MULTI-AGENT REINFORCEMENT LEARNING

Prof. Dr. A. Nowé

MULTI-AGENT SYSTEMS

A *multi-agent system* (**MAS**) consists of a set of agents *A* that are situated in an environment *E*

CHALLENGES

Agents are physically distributed, cannot necessarily observe each other

Communication can be expensive

Agents might have common as well as conflicting interests

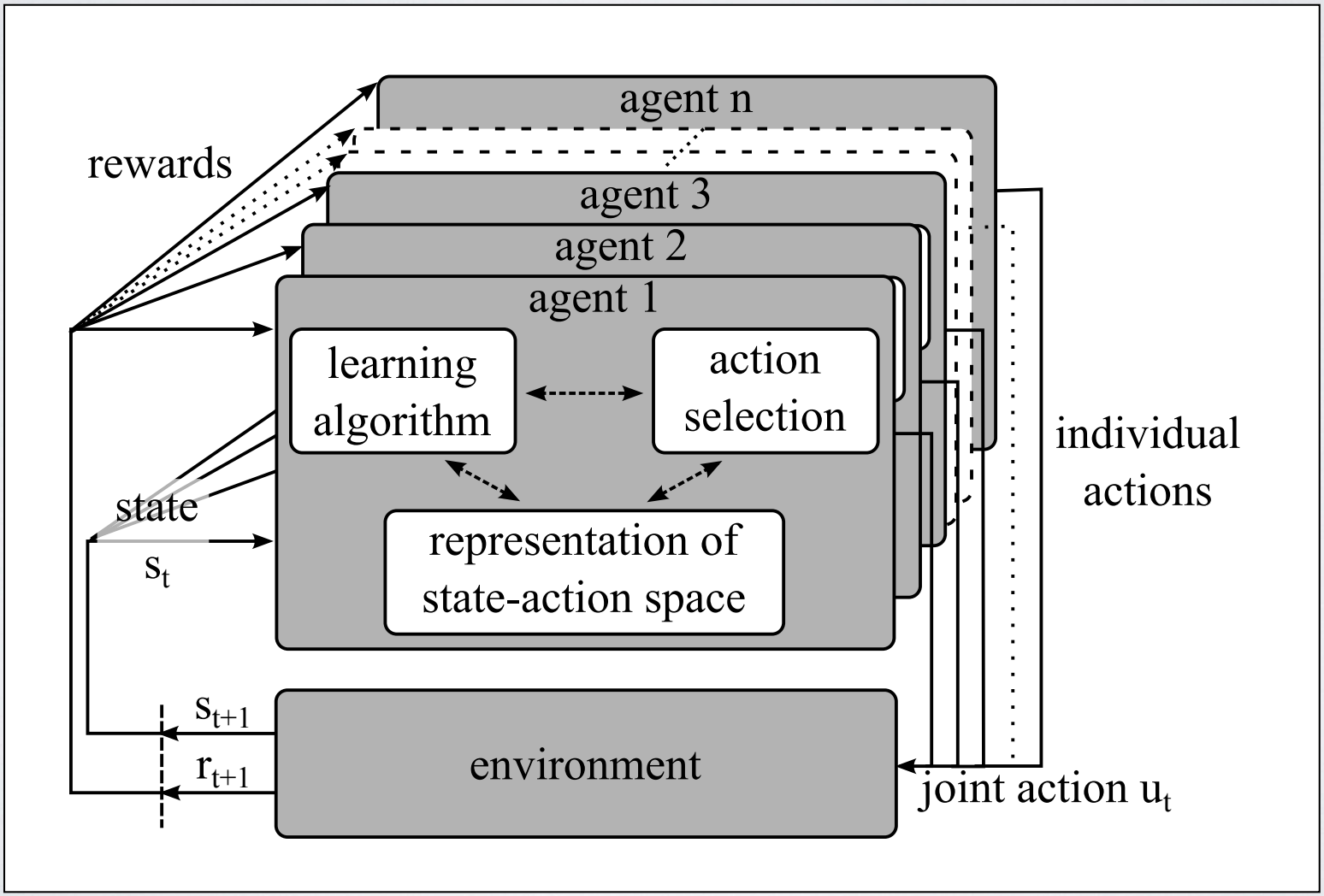
KEY QUESTIONS

Are RL algorithms guaranteed to converge in MAS settings?

If so, do they converge to (optimal) equilibria?

Are there differences between agents that learn as if there are no other agents (i.e. use single agents RL algorithms) and agents that attempt to learn both the values of specific joint actions and the strategies employed by other agents?

CONCEPTUAL OVERVIEW



RECAP

- Q-learning
- Exploration - Exploitation trade off

Q-LEARNING

One-step Q-learning:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s,a)]$$

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

Q-LEARNING

One-step Q-learning: single state, no long term reward

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s,a)]$$

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

ACTION/VALUE METHODS

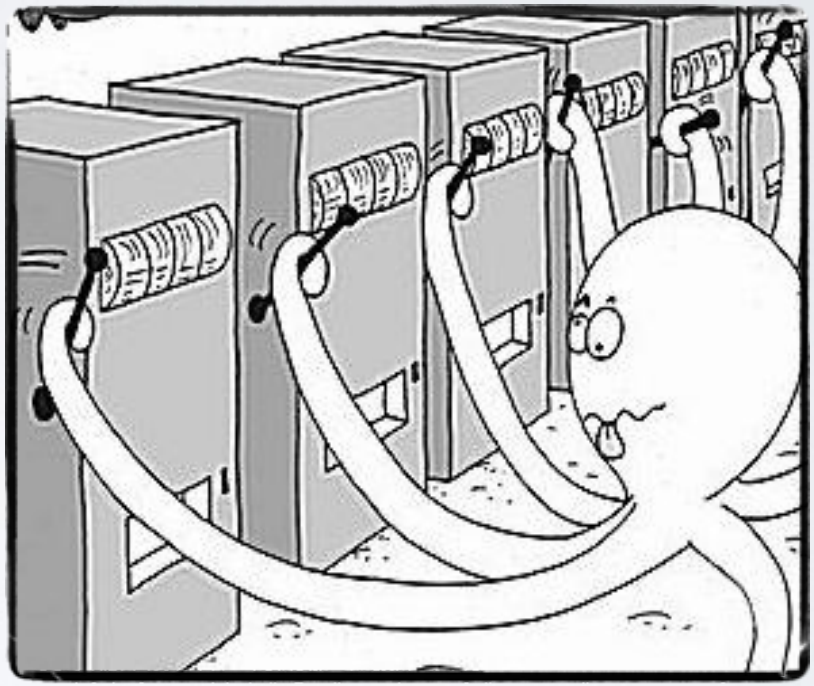
Methods that adapt action-value estimates and nothing else, e.g. suppose by the *t*-th play, action *a* has been chosen k_a times, producing rewards $r_1, r_2, \ldots, r_{k_a}$, then

$$Q_t(a) = \frac{r_1, r_2, \ldots, r_{k_a}}{k_a}$$ **Sample average**

$$\lim_{k_a \to \infty} Q_t(a) = Q^*(a)$$

N-ARMED BANDITS

Single state RL



EXPLORATION/EXPLOITATION DILEMMA

- Suppose you calculate estimates

  $Q_t(a) \approx Q^*(a)$ as t grows

- The greedy action at t is $a_t^* = \underset{a}{\operatorname{argmax}}\, Q_t(a)$

  $a_t = a_t^* \rightarrow exploitation$

  $a_t \neq a_t^* \rightarrow exploration$

- Constant exploration = bad idea
- Constant exploitation = bad idea
- Stop exploration = bad idea
- Reduce exploration = good idea (maybe)

RANDOM EXPLORATION

- Simplest form of action selection
- Very good for exploration
- Very bad for exploitation

$$a_t = \text{random action}$$

ϵ-GREEDY EXPLORATION

- The simplest way to balance exploration and exploitation

- Greedy action selection

$$a_t = a_t^* = \underset{a}{\operatorname{argmax}}\, Q_t(a)$$

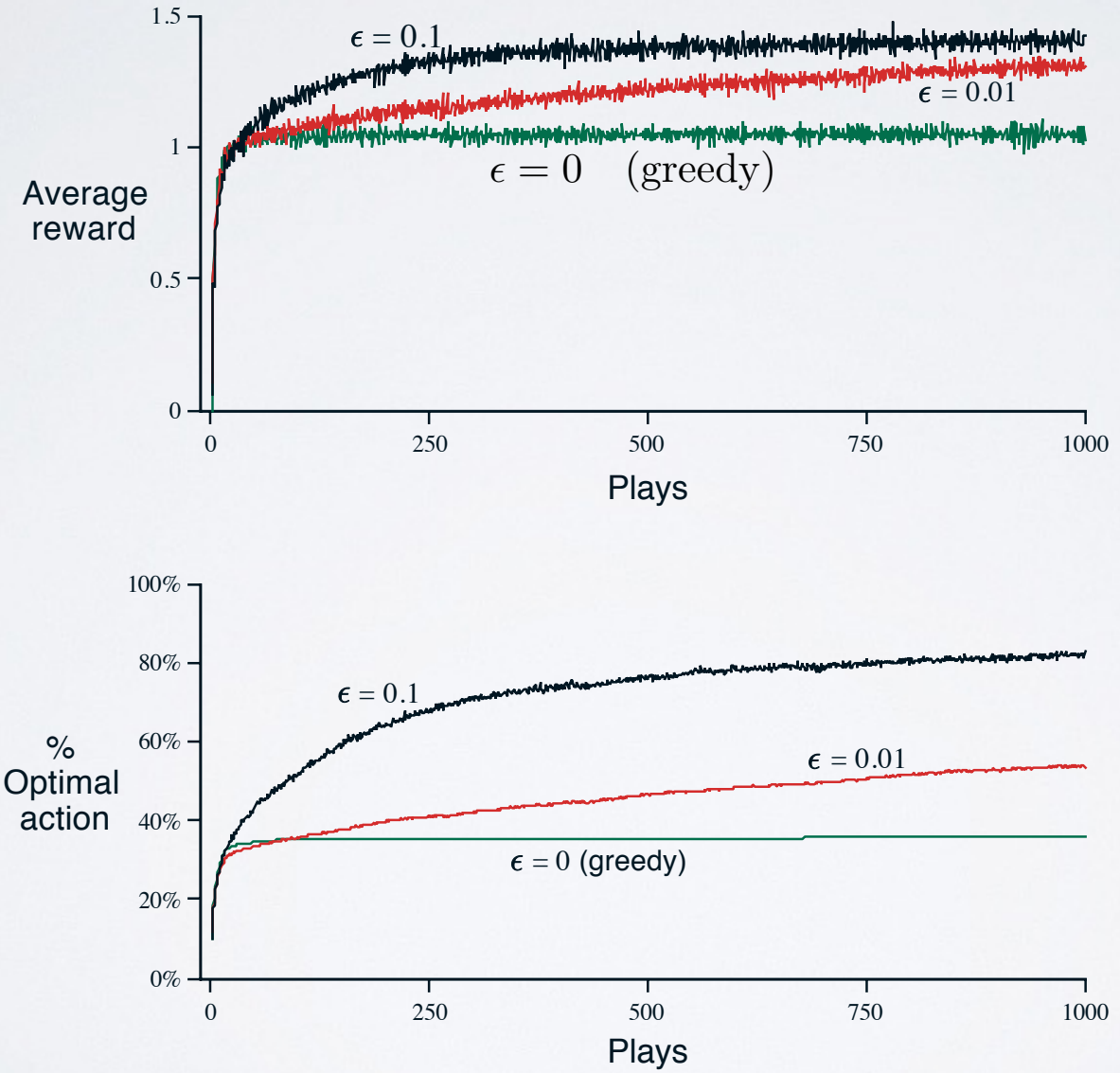
- ϵ-greedy action selection

$$a_t = \begin{cases} a_t^* & \text{with probability } 1 - \epsilon \\ \text{random action} & \text{with probability } \epsilon \end{cases}$$

10-ARMED TESTBED

- n = 10, so 10 possible actions
- Each $Q^*(a)$ is chosen randomly from a normal distribution $\eta(0, 1)$
- Each r_t is also normally distributed: $\eta(Q^*(a_t), 1)$
- 1000 plays
- Repeat the whole thing 2000 times and average the results

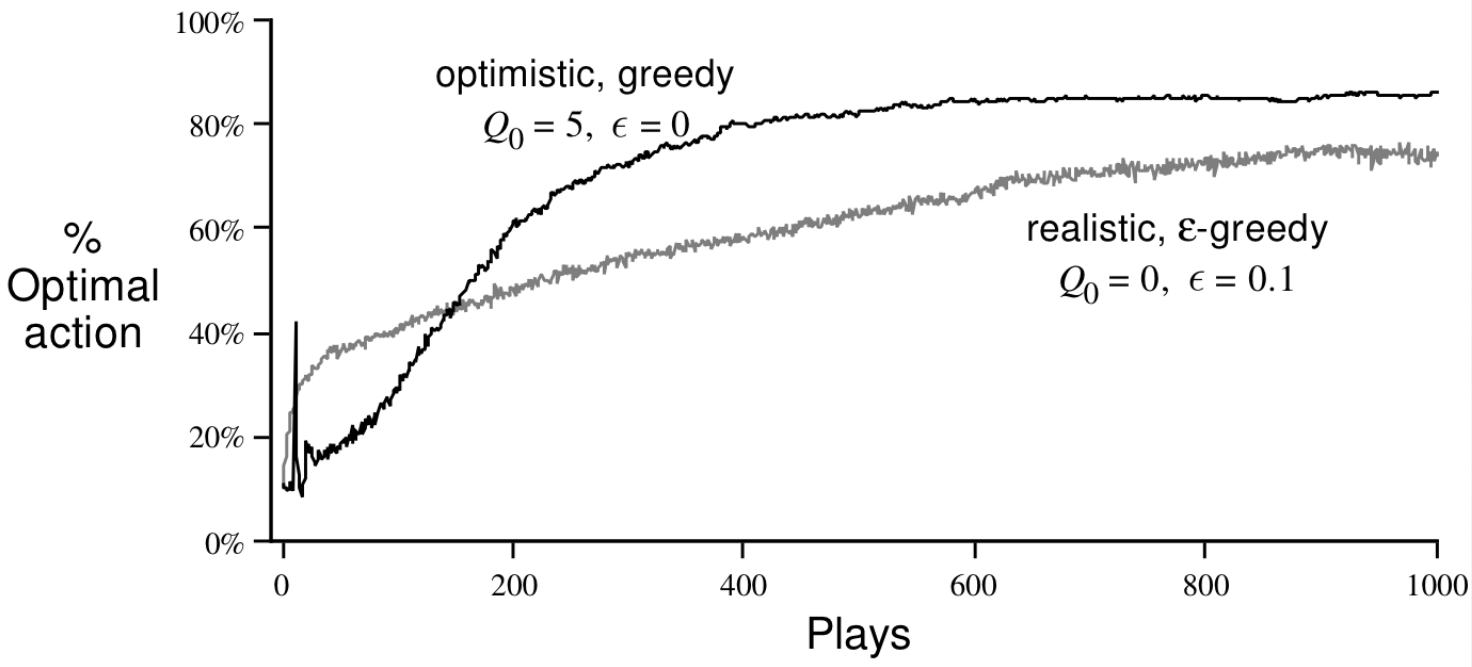
ϵ -GREEDY ON 10-ARMED TESTBED



PERFORMANCE

Is in general dependent on the interplay between the reward scheme, exploration strategy and initialisation of Q-values.

Assume same bandit, but $Q_0(a) = 5$ for all a



RL IN NORMAL FORM GAMES

- **Single stage setting**
- Common interest (Claus & Boutilier, Kapetanakis & Kudenko)
- Conflicting interest (based on Learning Automata)

SOFTMAX ACTION SELECTION

- Softmax action selection methods grade action probabilities by estimated values
- The most common softmax uses a Gibbs or Boltzmann distribution:

Choose action *a* on play *t* with probability

$$\frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}}$$

Where τ is the computational temperature

COMMON INTEREST

	a_0	a_1
b_0	x	0
b_1	0	y

If **x > y > 0** then
2 equilibria: (a_0, b_0) and (a_1, b_1) (1st one optimal)

if **x = y > 0** then
equilibrium selection problem

- Super RL agent (Q-values for joint actions and joint action selection)
  - No challenge, equivalent to single agent learning
  - Joint action learners (**JAL**) (Q-values for joint actions, actions are selected independently)
  - Independent learners (**IL**) (Q-values for individual actions, actions are selected independently)

SIMPLE SINGLE STAGE

Joint action learners (**JAL**) (Q-values for joint actions, actions are selected independently)

- Use Q-learning to learn $Q(a_0, b_0)$
  $Q(a_0, b_1)$
  $Q(a_1, b_0)$
  $Q(a_1, b_1)$

Assumption: actions taken by the other agents can be observed

- Action selection for individual agents:
  - The quality of an individual action depends on the action taken by the other agent
    ⇒ maintain beliefs about strategies of other agents

$$EV(a^i) = \sum_{a^{-i} \in A_{-i}} Q(a^{-i} \cup \{a^i\}) \prod_{j \neq 1} \{Pr^i_{a^{-i}[j]}\}$$

SIMPLE SINGLE STAGE

Independent learners (**IL**) (Q-values for individual actions, actions are selected independently)

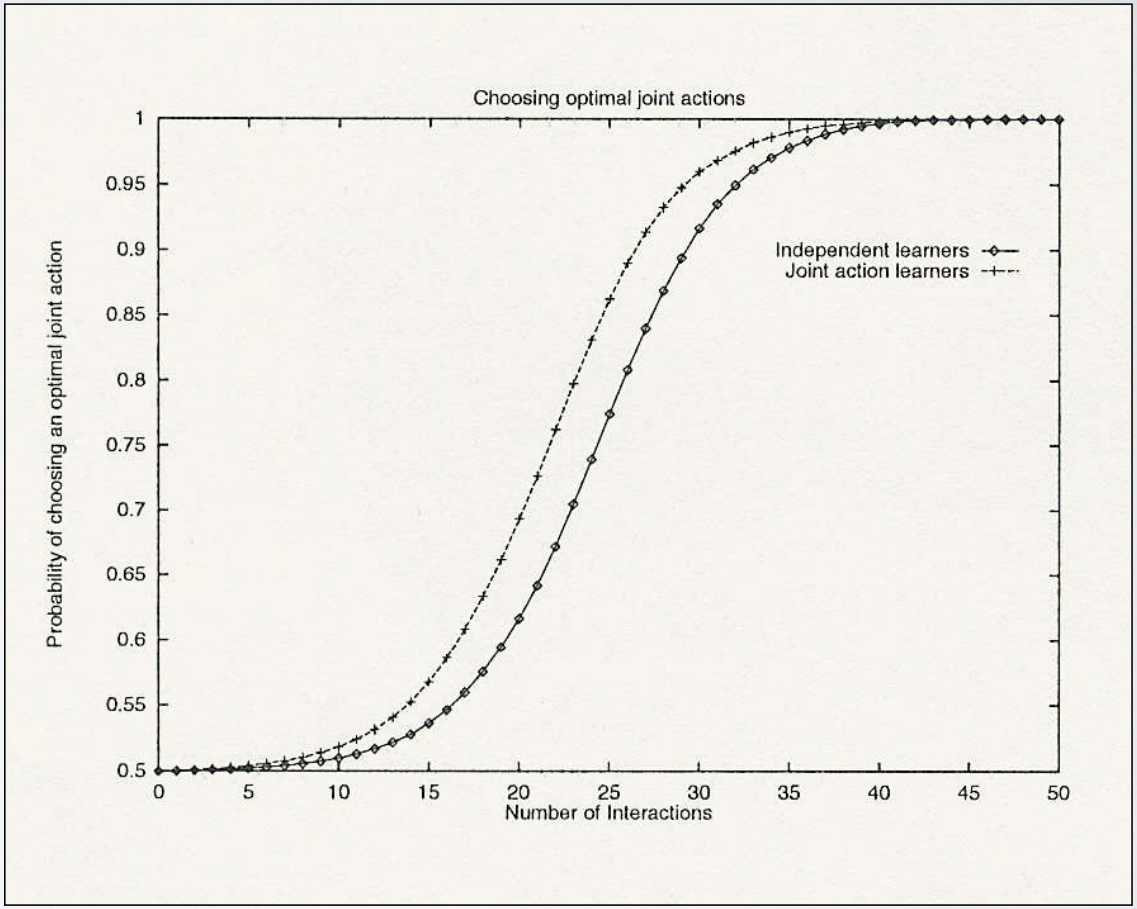
- Use Q-learning to learn $Q(a_0)$ $Q(a_1)$ $Q(b_0)$ $Q(b_1)$

Assumption: No need to observe actions taken by other agents

- Action selection for individual agents:
  - Exploration strategy is crucial
  - ~~Random~~
  - ⇒ Boltzmann with decreasing τ

COMPARING IL AND JAL

	a_0	a_1
b_0	10	0
b_1	0	10



THE PENALTY GAME

	a_0	a_1	a_2
b_0	10	0	k
b_1	0	2	0
b_2	k	0	10

$k < 0$

3 Nash Equilibria , 2 optimal

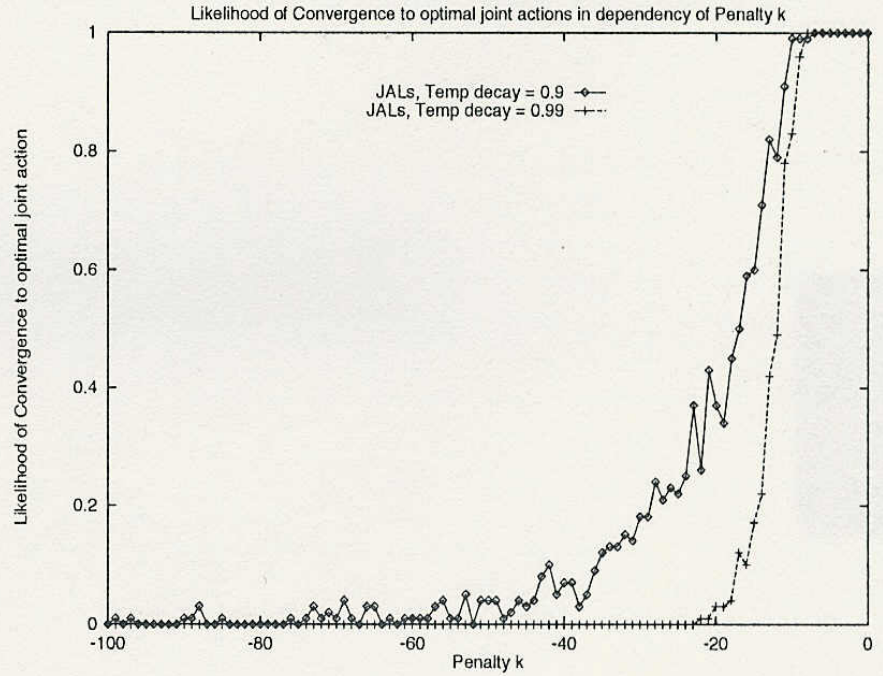
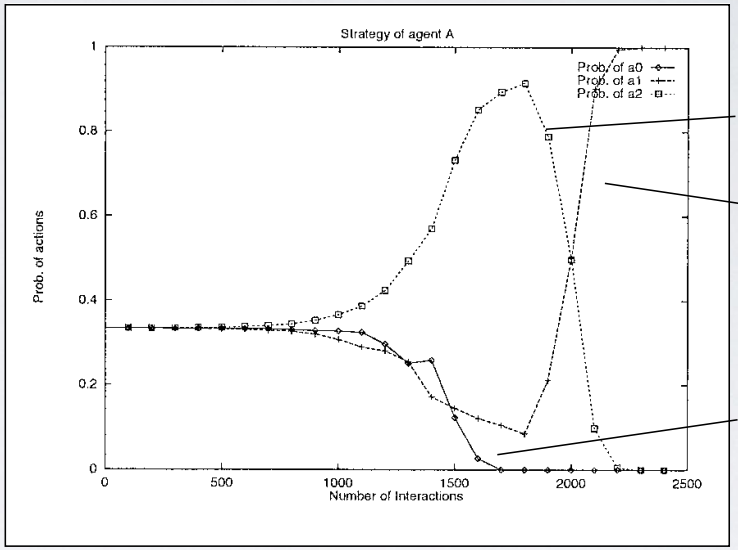


Figure 2: Likelihood of convergence to opt. equilibrium as a function of penalty k (averaged over 100 trials).

CLIMBING GAME

	a_0	a_1	a_2
b_0	11	-30	0
b_1	-30	7	6
b_2	0	0	5

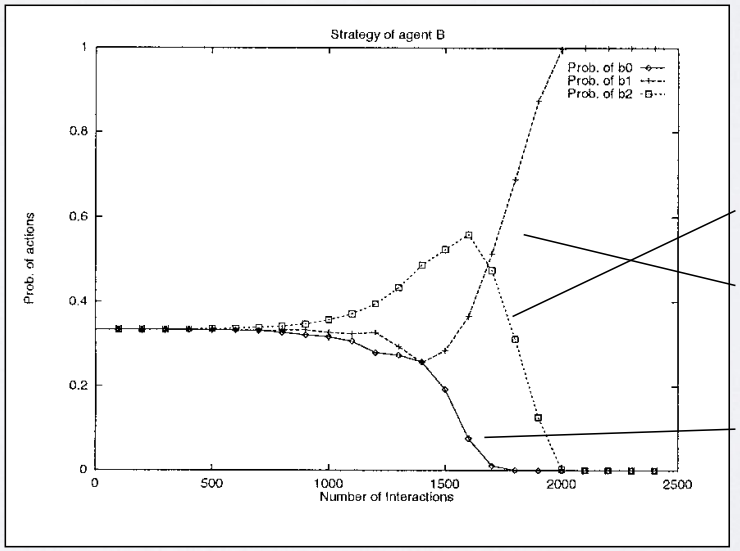
2 Nash Equilibria , 1 optimal



Action a2

Action a1

Action a0



Action b2

Action b1

Action b0

initial temperature 10000 is decayed at rate 0.995

FMQ HEURISTIC

Observation

Kapetanakis & Kudenko

- The setting of the temperature in the Boltzmann strategy for independent learners is crucial
- Convergence to some equilibrium, but not necessary optimal
- FMQ: Frequency maximum Q-value heuristic

$$EV(a) = Q(a) + c \times freq(\max R(a)) \times \max R(a)$$

Controls weight of heuristic

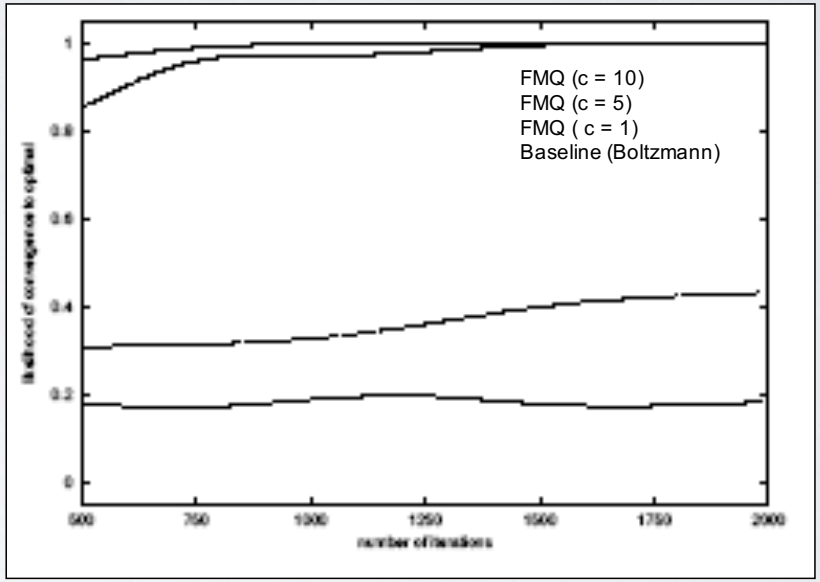
Fraction of time maxR(*a*)

Max reward so far for action *a*

FMQ HEURISTIC

	a_0	a_1	a_2
b_0	11	-30	0
b_1	-30	7	6
b_2	0	0	5

The climbing game



ACTIONS ARE SELECTED ACCORDING TO BOLTZMANN WITH DECREASING TEMPERATURE

FMQ HEURISTIC

The stochastic climbing game (50%)

	a_0	a_1	a_2
b_0	10/12	5/-65	8/-8
b_1	5/-65	14/0	12/0
b_2	5/-5	5/-5	10/0

Goal is stochastic

FMQ heuristic is not robust to stochastic reward games

Improvement: commitment sequences

COMMITMENT SEQUENCES

Kapetanakis & Kudenko

- **Motivation**
  - Difficult to distinguish between the two sources of uncertainty (other agents, multiple rewards)
- **Definition**
  - A commitment sequence is some list of time slots for which an agent is committed to taking the same action
- **Condition**
  - An exponentially increasing time interval between successive time slots

COMMITMENT SEQUENCES

Kapetanakis & Kudenko

- Sequence 1
  - 1, 3, 6, 10, 15, 22, ...
- Sequence 2
  - 2, 5, 9, 14, 20, 28, ...
- Sequence 3
  - 4, ...

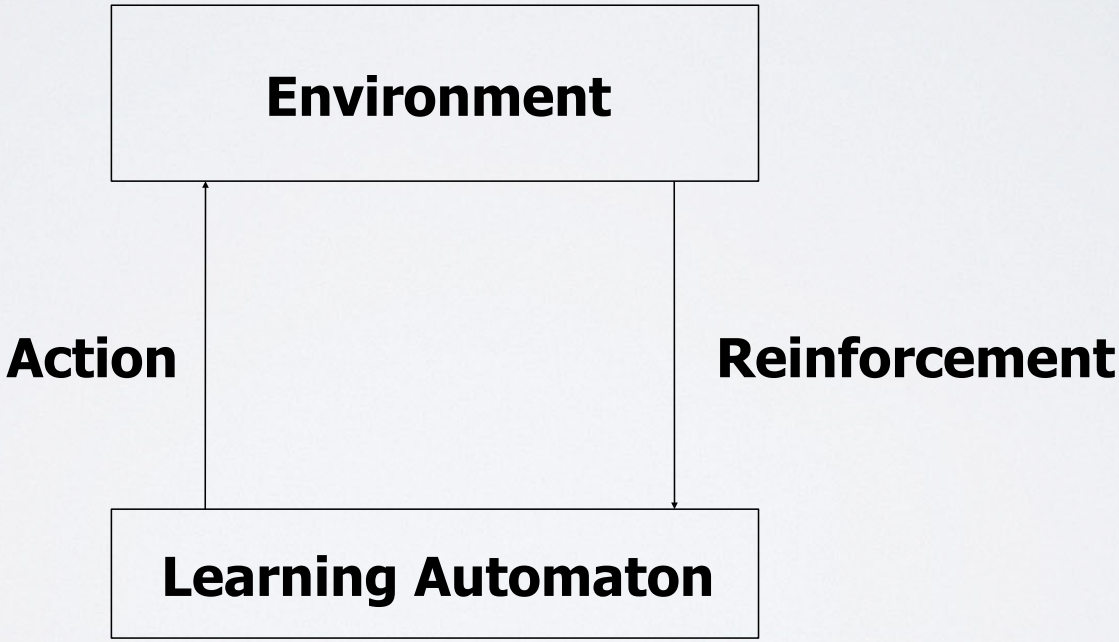
Assumptions:

1. Common global clock
2. Common protocol for defining commitment sequences

CONFLICTING INTEREST?

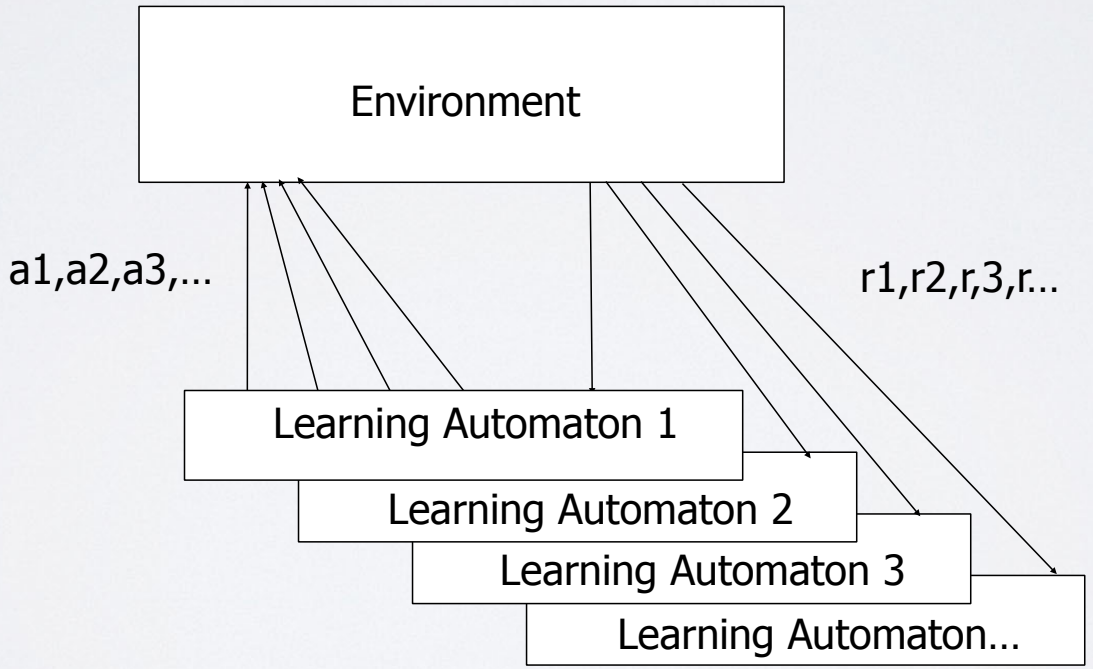
- Narendra and Wheeler (1989)
  - Players in an n-person non-zero sum game who use independently a reward-inaction update scheme with an arbitrary small step size will always converge to a pure equilibrium point
  - If the game has a pure Nash Equilibrium, the equilibrium point will not be part of one of the Nash Equilibrium
  - Convergence to Pareto Optimal (Nash) equilibrium is not guaranteed

LEARNING AUTOMATA



LEARNING AUTOMATA

Single stage, multi automata



LEARNING AUTOMATA

Single Stage, Single Agent

Assume **binary** feedback, and l actions
When feedback signal is **positive**,

$$p_i(k+1) = p_i(k) + a\left[1 - p_i(k)\right] \text{ if i}^{\text{th}} \text{ action is taken at time k}$$
$$p_j(k+1) = (1-a)p_j(k), \text{ for all } j \neq i$$

with a in $]0,1[$

When feedback signal is **negative**,

$$p_i(k+1) = (1-b)p_i(k), \text{ if i}^{\text{th}} \text{ action is taken at time k}$$
$$p_j(k+1) = b/(l-1) + (1-b)p_j(k), \text{ for all } j \neq i$$

Reward-penalty, L_{R-P}

Reward-ε penalty, $L_{R-\varepsilon P}$ $b << a$

LEARNING AUTOMATA, CONT.

When updates only happen at positive feedback, (or ***b = 0***)

$$p_i(k+1) = p_i(k) + a\left[1 - p_i(k)\right] \text{ if } i^{th} \text{ action is taken at time k}$$
$$p_j(k+1) = (1-a)p_j(k), \text{ for all } j \neq i$$

Reward-in-action, L_{R-I}

Some more types:

- Binary feedback : P-model
- Discrete valued feedback: Q-model
- Continuous valued feedback : S-model
- Finite action Learning Automata : FALA
- Continuous action Learning Automata : CALA

GENERAL S-MODEL

Reward penalty, L_{R-P}

$$p_i(k+1) = p_i(k) + a \cdot r(k)\big(1 - p_i(k)\big) - b \cdot \big(1 - r(k)\big)p_i(k), \text{ with } i \text{ the action taken}$$

$$p_j(k+1) = p_j(k) - a \cdot r(k)p_j(k) + b \cdot \big(1 - r(k)\big)\Big[\big(l-1\big)^{-1} - p_j(k)\Big], \text{ for all } j \neq i$$

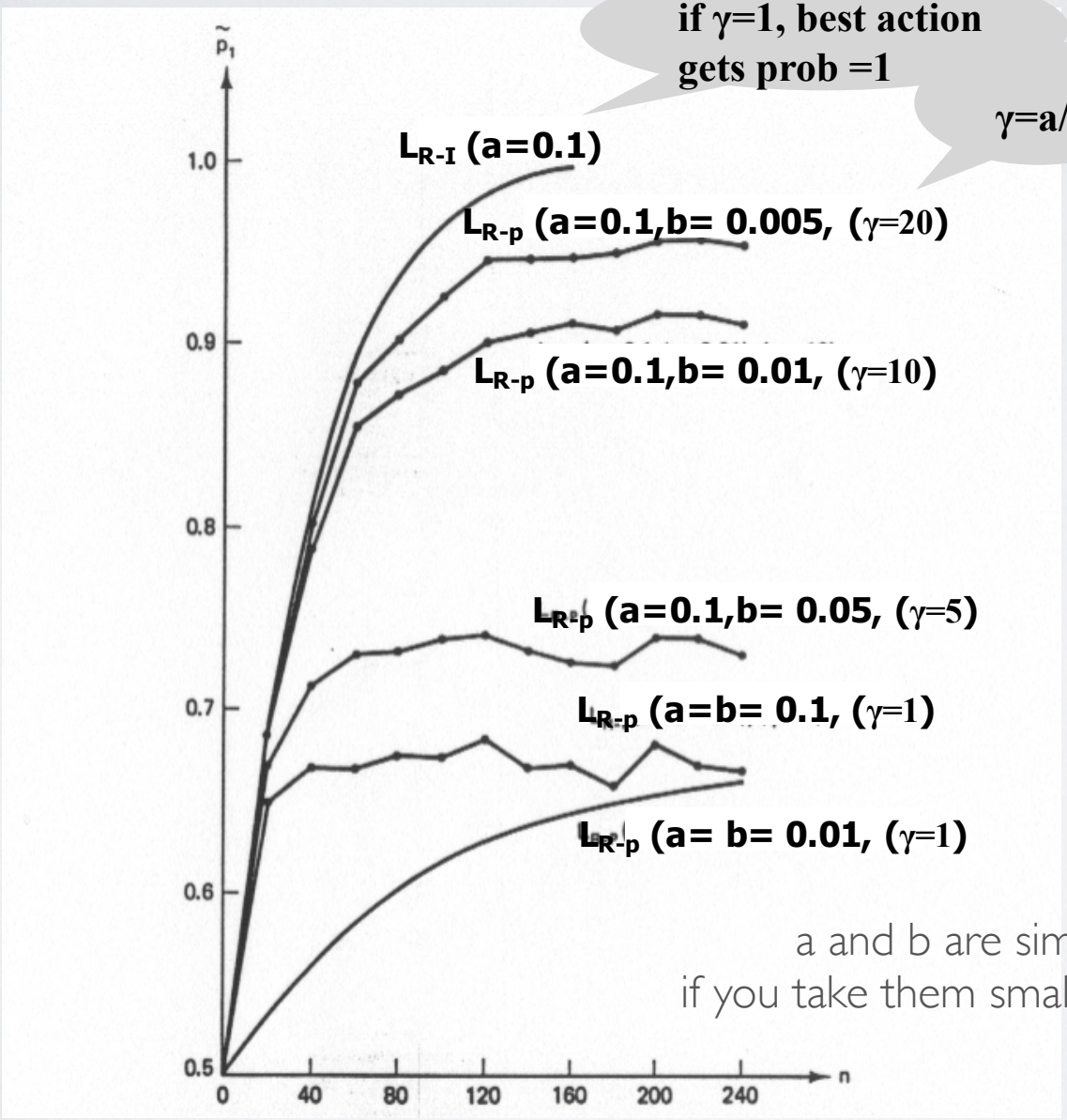
with r(k) real valued reward signal

If b << a : Reward- ε penalty, $L_{R-\varepsilon P}$

If b = 0 : Reward-in-action, L_{R-I}

Action selection for LA is implicit,
i.e. directly based on the action probabilities

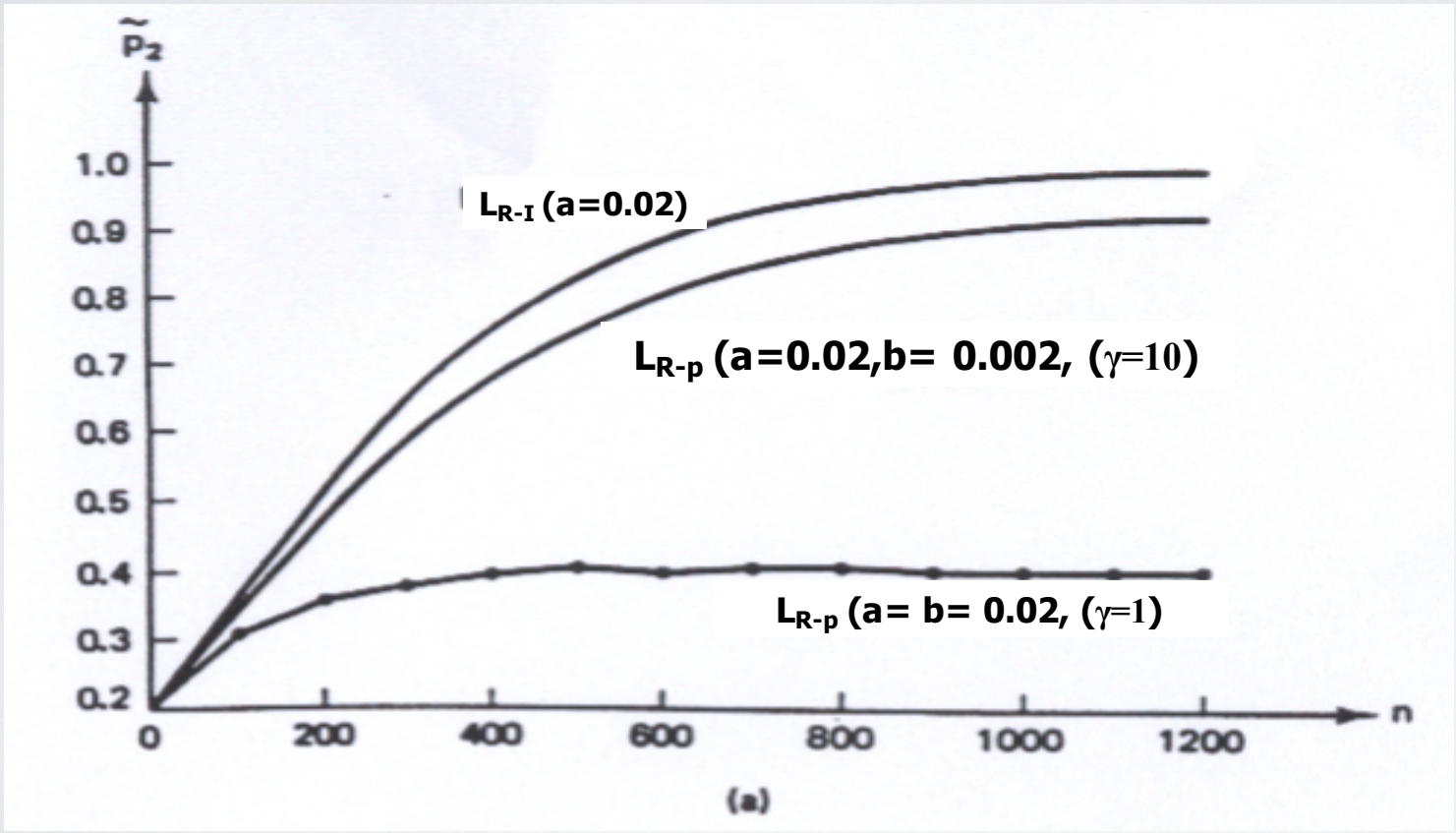
LEARNING AUTOMATA, A SIMULATION



2 actions
reward probabilities :
c_1 = 0.6, c_2 = 0.2

a and b are similar to learning rates,
if you take them small, convergence is smoother

LEARNING AUTOMATA, A SIMULATION



5 actions
reward probabilities:
c_1 = 0.35, c_2 = 0.8,
c_3 = 0.5, c_4 = 0.6,
c_5 = 0.15

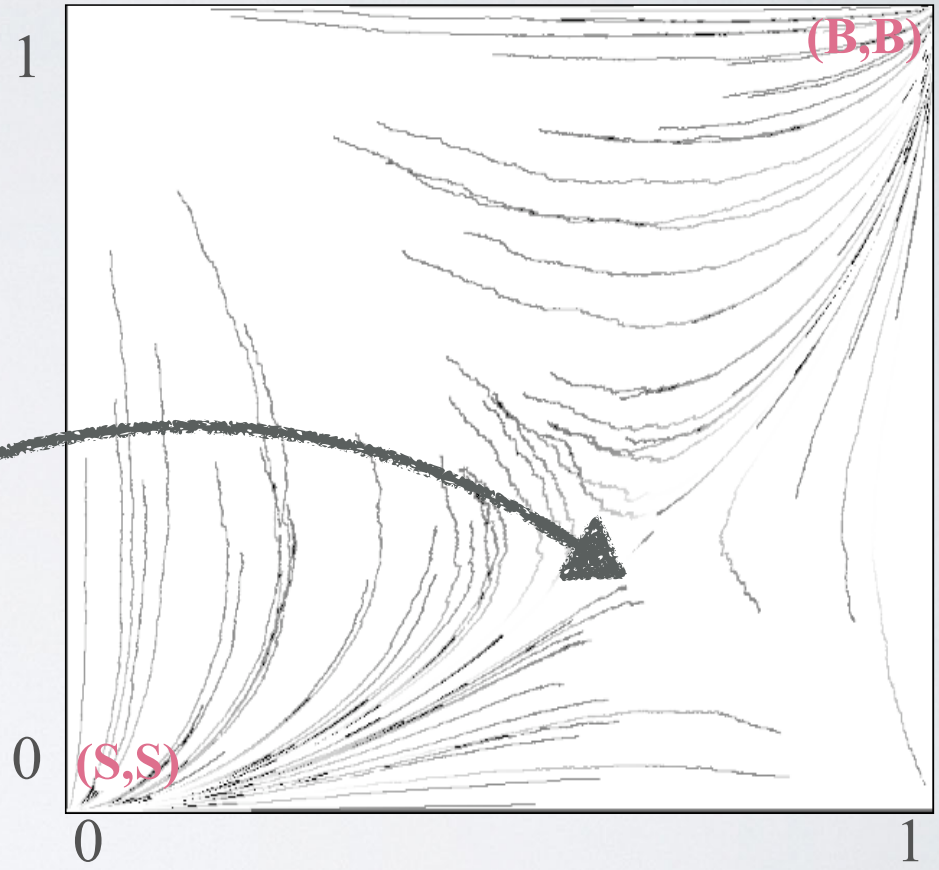
LA IN STRATEGIC GAMES

Battle of the sexes

	Bach	Strav.
Bach	2,1	0
Strav.	0	1,2

2 pure Nash equilibria

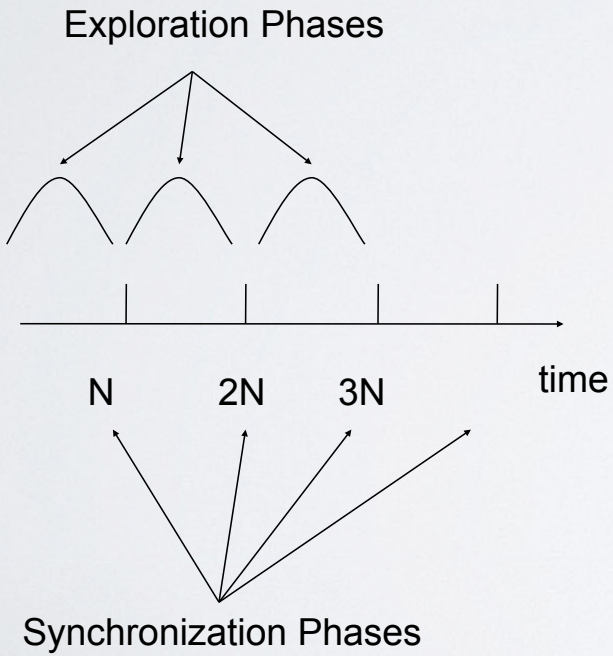
1 mixed Nash equilibrium
((2/3 B, 1/3 S) , (1/3 B, 2/3 S))



Paths induced by a linear reward-inaction LA.
Starting points are chosen randomly
x-axis = prob. of the first player to play Bach
y-axis = prob. of the second player to play Bach

EXPLORING SELFISH RL

Exploring selfish reinforcement learners (**ESRL**)



Basic idea: 2 phases

Exploration: Be Selfish

- Independent learning
- Convergence to different NE and Pareto Optimal non-NE

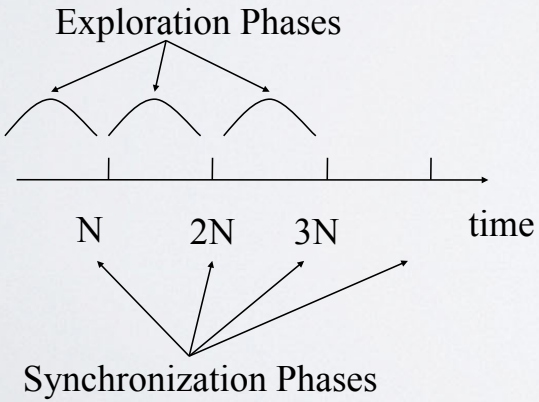
ESRL AND STRATEGIC GAMES

The Penalty Game

Player B

Player A			
	10,10	0,0	k,k
	0,0	2,2	0,0
	k,k	0,0	10,10

Witk k < 0



Synchronization: Be Social

- Exclusion phase: shrink the action space by excluding an action

ESRL AND STRATEGIC GAMES

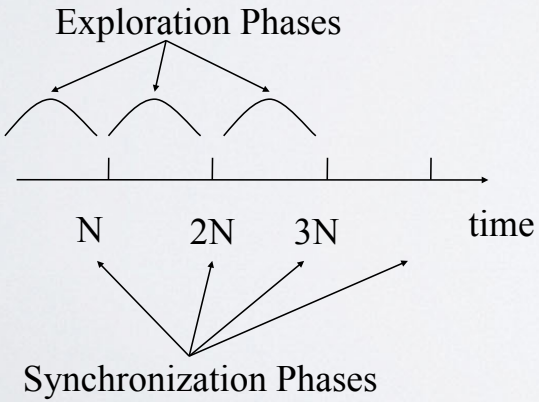
The Penalty Game

Player B

Player A

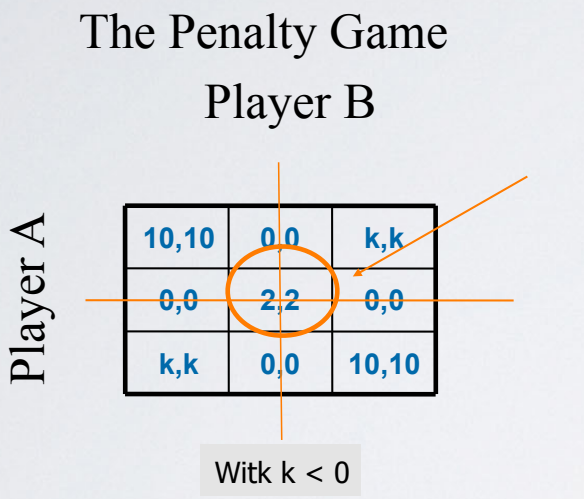
10,10	0,0	k,k
0,0	2,2	0,0
k,k	0,0	10,10

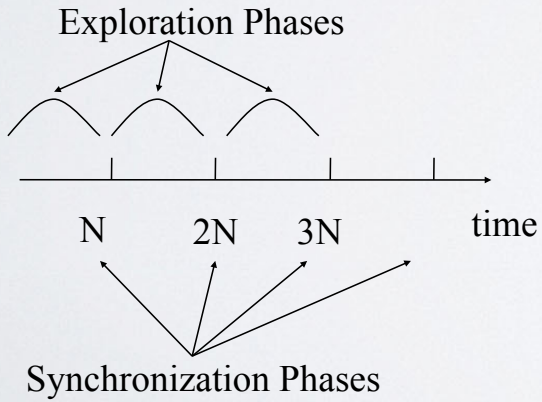
Witk k < 0



- Exploration
  - Use LRI → the agents converge to pure (Nash) joint action
- Synchronization
  - Exclude action *a* and explore again if empty *action* set → RESET
- If done, select BEST

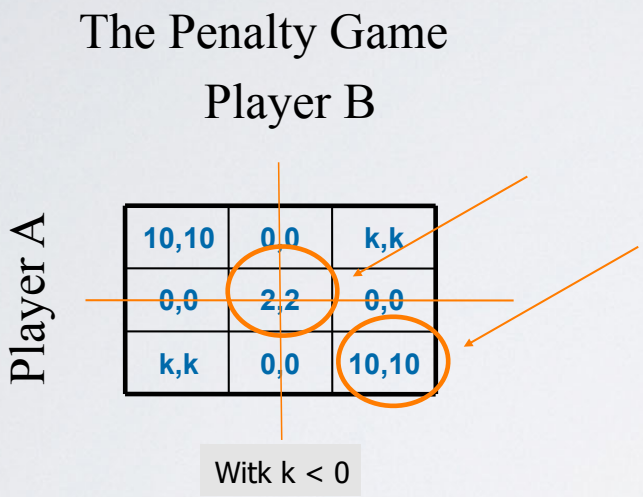
ESRL AND STRATEGIC GAMES

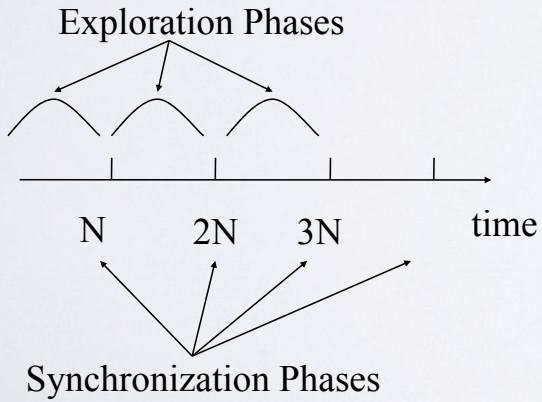




- Exploration
  - Use LRI → the agents converge to pure (Nash) joint action
- Synchronization
  - Exclude action *a* and explore again if empty *action* set → RESET
- If done, select BEST

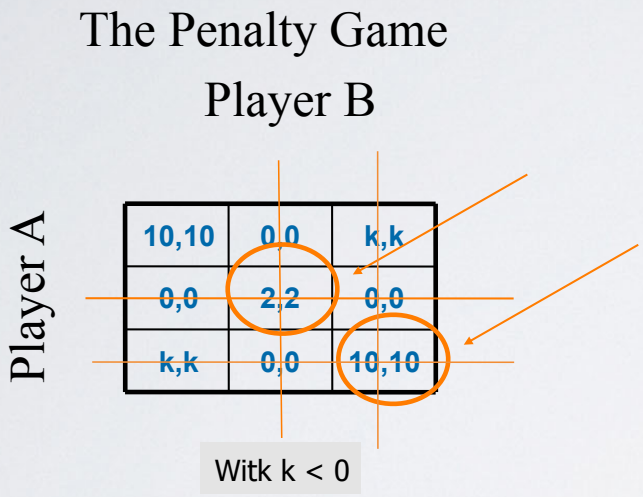
ESRL AND STRATEGIC GAMES

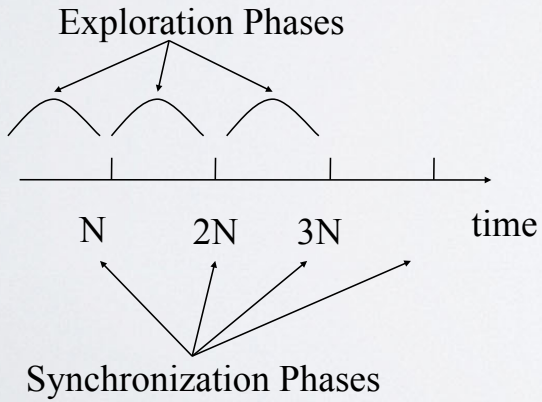




- Exploration
  - Use LRI → the agents converge to pure (Nash) joint action
- Synchronization
  - Exclude action *a* and explore again if empty *action* set → RESET
- If done, select BEST

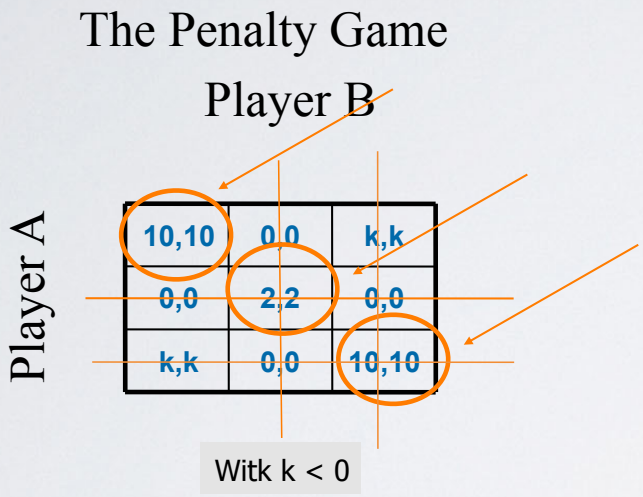
ESRL AND STRATEGIC GAMES

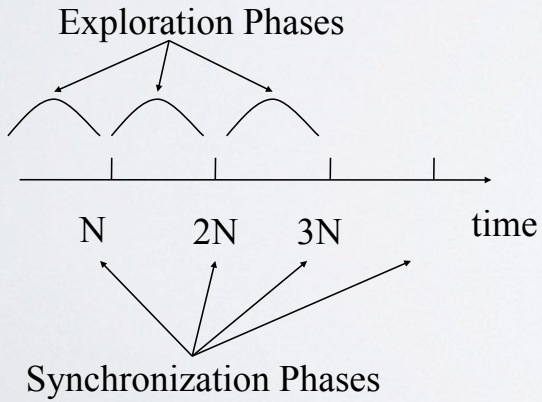




- Exploration
  - Use LRI → the agents converge to pure (Nash) joint action
- Synchronization
  - Exclude action *a* and explore again if empty *action* set → RESET
- If done, select BEST

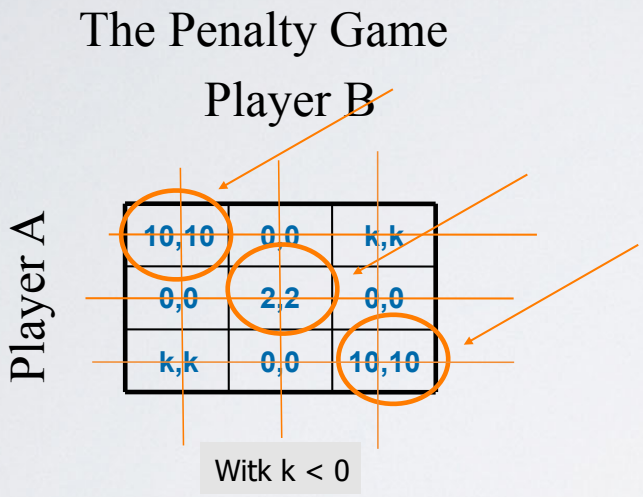
ESRL AND STRATEGIC GAMES

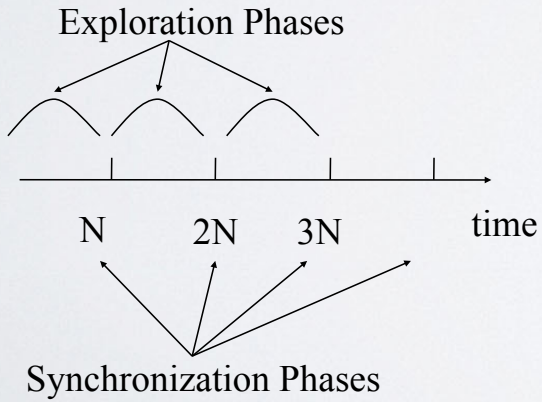




- Exploration
  - Use LRI → the agents converge to pure (Nash) joint action
- Synchronization
  - Exclude action *a* and explore again if empty *action* set → RESET
- If done, select BEST

ESRL AND STRATEGIC GAMES





- Exploration
  - Use LRI → the agents converge to pure (Nash) joint action
- Synchronization
  - Exclude action *a* and explore again if empty *action* set → RESET
- If done, select BEST

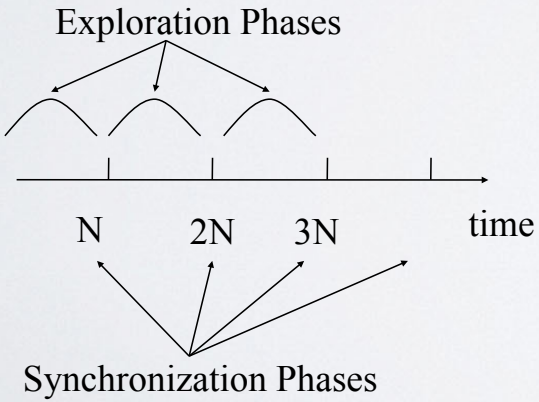
ESRL AND STRATEGIC GAMES

The Penalty Game

Player B

Player A			
	10,10	0,0	k,k
	0,0	2,2	0,0
	k,k	0,0	10,10

Witk k < 0



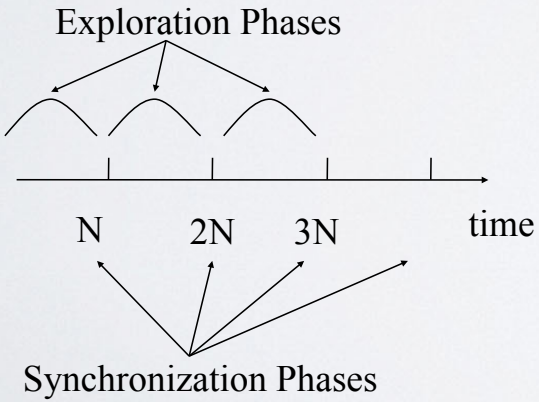
- Exploration
  - Use LRI → the agents converge to pure (Nash) joint action
- Synchronization
  - Exclude action *a* and explore again if empty *action* set → RESET
- If done, select BEST

ESRL IN CONFLICTING INTEREST GAMES

Battle of the sexes

Player 2

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



- Exploration
  - Use LRI → the agents converge to pure (Nash) joint action
- Synchronization
  - Exclude action *a* and explore again if empty *action* set → RESET
- If done, select BEST

Can be extended to more players see: Verbeeck 2004

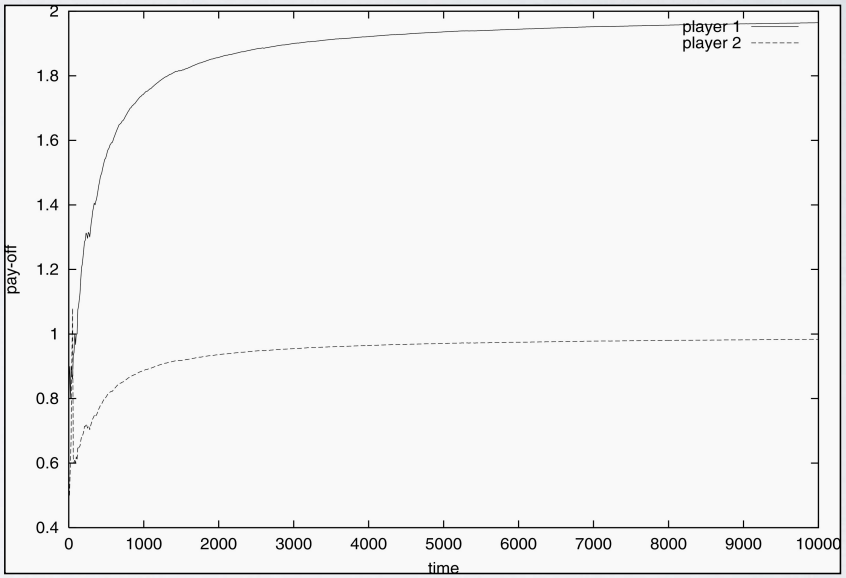
ESRL IN CONFLICTING INTEREST GAMES

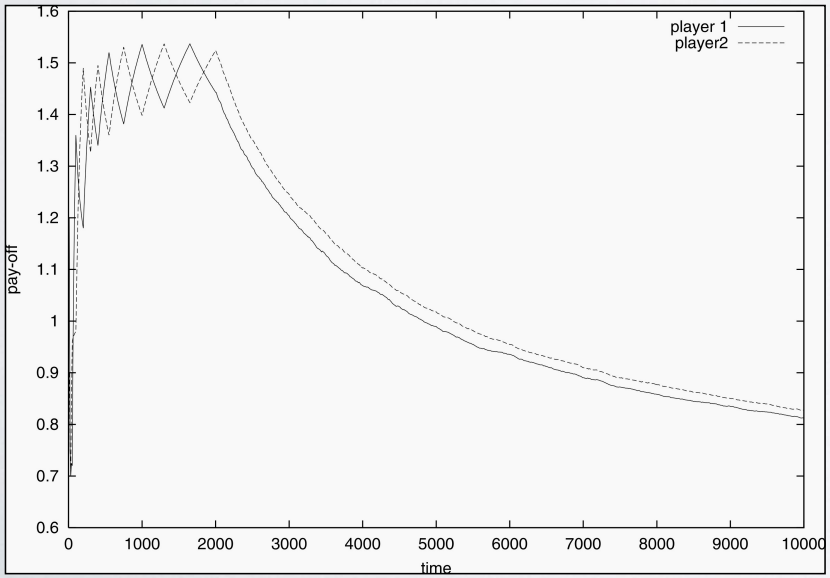
Battle of the sexes

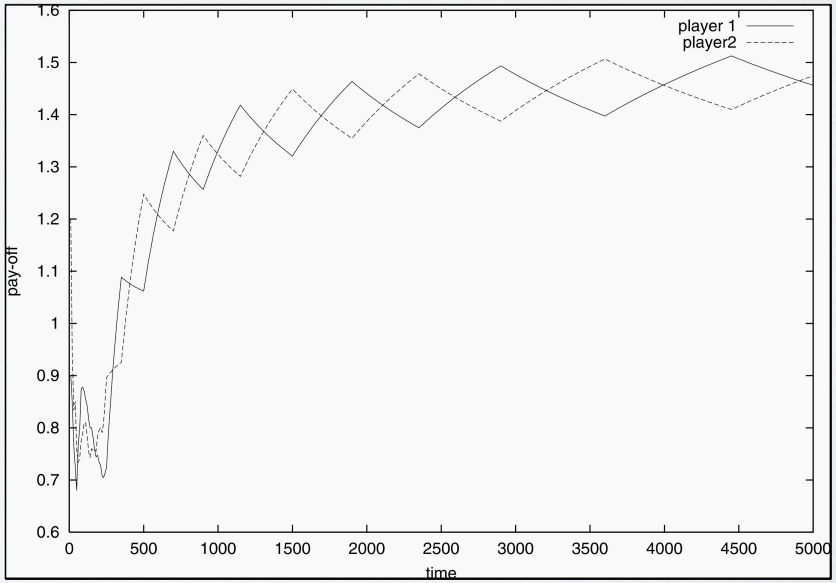
Player 2

Player 1

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







SUMMARY

- "Best" algorithm depends on the type of game (common interest vs conflicting interest, deterministic vs stochastic, (a-)synchronous action taking, observability of other agents, etc.)
- Other approaches for co-ordination in networks, Markov Games, **graphical games, routing games and sparse interactions**