Player 1 wins if both choose different sides
- Otherwise player 2 wins

	H	T
H	-1, 1	1, -1
T	1, -1	-1, 1

- **Nash equilibrium**: choose either action with equal probability

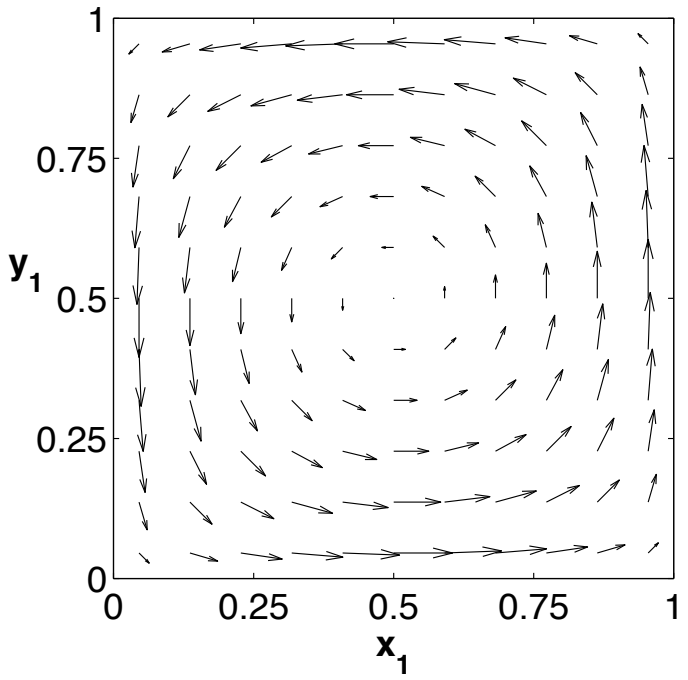
The Link to Multi-Agent Learning

- Two population replicator dynamics

$$\dot{x}_i = x_i[(Ay)_i - x^\mathsf{T} Ay]$$

$$\dot{y}_i = y_i[(x^\mathsf{T} B)_i - x^\mathsf{T} By]$$

	H	T
H	-1, 1	1, -1
T	1, -1	-1, 1



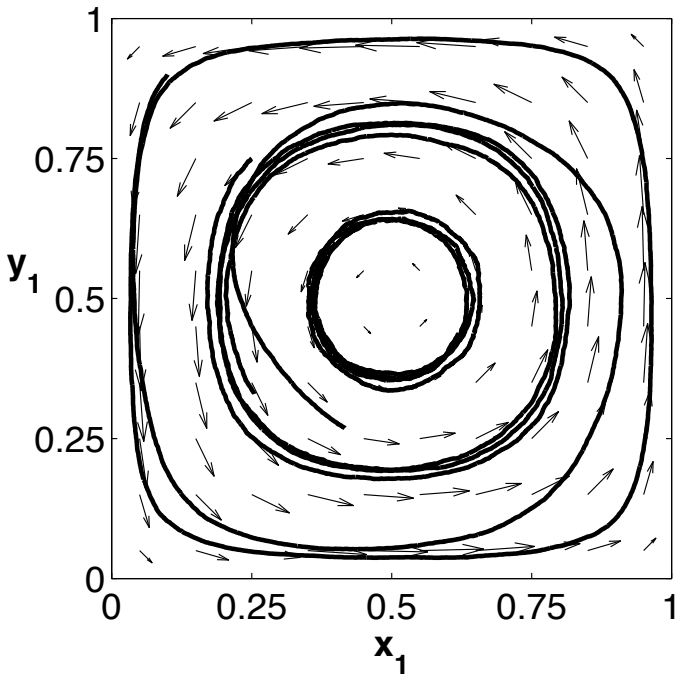
The Link to Multi-Agent Learning

Learning to play the game

	H	T
H	-1, 1	1, -1
T	1, -1	-1, 1

- **Reinforcement learning**
  - Players iteratively improve their policy over time
  - E.g. **Learning Automata**

- Policy traces match precisely with the replicator dynamics!



Formalising the Link

- **Cross' Learning**:

$$\pi(a) \leftarrow \pi(a) + \alpha \begin{cases} r - \pi(a)r & \text{if taken action } a \\ -\pi(a)r & \text{otherwise} \end{cases}$$

- What is the expected policy update of this algorithm?

Cross, 1973. A stochastic learning model of economic behavior. *The Quarterly Journal of Economics*, *87*(2), pp.239-266.

Formalising the Link

$$\pi(a) \leftarrow \pi(a) + \alpha \begin{cases} r - \pi(a)r & \text{if taken action } a \\ -\pi(a)r & \text{otherwise} \end{cases}$$

- Expected update to $\pi(a)$ after taking action a:

$$E[\Delta\pi(a)] = \alpha\left[E_a[r] - \pi(a)E_a[r]\right]$$

- Expected update to $\pi(a)$ after taking action $b \neq a$:

$$E[\Delta\pi(a)] = \alpha\left[-\pi(a)E_b[r]\right]$$

Börgers & Sarin, 1997. Learning through reinforcement and replicator dynamics. *J. of Economic Theory*, 77(1), pp.1-14.
Bloembergen, *et al.*., 2015. Evolutionary Dynamics of Multi-Agent Learning: A Survey. *JAIR*, *53*, pp.659-697.

Formalising the Link

$$\pi(a) \leftarrow \pi(a) + \alpha \begin{cases} r - \pi(a)r & \text{if taken action } a \\ -\pi(a)r & \text{otherwise} \end{cases}$$

- Combining those, weighted by their probability of occurring:

$$E[\Delta\pi(a)] = \alpha\pi(a)\left[E_a[r] - \pi(a)E_a[r]\right] + \alpha \sum_{b \neq a} \pi(b)\left[-\pi(a)E_b[r]\right]$$

probability of taking action a

probability of taking any other action

Formalising the Link

$$E[\Delta\pi(a)] = \alpha\pi(a)\left[E_a[r] - \pi(a)E_a[r]\right] + \alpha\sum_{b\neq a}\pi(b)\left[-\pi(a)E_b[r]\right]$$

We can rewrite this equation as:

$$E[\Delta\pi(a)] = \alpha\pi(a)\left[E_a[r]\right] - \alpha\left[\pi(a)\pi(a)E_a[r]\right] - \alpha\sum_{b\neq a}\left[\pi(b)\pi(a)E_b[r]\right]$$

$$= \alpha\pi(a)\left[E_a[r]\right] - \alpha\sum_{b}\left[\pi(b)\pi(a)E_b[r]\right]$$

Formalising the Link

$$E[\Delta\pi(a)] = \alpha\pi(a)\left[E_a[r]\right] - \alpha\left[\pi(a)\pi(a)E_a[r]\right] - \alpha\sum_{b\neq a}\left[\pi(b)\pi(a)E_b[r]\right]$$

$$= \alpha\pi(a)\left[E_a[r]\right] - \alpha\sum_{b}\left[\pi(b)\pi(a)E_b[r]\right]$$

$$= \alpha\pi(a)\left[E_a[r]\right] - \alpha\pi(a)\sum_{b}\left[\pi(b)E_b[r]\right]$$

$$= \alpha\pi(a)\left[E_a[r] - \sum_{b}\left[\pi(b)E_b[r]\right]\right]$$

Formalising the Link

- So, the expected policy change of Cross' Learning is

$$E[\Delta\pi(a)] = \alpha\pi(a)\left[E_a[r] - \sum_b \left[\pi(b)E_b[r]\right]\right]$$

- Now assume very small update steps δ such that

$$\pi_{t+\delta}(a) = \pi_t(a) + \delta\Delta\pi_t(a)$$

- Taking the limit $\delta \to 0$ gives

$$\dot{\pi}(a) = \alpha\pi(a)\left[E_a[r] - \sum_b \left[\pi(b)E_b[r]\right]\right]$$

Formalising the Link

- Now we have the policy gradient of Cross' Learning

$$\dot{\pi}(a) = \alpha\pi(a)\left[E_a[r] - \sum_b \left[\pi(b)E_b[r]\right]\right]$$

- Suppose the learner interacts in a matrix game
  - with payoff matrix A
  - using the simplified stateless policy x with $x_a = \pi(a)$
  - against an opponent playing policy y

$$\dot{x}_a = \alpha x_a[(Ay)_a - x^\intercal Ay]$$

Looks familiar??

Formalising the Link

$$\pi(a) \leftarrow \pi(a) + \alpha \begin{cases} r - \pi(a)r & \text{if taken action } a \\ -\pi(a)r & \text{otherwise} \end{cases}$$

"EQUALS"

$$\dot{x}_a = \alpha x_a [(Ay)_a - x^\intercal Ay]$$

Cross' Learning behaviour matches the Replicator Dynamics in expectation!

The Link between RL and EGT

The formal link between EGT and RL has many advantages

- Modifications allow to model various RL algorithms
- Provides insight into the black box of RL
- Simplifies parameter tuning
- Allows to create new learning algorithms by first designing the preferred dynamics

Dynamics of Q-learning

We can modify the replicator dynamics to match the simplified form of **Q-learning**

$$\mathrm{Q}(a) \leftarrow \mathrm{Q}(a) + \alpha[r - \mathrm{Q}(a)]$$

with Boltzman / softmax action selection, which generates policy π as

$$\pi(a) = \frac{e^{Q(a)/\tau}}{\sum_b e^{Q(b)/\tau}}$$

with temperature τ balancing exploration and exploitation

Dynamics of Q-learning

The modified replicator dynamics that match **Boltzman Q-learning**

$$\dot{x}_i = \frac{\alpha x_i}{\tau}\underbrace{[(Ay)_i - x^\intercal Ay]}_{\text{selection}} - \alpha x_i \underbrace{\left[\log x_i - \sum_j x_j \log x_j\right]}_{\text{mutation}}$$

selection following the replicator dynamics → **exploitation**

mutation as a result of Boltzman action selection → **exploration**

Comparing RL Algorithms

- In a similar way we can derive the dynamics of many reinforcement learning algorithms
  - Cross' Learning
  - Q-learning
  - Infinitesimal Gradient Ascent (IGA)
  - Win-or-Learn-Fast IGA (WoLF)
  - Weighted Policy Learner (WPL)

- **Using the dynamics we can easily compare their learning behaviour!**

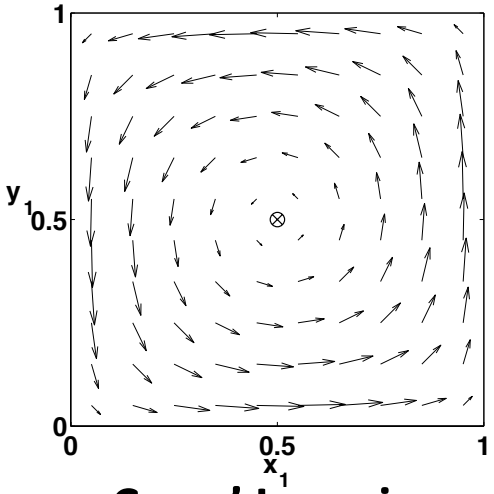
Comparing RL Algorithms

Simplified dynamics in 2 player 2 action games

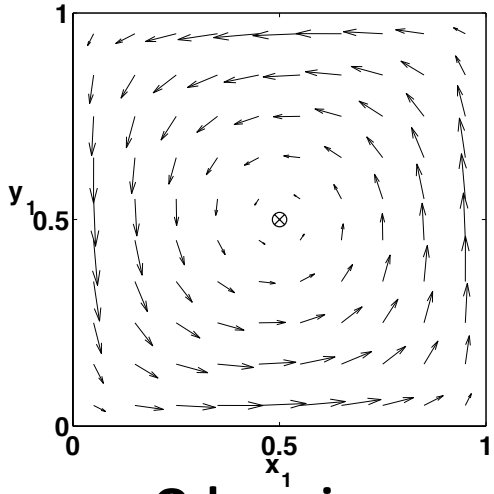
Algorithm	$\dot{x}$
IGA	$\alpha\eth$
IGA-WoLF	$\eth \cdot \begin{cases} \alpha_{min} & \text{if } V(\mathbf{x}) > V(\mathbf{x}^*) \\ \alpha_{max} & \text{otherwise} \end{cases}$
WPL	$\alpha\eth \cdot \begin{cases} x & \text{if } \eth < 0 \\ (1-x) & \text{otherwise} \end{cases}$
CL	$x(1-x)\ \eth$
Q-learning	$\alpha x(1-x)\left[\eth \cdot \tau^{-1} - \log \frac{x}{1-x}\right]$

where ð is the gradient of the value function w.r.t. $x = x_1$

Matching Pennies

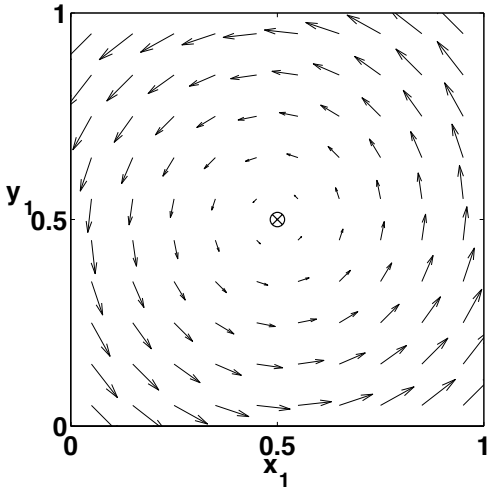


Cross' Learning

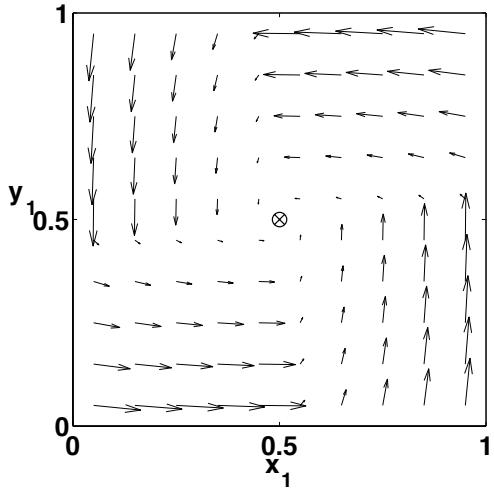


Q-learning

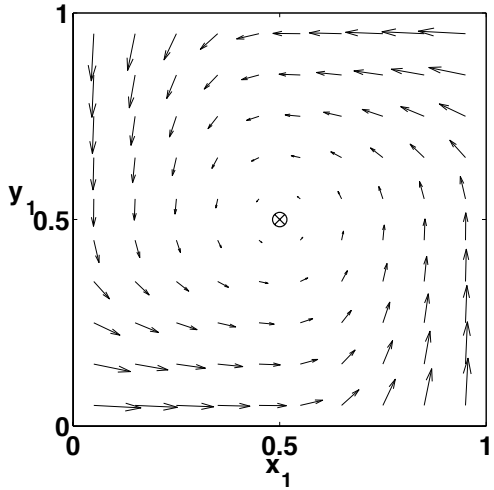
	H	T
H	-1, 1	1, -1
T	1, -1	-1, 1



IGA



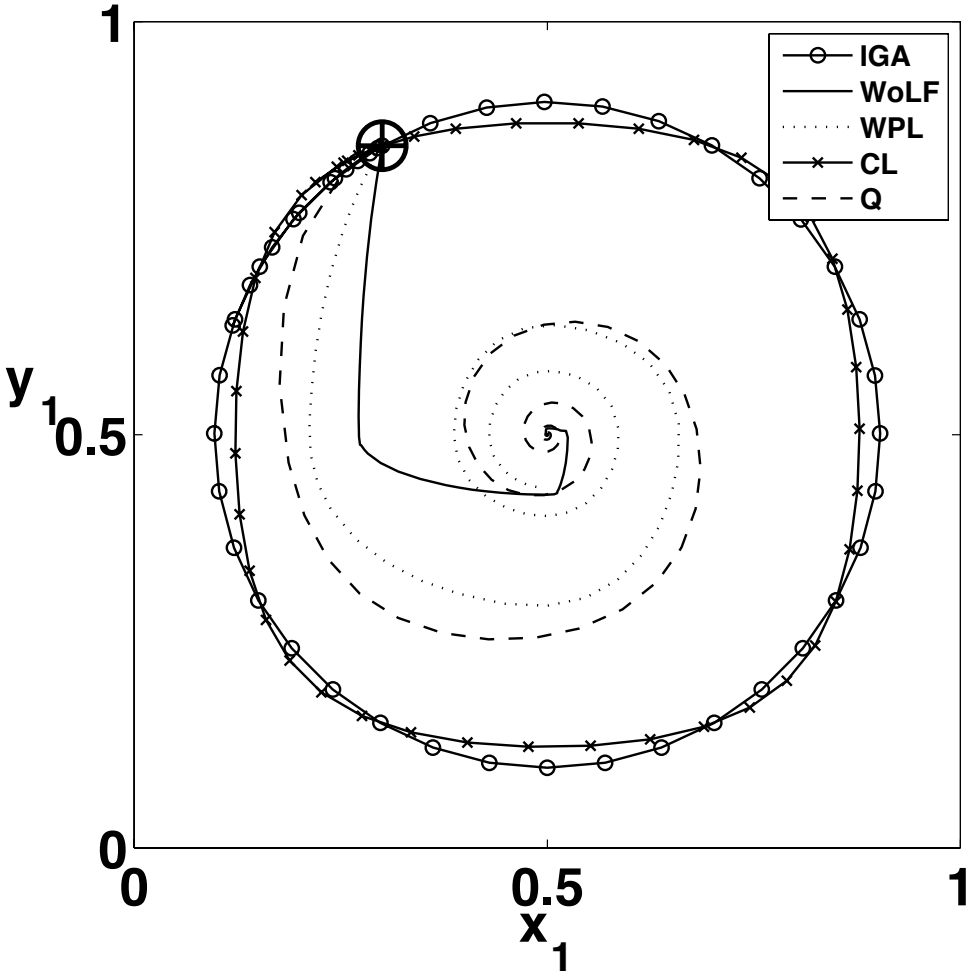
WoLF



WPL

Figure 7 in Bloembergen, *et al.*, survey paper, *JAIR* 2015.

Matching Pennies



"Tracing" the dynamics gives detailed insights into convergence!

The Dynamics of MARL

- So far, restricted to **stateless games** with **discrete actions**

- However, promising steps have been taken to extend the evolutionary model
  - to **continuous action spaces**
  - to **multi-state games**
  - to **extensive form games**

Applications of the Evolutionary Model

- The evolutionary model is useful to study dynamics of a given learner ..
  .. which can facilitate **parameter tuning**

- **Reverse approach**: design a learning algorithm that exhibits desired dynamics

- Study **complex strategic interactions** that normally defy formal analysis

Wrapping up

- Formal link between RL and EGT
  - Deriving the dynamics of Cross Learning
  - Extension to other RL algorithms, e.g., Boltzman Q-learning
- Next lecture: Applications EGT models