Multiagent Learning in the Presence of Limited Agents

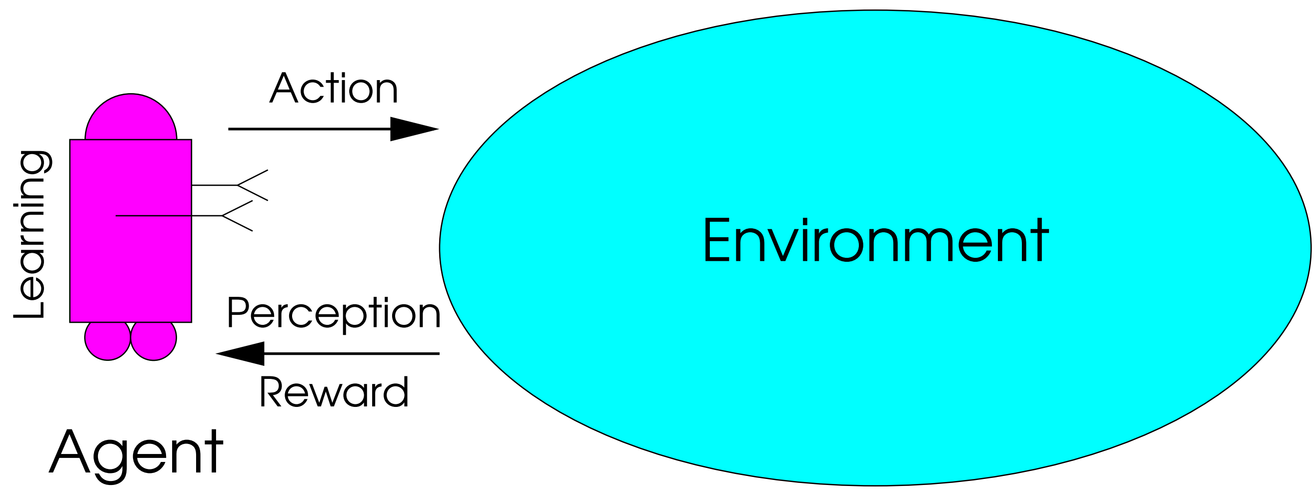
Michael Bowling
Computer Science Department
Carnegie Mellon University

Job Talk

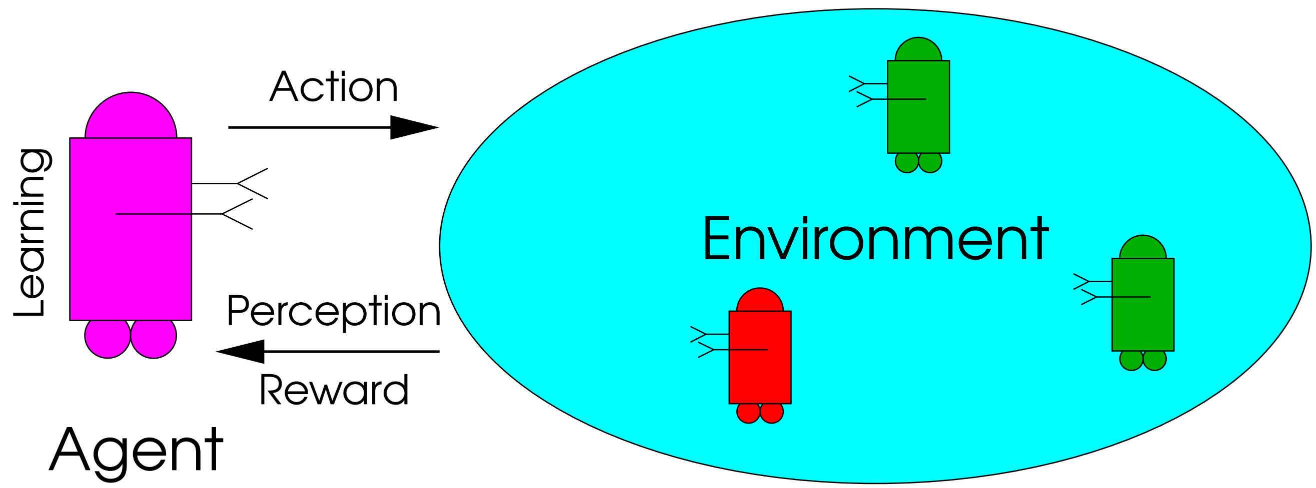
Based on joint work with Manuela Veloso.

What is Multiagent Learning?

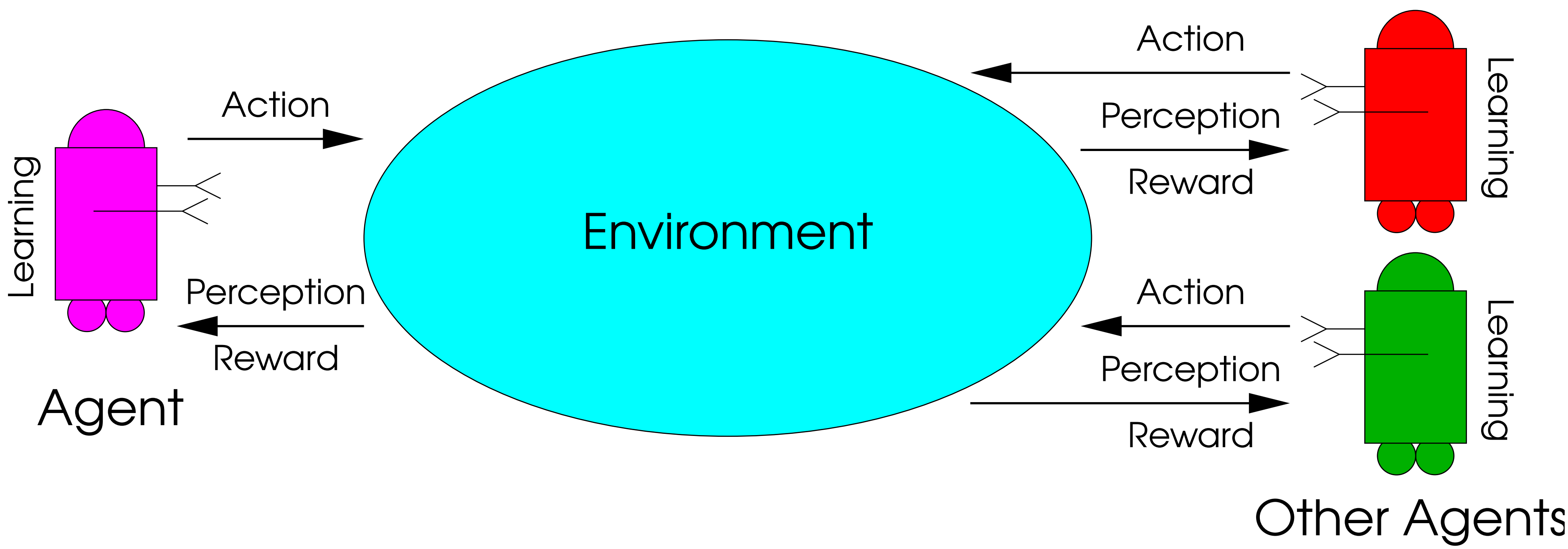
What is Multiagent Learning?



What is Multiagent Learning?



What is Multiagent Learning?



Why Limited Agents?

Why Limited Agents?

- Optimal action selection is impractical.
- Agent behavior is at best near optimal.
- "Bounded Rationality"

Examples – Goofspiel

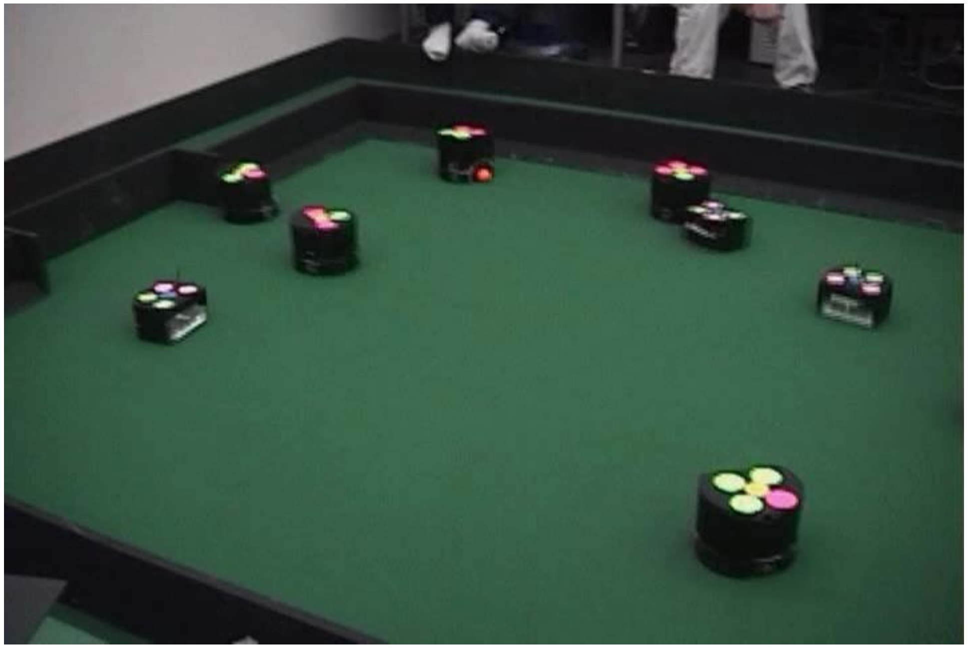
- Players hands and the deck have cards $1 \ldots n$.
- Card from the deck is bid on secretly.
- Highest card played gets points equal to the deck card.
- Both players discard the cards bid.
- Repeat for all n deck cards.

Examples – Goofspiel

- Players hands and the deck have cards $1 \dots n$.
- Card from the deck is bid on secretly.
- Highest card played gets points equal to the deck card.
- Both players discard the cards bid.
- Repeat for all n deck cards.

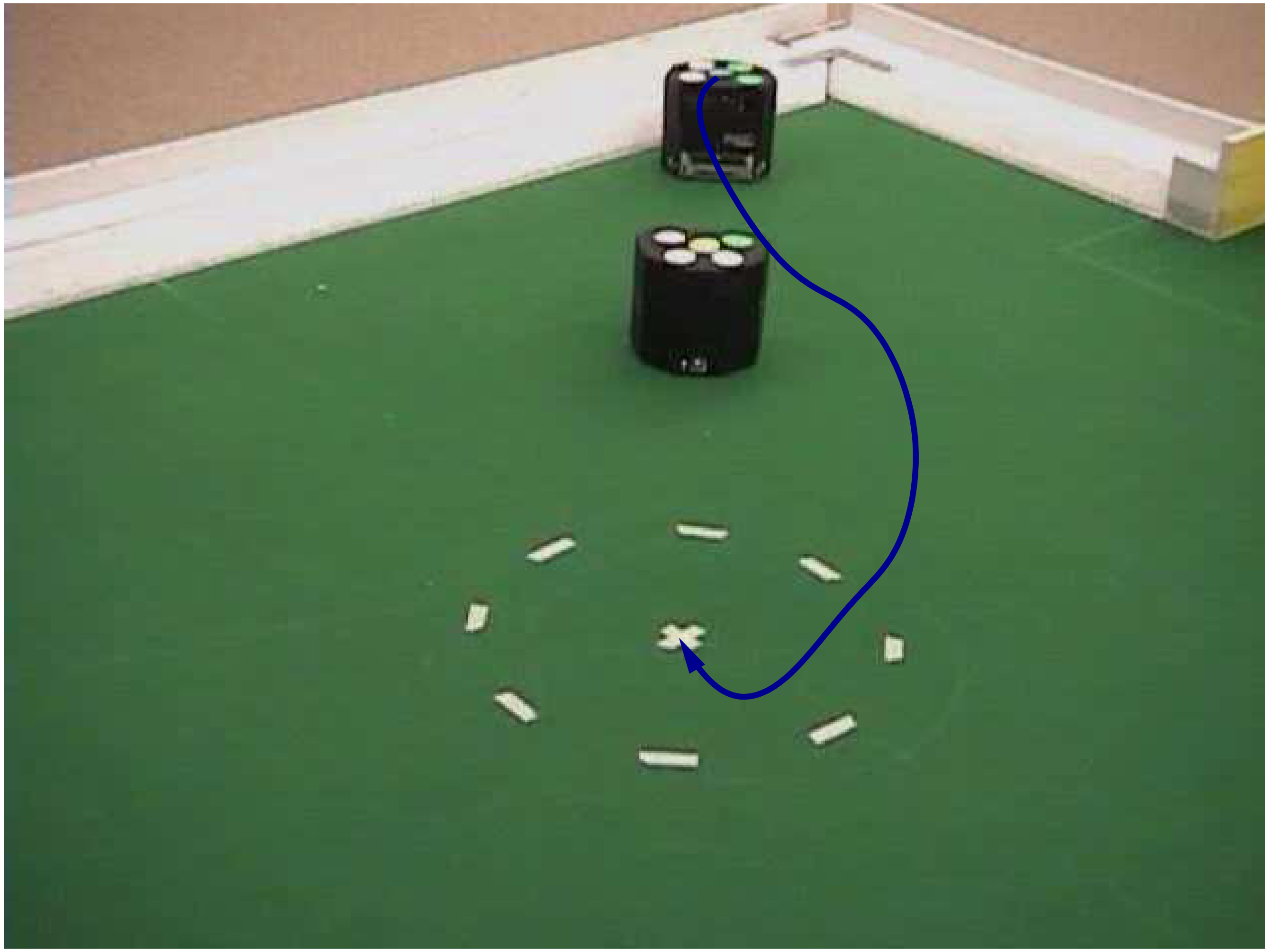
n	$\|S\|$	$\|S \times A\|$	SIZEOF(π or Q)	VALUE(det)	VALUE(random)
4	692	15150	$\sim$ 59KB	-2	-2.5
8	3×10^6	1×10^7	$\sim$ 47MB	-20	-10.5
13	1×10^{11}	7×10^{11}	$\sim$ 2.5TB	-65	-28

Examples – Robot Soccer

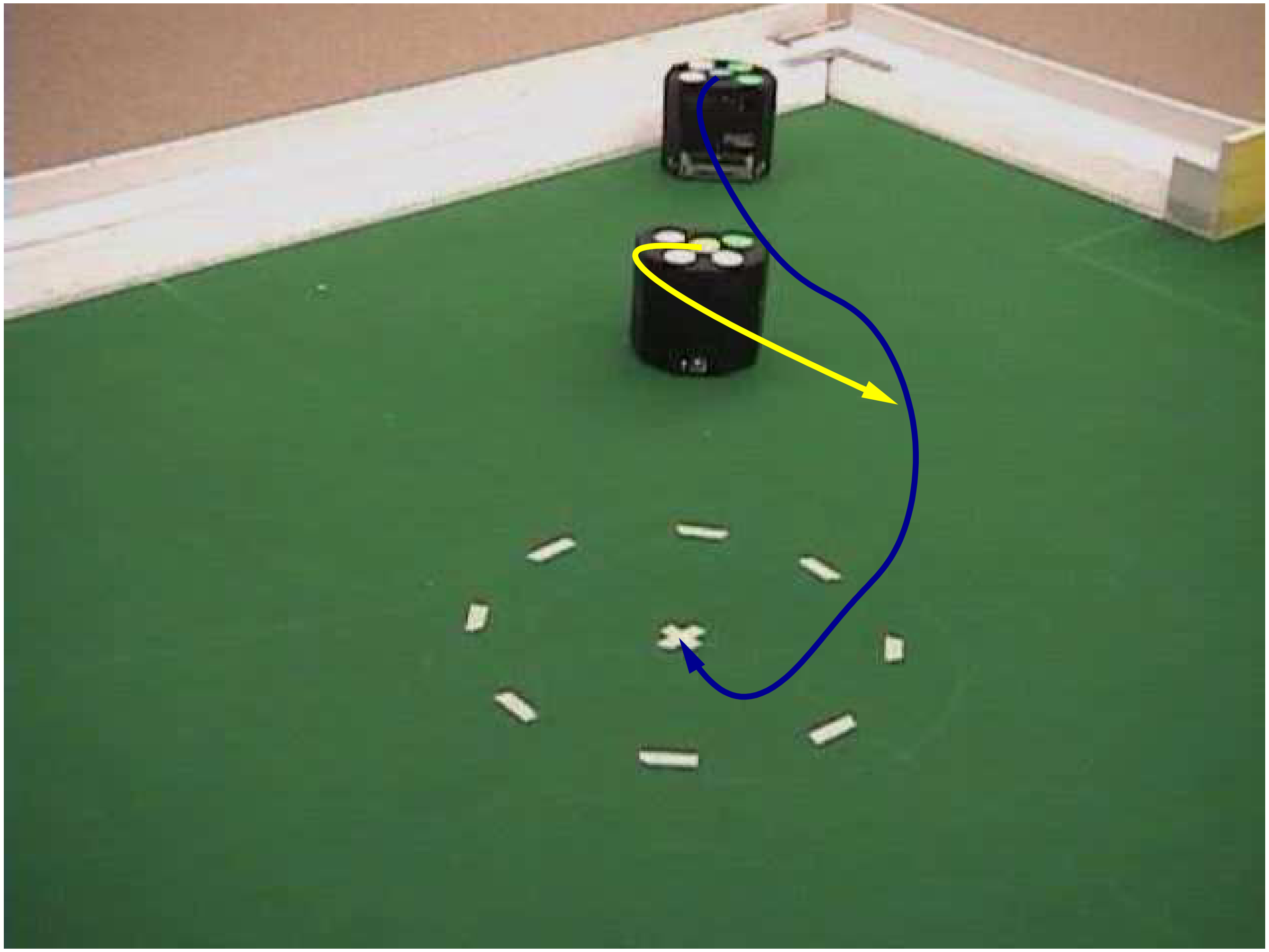


Examples – Keepout

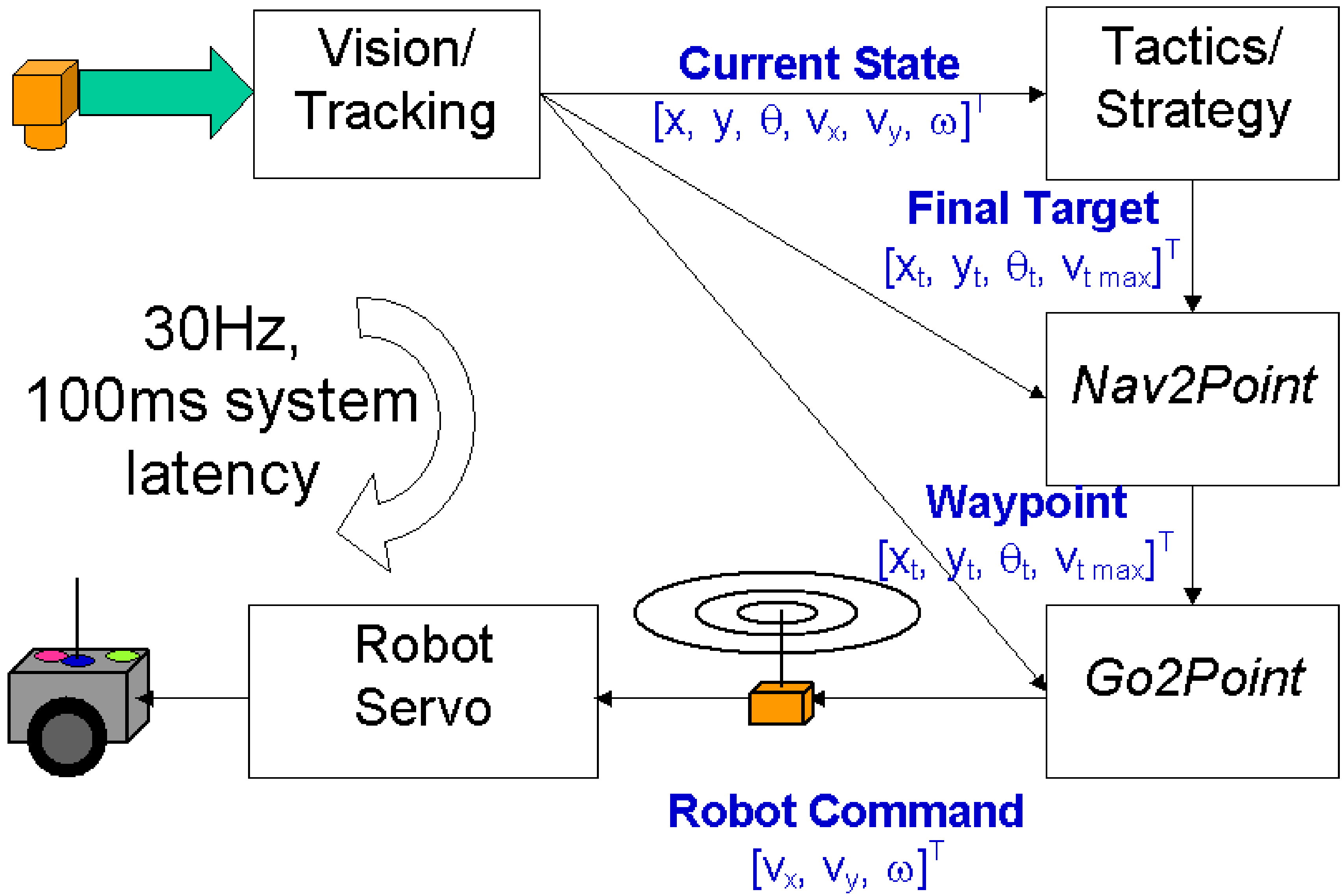
Examples – Keepout



Examples – Keepout



Examples – Keepout – 2



Why is This Hard?

Why is This Hard?

- Multiagent Learning
  - Optimal behavior depends on the other agents.
  - Other agents may be learning as well.
  - Deterministic policies can often be exploited.

- Limitations
  - Agents cannot act optimally.
    * Intractably large or continuous state spaces.
    * Situated learning among fixed components.
    * Latency as Partial Observability
  - Applies to “us” as well as “them”.

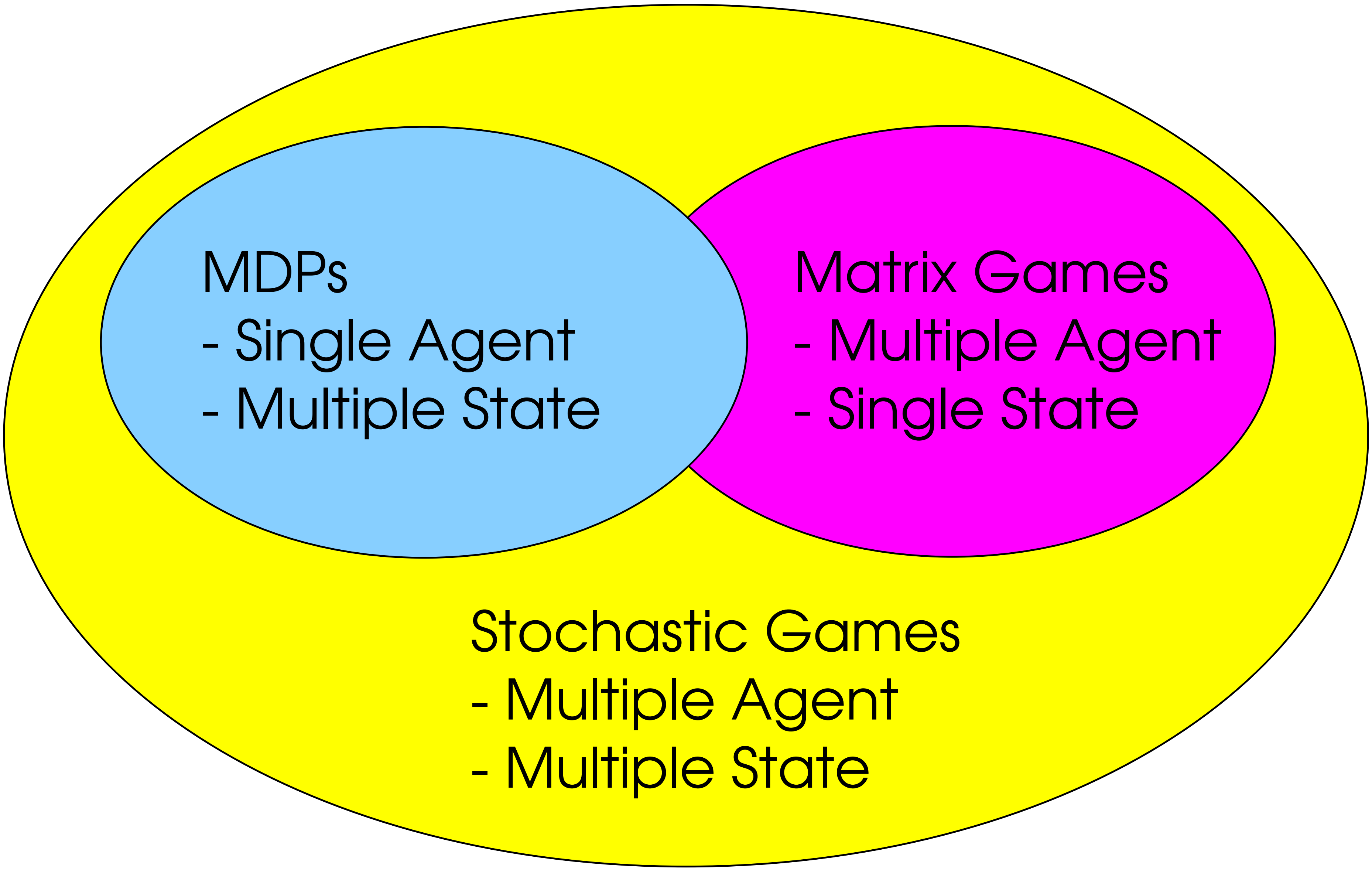
Examples – Other Applications

- Keepout
  - Robot Soccer
  - Search and Rescue
  - Automated Driving
- Goofspiel
  - Auctions with Limited Resources
  - Electronic Commerce
  - Artificial Markets
- Many environments involve goal-directed agents
  - Personal Assistants, Negotiators
  - Computer Games (Agent-Human Interaction)

Overview

- Motivation
- Stochastic Games
- WoLF: Rational and Convergent Learning
  - Theoretical results in matrix games
  - Empirical results in “small” stochastic games
- GraWoLF: Learning with Limitations
  - Limitations and Equilibria
  - GraWoLF Algorithm
  - Empirical results of learning in Goofspiel and Keepout
- Summary and Future Work

Stochastic Games



SCS Carnegie Mellon

Matrix Games – Examples

- Matching Pennies
  - Players: Two
  - Actions: Heads (H) or Tails (T)
  - The rules:

Player One	wins	if actions are the same
Player Two	wins	if actions are different

Matrix Games – Examples

- Matching Pennies
  - Players: Two
  - Actions: Heads (H) or Tails (T)
  - The rules:

Player One	wins	if actions are the same
Player Two	wins	if actions are different

$$R_1 = \begin{array}{c} \\ \text{H} \\ \text{T} \end{array} \begin{array}{c} \begin{array}{cc} \text{H} & \text{T} \end{array} \\ \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \end{array} \qquad R_2 = \begin{array}{c} \\ \text{H} \\ \text{T} \end{array} \begin{array}{c} \begin{array}{cc} \text{H} & \text{T} \end{array} \\ \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \end{array}$$

Matrix Games – Examples – 2

- Rock-Paper-Scissors
  - Players: Two
  - Actions: Rock (R), Paper (P), or Scissors (S)
  - The rules:

Rock	beats	Scissors
Scissors	beats	Paper
Paper	beats	Rock

$$R_1 = \begin{array}{c} \\ \mathsf{R} \\ \mathsf{P} \\ \mathsf{S} \end{array} \begin{array}{c} \begin{array}{ccc} \mathsf{R} & \mathsf{P} & \mathsf{S} \end{array} \\ \left(\begin{array}{rrr} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{array} \right) \end{array} \qquad R_2 = \begin{array}{c} \\ \mathsf{R} \\ \mathsf{P} \\ \mathsf{S} \end{array} \begin{array}{c} \begin{array}{ccc} \mathsf{R} & \mathsf{P} & \mathsf{S} \end{array} \\ \left(\begin{array}{rrr} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{array} \right) \end{array}$$

Matrix Games – Equilibria

- No optimal opponent independent strategies.
- Best-responses

  The set of all strategies that are optimal given the strategies of the other players.

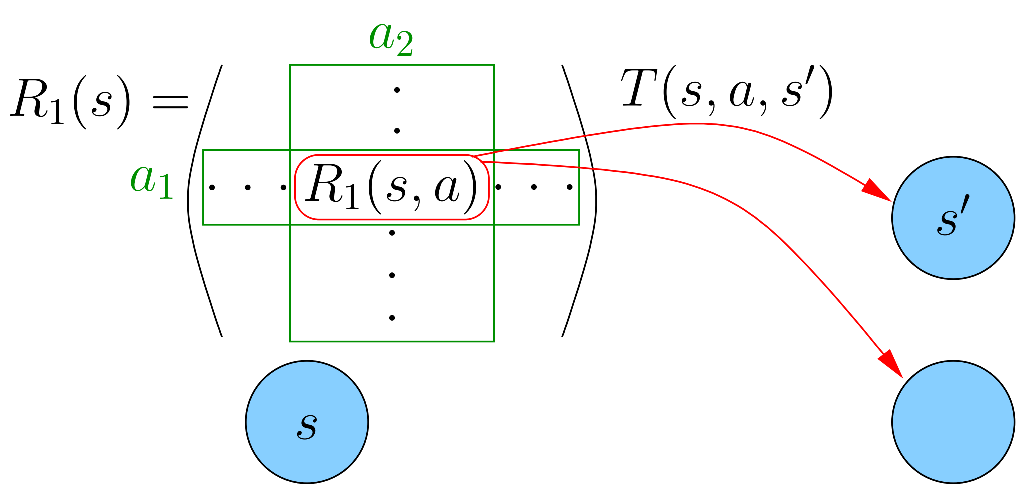
- Nash Equilibrium (Nash, 1950)

  A strategy for each player, such that each is playing a best-response to the others' strategies. No player wants to deviate.

Stochastic Games

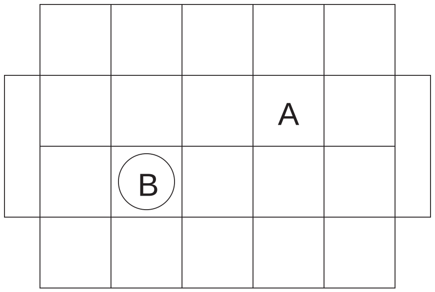
A *stochastic game* is a tuple $(n, \mathcal{S}, \mathcal{A}_{1\ldots n}, T, R_{1\ldots n})$,

- n is the number of agents,
- $\mathcal{S}$ is the set of states,
- $\mathcal{A}_i$ is the set of actions available to agent i,
  - $\mathcal{A}$ is the joint action space $\mathcal{A}_1 \times \ldots \times \mathcal{A}_n$,
- T is the transition function $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$,
- R_i is the reward function for the ith agent $\mathcal{S} \times \mathcal{A} \rightarrow \Re$.



Stochastic Games – Example

(Littman, 1994)



- Players: Two
- States: Player positions and ball possession (780).
- Actions: N, S, E, W, Hold (5).
- Transitions:
  - Simultaneous action selection, random execution.
  - Collision could change ball possession.
- Rewards: Ball enters a goal.

Stochastic Games – Equilibria

- Goal is to learn a policy, $\pi : \mathcal{S} \rightarrow PD(A_i)$.
- No optimal opponent independent policies.
- Best-responses

  The set of all policies that are optimal given the policies of the other players.
- Nash Equilibrium (Shapley, 1953; Fink 1964)

  A policy for each player, such that each is playing a best-response to the others' policies. No player wants to deviate.

Overview

- Motivation
- Stochastic Games
- WoLF: Rational and Convergent Learning
  - Theoretical results in matrix games
  - Empirical results in “small” stochastic games
- GraWoLF: Learning with Limitations
  - Limitations and Equilibria
  - GraWoLF Algorithm
  - Empirical results of learning in Goofspiel and Keepout
- Summary and Future Work

Learning Properties

Learning Properties

- Rational.
  We want to learn best responses, if possible.

Learning Properties

- Rational.
  We want to learn best responses, if possible.
- Convergent.
  We want to converge, if possible.
- Convergent in Self-Play.
  Opponents are at least as sophisticated as ourselves.

Learning Properties

- Rational.
  We want to learn best responses, if possible.

- Convergent.
  We want to converge, if possible.

- Convergent in Self-Play.
  Opponents are at least as sophisticated as ourselves.

If all players are rational and their policies converge, it must be to an equilibrium.

Learning Properties – 2

- How do previous algorithms do?

Learning Properties – 2

- How do previous algorithms do?
  - Single-Agent Learners (e.g., Q-learning, TD(λ))

Learning Properties – 2

- How do previous algorithms do?
  - Single-Agent Learners (e.g., Q-learning, TD(λ))

    Rational   Not Convergent

Learning Properties – 2

- How do previous algorithms do?
  - Single-Agent Learners (e.g., Q-learning, TD(λ))

    Rational Not Convergent

  - Best-Response Learners (e.g., JALs, Fictitious-Play)

Learning Properties – 2

- How do previous algorithms do?
  - Single-Agent Learners (e.g., Q-learning, TD(λ))

    Rational Not Convergent

  - Best-Response Learners (e.g., JALs, Fictitious-Play)

    Rational Not Convergent

Learning Properties – 2

- How do previous algorithms do?
  - Single-Agent Learners (e.g., Q-learning, TD(λ))

    Rational Not Convergent
  - Best-Response Learners (e.g., JALs, Fictitious-Play)

    Rational Not Convergent
  - Equilibrium Learners (e.g., Minimax-Q, Nash-Q, CE-Q)

Learning Properties – 2

- How do previous algorithms do?
  - Single-Agent Learners (e.g., Q-learning, TD(λ))

    Rational   Not Convergent
  - Best-Response Learners (e.g., JALs, Fictitious-Play)

    Rational   Not Convergent
  - Equilibrium Learners (e.g., Minimax-Q, Nash-Q, CE-Q)

    Not Rational   Convergent

- Goal: We want the rationality of best-response learners and the convergence of equilibrium learners.

WoLF: Win or Learn Fast

- Intuition: Don't want to "overfit" to a changing policy.
- Intuition: Learning should be cautious if doing too well.

WoLF: Win or Learn Fast

- Intuition: Don't want to "overfit" to a changing policy.
- Intuition: Learning should be cautious if doing too well.

- Idea #1: Variable Learning Rate.
  - Change the speed of learning over time.
- Idea #2: WoLF — "Win or Learn Fast".
  - If we're winning, we learn cautiously.
  - If we're losing, we learn quickly.
  - Winning == Doing better than playing the equilibrium.

- Can make rational, non-convergent algorithms converge!
  - Theoretical Results.
  - Empirical Results.

Theoretical Results

- Learning in two-player, two-action matrix games.
- Gradient ascent (Singh, Kearns, & Mansour, 2000)
- Modify with WoLF.

Gradient Ascent

$$R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \qquad C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

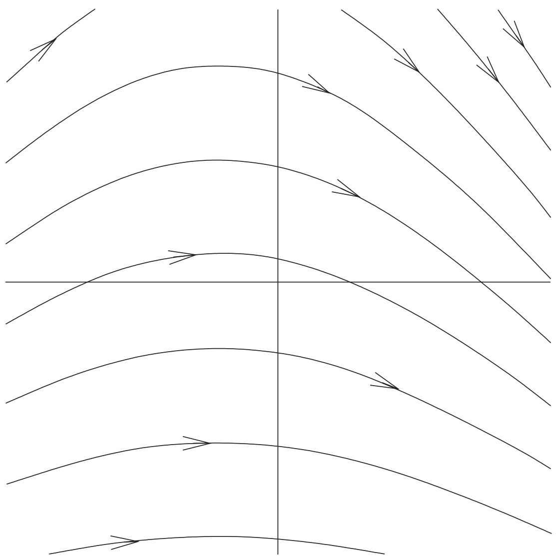
Gradient Ascent

$$R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \quad C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

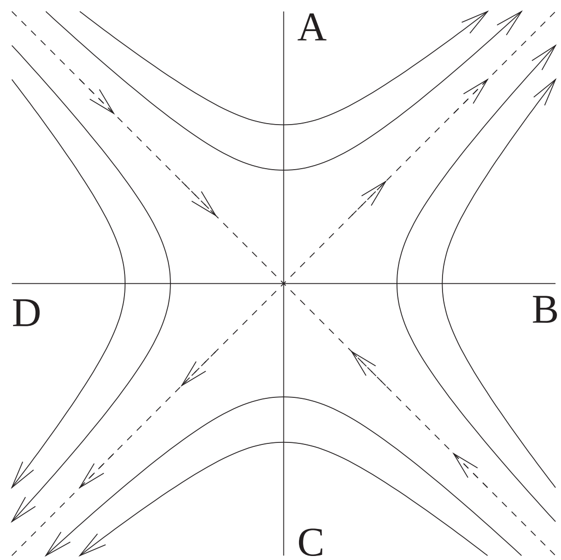
$$\begin{aligned} \alpha_{k+1} &= \alpha_k + \eta \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \alpha_k} \\ \beta_{k+1} &= \beta_k + \eta \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \beta_k} \end{aligned}$$

Gradient Ascent – 3

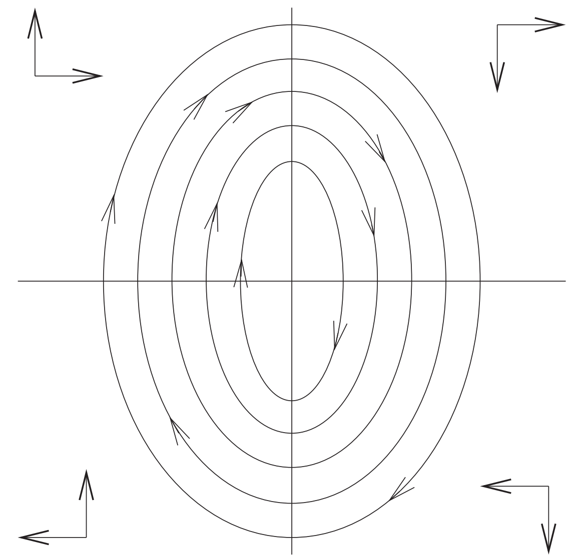
$$\begin{bmatrix} \frac{\partial \alpha}{\partial t} \\ \frac{\partial \beta}{\partial t} \end{bmatrix} = \begin{bmatrix} 0 & u \\ u' & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} (r_{12} - r_{22}) \\ (c_{21} - c_{22}) \end{bmatrix}.$$



U is not invertible
$u = 0$ or $u' = 0$



U has real eigenvalues
$uu' < 0$



U has imaginary eigenvalues
$uu' > 0$

WoLF Gradient Ascent

$$
\begin{aligned}
\alpha_{k+1} &= \alpha_k + \eta \ell_k^r \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \alpha} \\
\beta_{k+1} &= \beta_k + \eta \ell_k^c \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \beta}
\end{aligned}
$$

$$\ell_k^{r,c} \in [\ell_{\min}, \ell_{\max}] > 0$$

WoLF Gradient Ascent – 2

$$\alpha_{k+1} = \alpha_k + \eta \ell_k^r \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \alpha}$$

$$\beta_{k+1} = \beta_k + \eta \ell_k^c \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \beta}$$

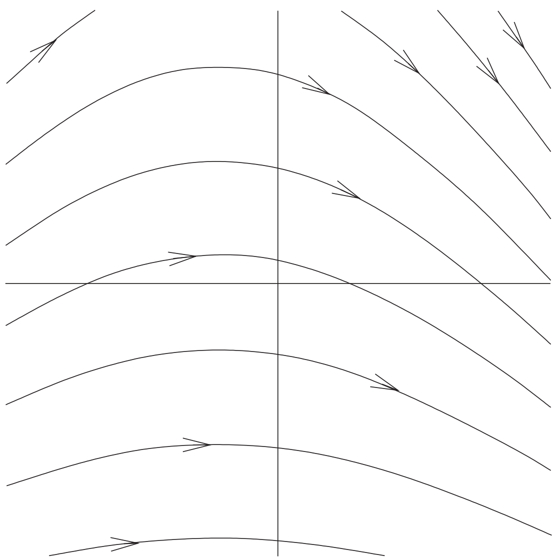
WoLF = Win or Learn Fast!

$$\ell_k^r = \begin{cases} \ell_{\min} & \text{WINNING} \quad \text{if } V_r(\alpha_k, \beta_k) > V_r(\alpha^*, \beta_k) \\ \ell_{\max} & \text{LOSING} \quad \text{otherwise} \end{cases}$$

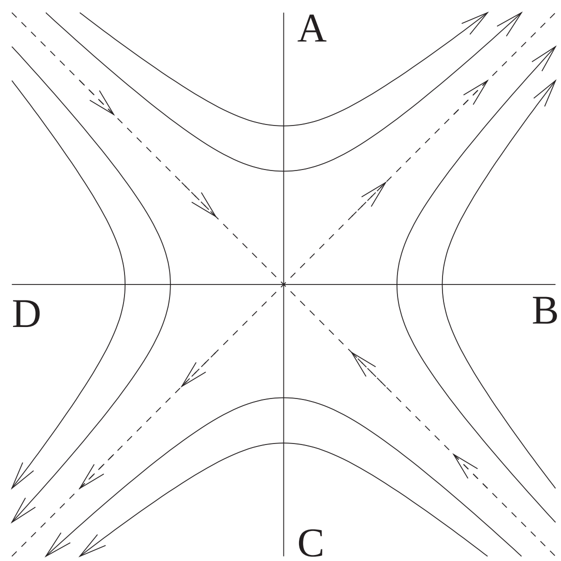
$$\ell_k^c = \begin{cases} \ell_{\min} & \text{WINNING} \quad \text{if } V_c(\alpha_k, \beta_k) > V_c(\alpha_k, \beta^*) \\ \ell_{\max} & \text{LOSING} \quad \text{otherwise} \end{cases}$$

WoLF Gradient Ascent – 3

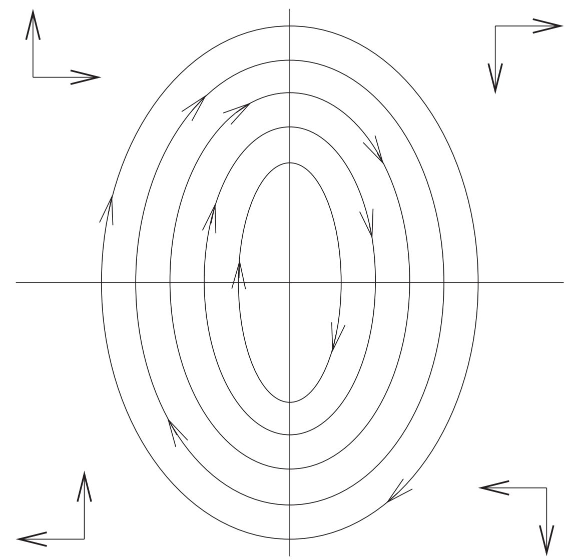
$$\begin{bmatrix} \frac{\partial \alpha}{\partial t} \\ \frac{\partial \beta}{\partial t} \end{bmatrix} = \begin{bmatrix} 0 & u\ell_r(t) \\ u'\ell_c(t) & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \ell_r(t)(r_{12} - r_{22}) \\ \ell_c(t)(c_{21} - c_{22}) \end{bmatrix}.$$



U is not invertible
$u = 0$ or $u' = 0$



U has real eigenvalues
$uu' < 0$



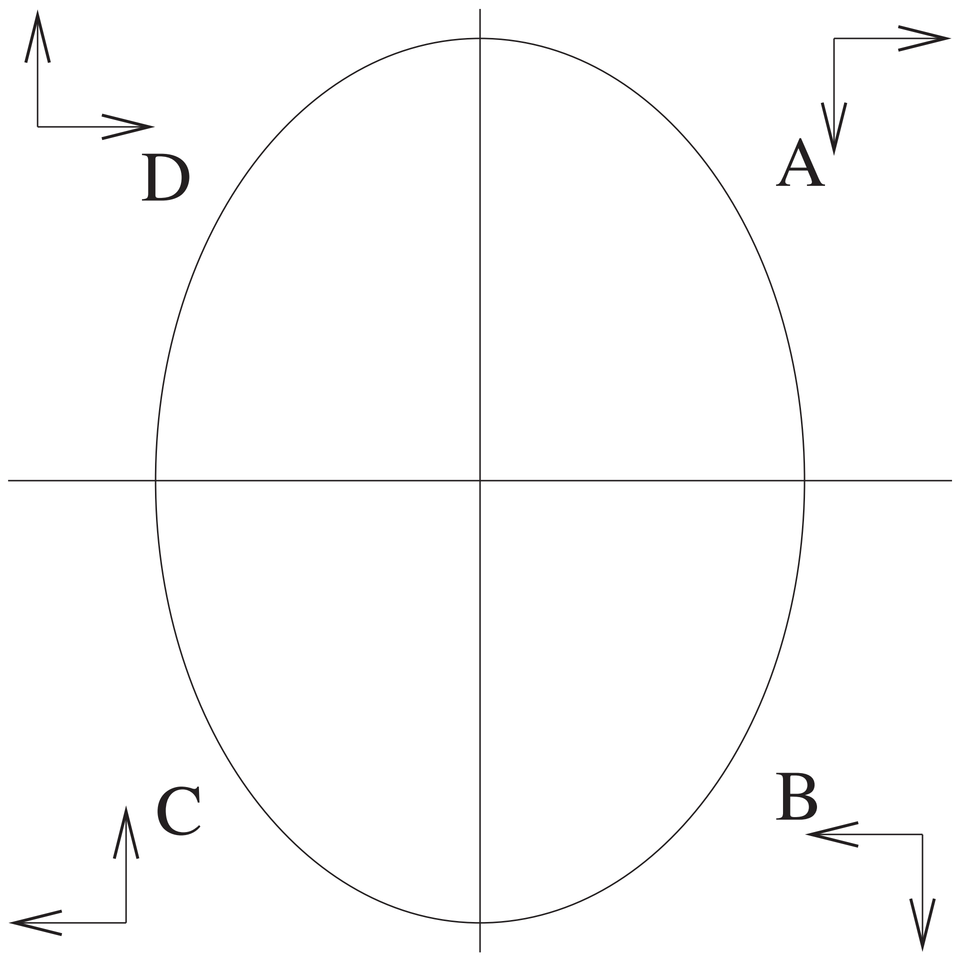
U has imaginary eigenvalues
$uu' > 0$

WoLF Gradient Ascent – 4

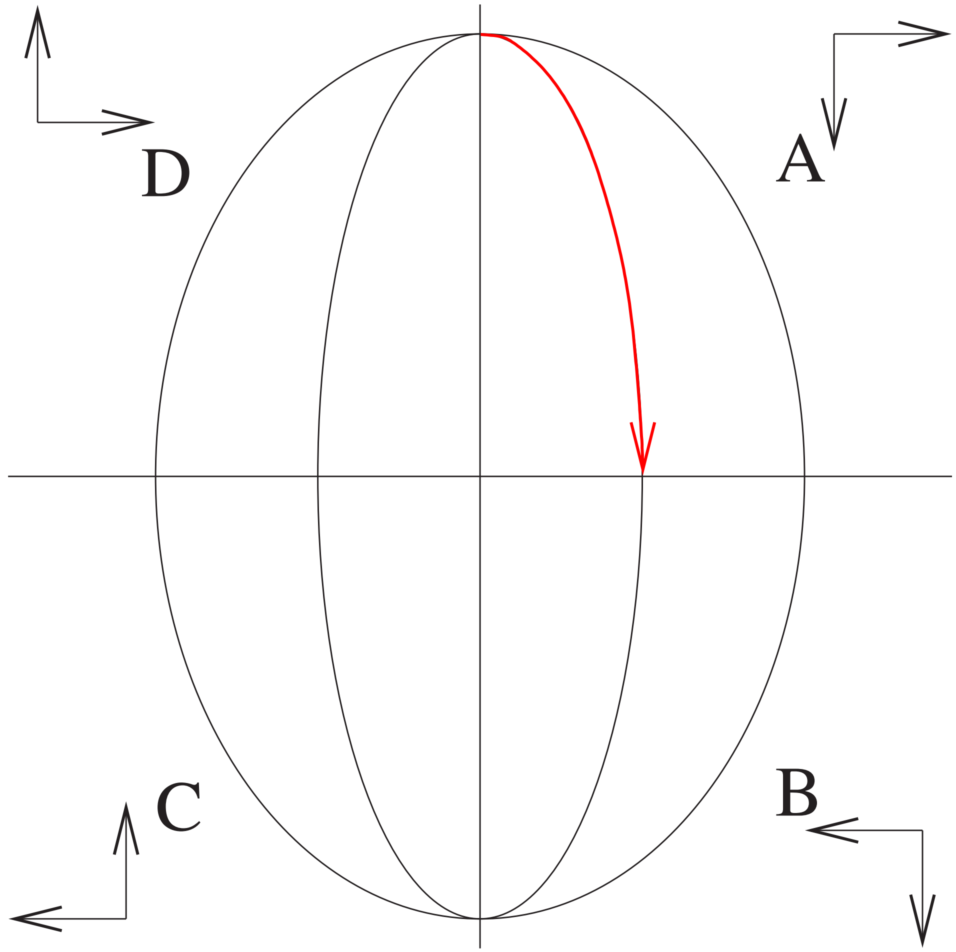
Lemma. A player's strategy is moving away from the equilibrium *if and only if* they are "winning".

I.e., $V_r(\alpha, \beta) - V_r(\alpha^*, \beta) > 0 \iff (\alpha - \alpha^*)\frac{\partial V_r(\alpha,\beta)}{\partial \alpha} > 0.$

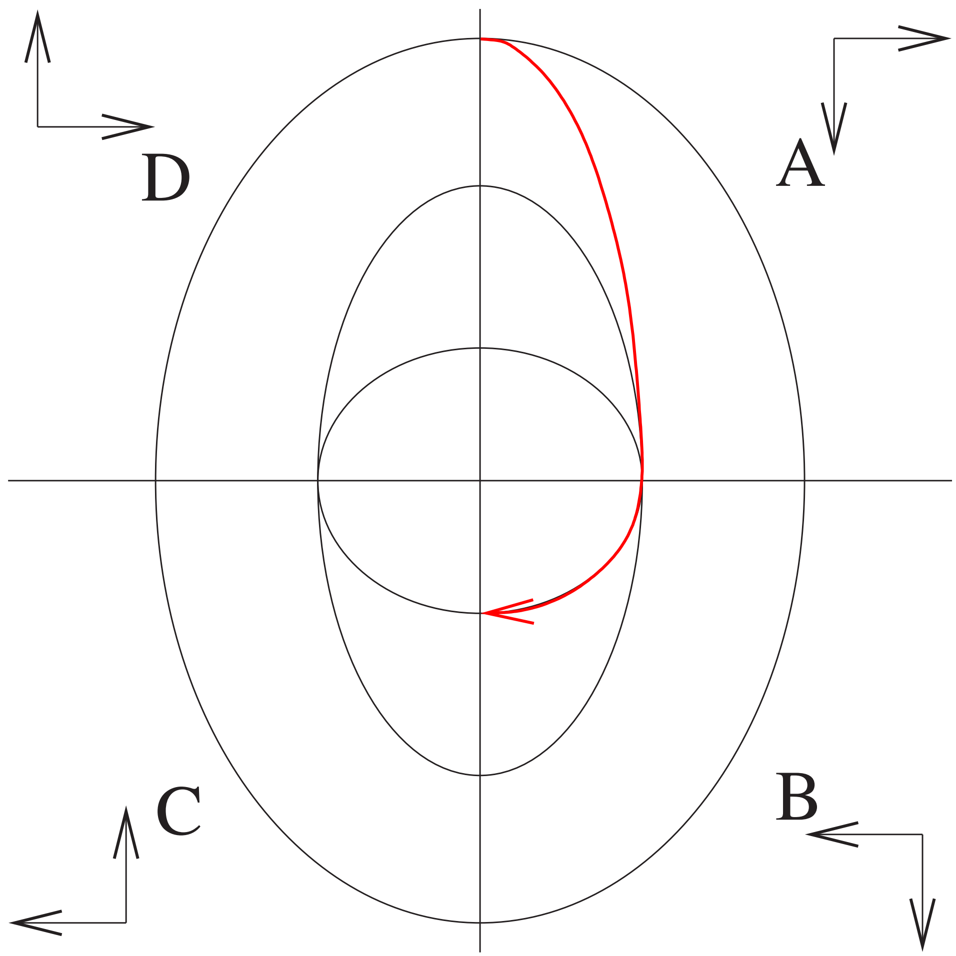
WoLF Gradient Ascent – 5



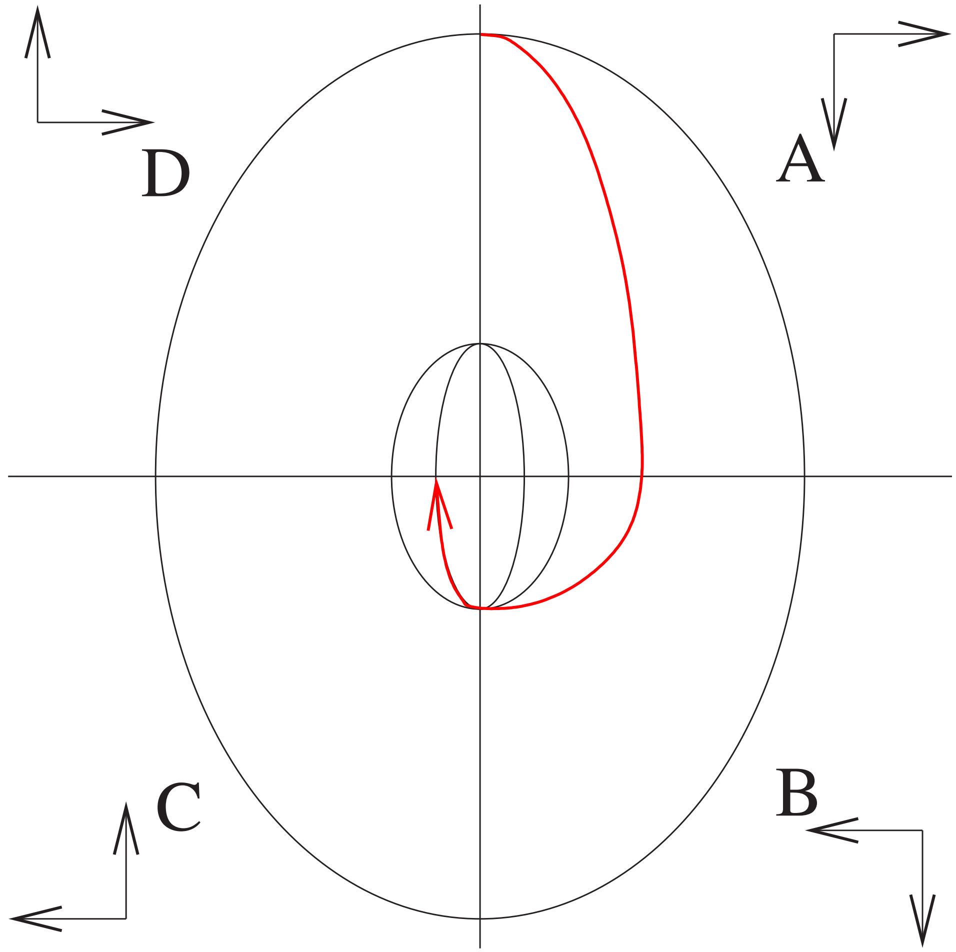
WoLF Gradient Ascent – 6



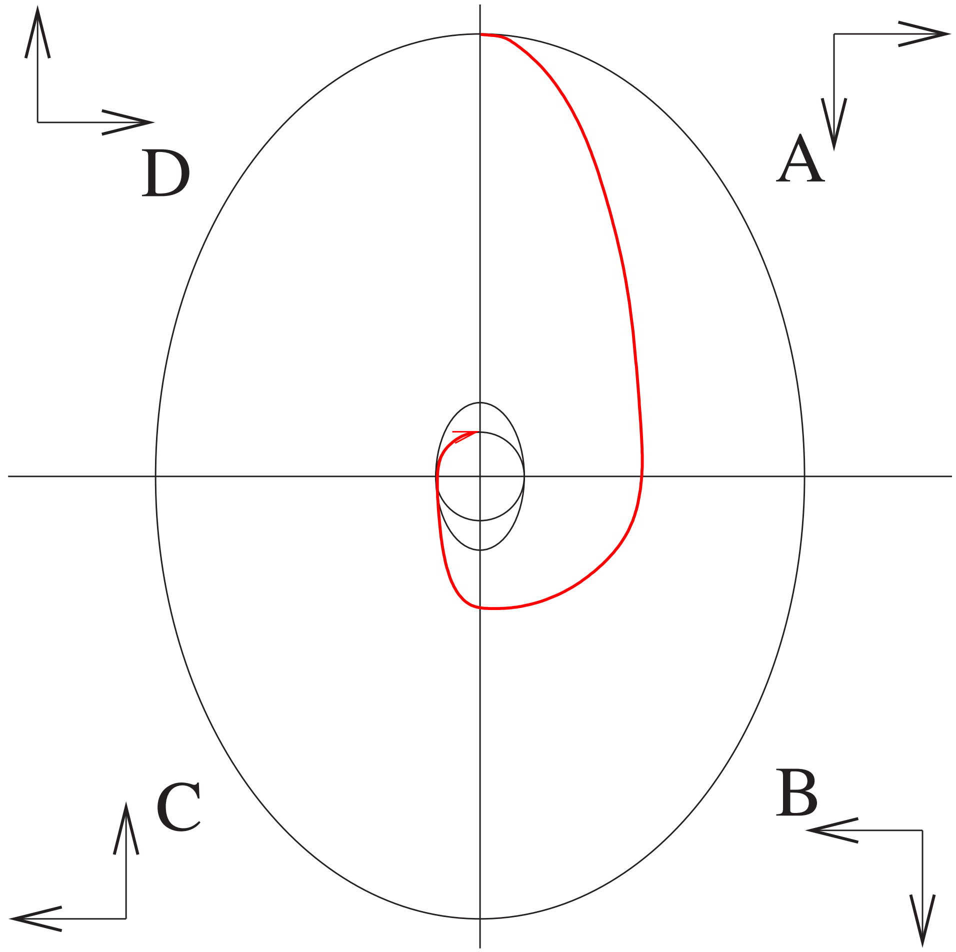
WoLF Gradient Ascent – 7



WoLF Gradient Ascent – 8

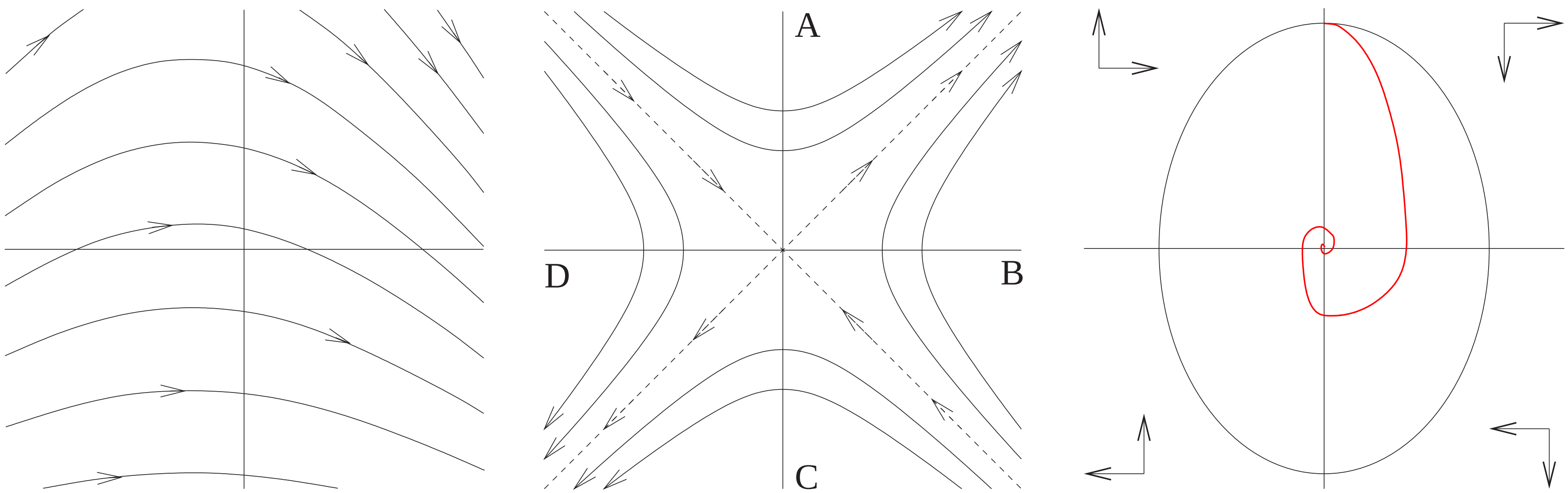


WoLF Gradient Ascent – 9



WoLF Gradient Ascent – Summary

Theorem 1. *If both players follow WoLF gradient ascent then their strategies will converge to a Nash equilibrium.*



U is not invertible
$u = 0$ or $u' = 0$

U has real eigenvalues
$uu' < 0$

U has imaginary eigenvalues
$uu' > 0$

Empirical Results

Policy Hill Climbing

- Q-Learning, but maintain a separate policy.
- Step policy towards maximizing Q-values.

$$
\begin{aligned}
Q(s,a) &\leftarrow (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a')\right) \\
\pi(s,a) &\leftarrow \pi(s,a) + \begin{cases} \delta & \text{if } a = \mathrm{argmax}_{a'} Q(s,a') \\ \frac{-\delta}{|A_i|-1} & \text{Otherwise} \end{cases}
\end{aligned}
$$

- Rational, but not Convergent.

WoLF Policy Hill-Climbing

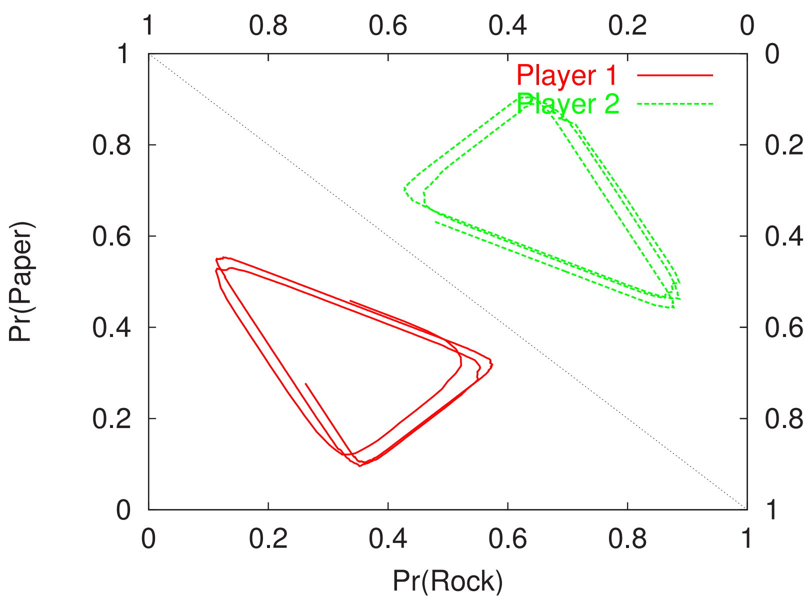
- Adjust δ based on winning and losing.
- Compare current policy to the average policy while learning.

$$\delta = \begin{cases} \delta_w & \text{if } \sum_{a'} \pi(s,a')Q(s,a') > \sum_{a'} \bar{\pi}(s,a')Q(s,a') \\ \delta_l & \text{otherwise} \end{cases} .$$

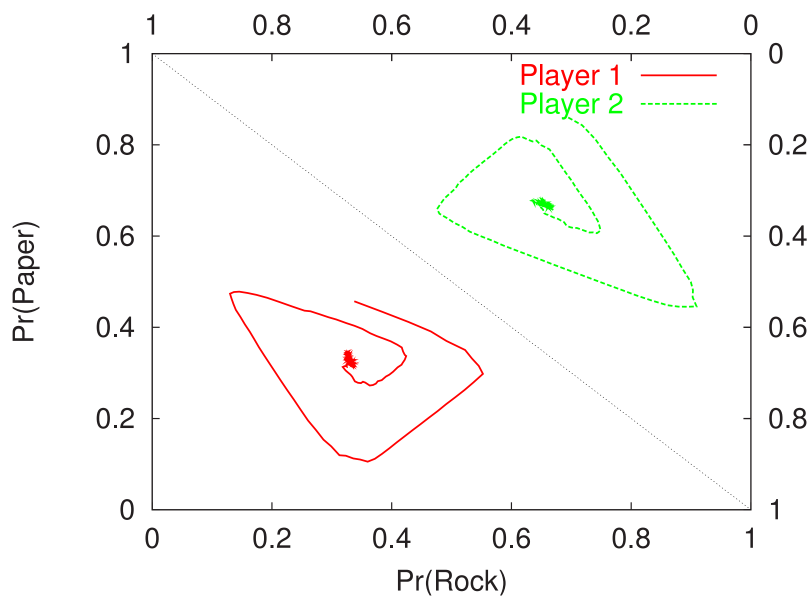
- Makes PHC converge in practice!

Results – Rock-Paper-Scissors

$$R_1 = \begin{array}{c|ccc} & R & P & S \\ \hline R & 0 & -1 & 1 \\ P & 1 & 0 & -1 \\ S & -1 & 1 & 0 \end{array} \qquad R_2 = \begin{array}{c|ccc} & R & P & S \\ \hline R & 0 & 1 & -1 \\ P & -1 & 0 & 1 \\ S & 1 & -1 & 0 \end{array}$$

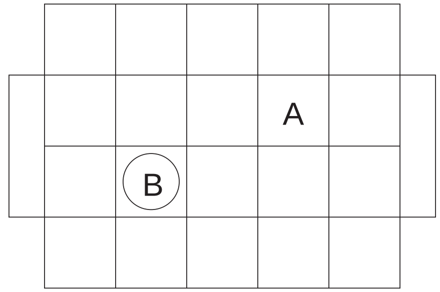


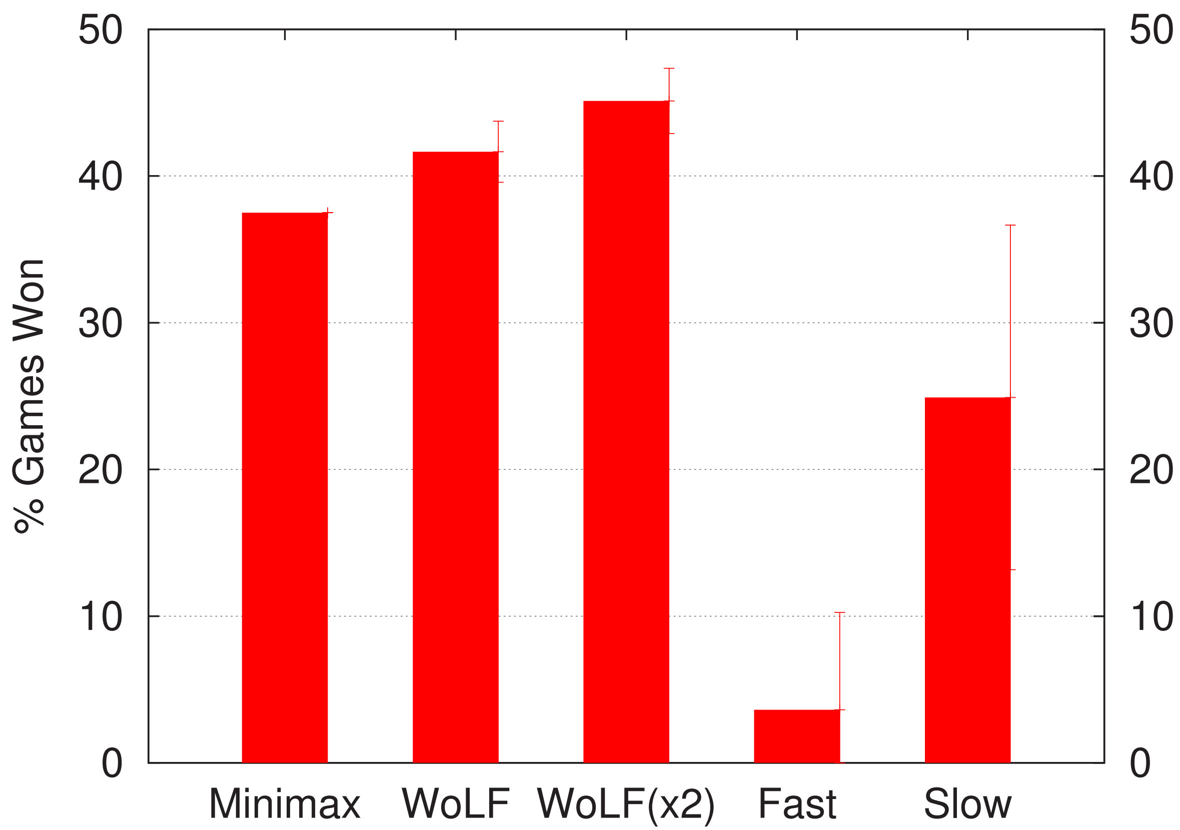
PHC



WoLF PHC

Results – Soccer





Overview

- Motivation
- Stochastic Games
- WoLF: Rational and Convergent Learning
  - Theoretical results in matrix games
  - Empirical results in "small" stochastic games
- GraWoLF: Learning with Limitations
  - Limitations and Equilibria
  - GraWoLF Algorithm
  - Empirical results of learning in Goofspiel and Keepout
- Summary and Future Work

Limitations and Equilibria

- Limitations may restrict an agent from playing the equilibrium.

- Restricted equilibria may exist. (Bowling & Veloso, 2002)
  - Guaranteed only under stringent assumptions.
  - Restricted equilibria can be learned by WoLF-PHC.

- In general, limitations do not preserve equilibria.
  - Can agents still learn?
  - How do we evaluate learning agents?

GraWoLF

- Intuition: Use parameterized policy gradient techniques.
- Intuition: Combine with WoLF.

GraWoLF

- Intuition: Use parameterized policy gradient techniques.
- Intuition: Combine with WoLF.
- Idea #1: Policy Gradient Ascent (Sutton et al., 2000)

$$\theta \leftarrow \theta + \delta \frac{\partial V^{\pi_\theta}}{\partial \theta}$$

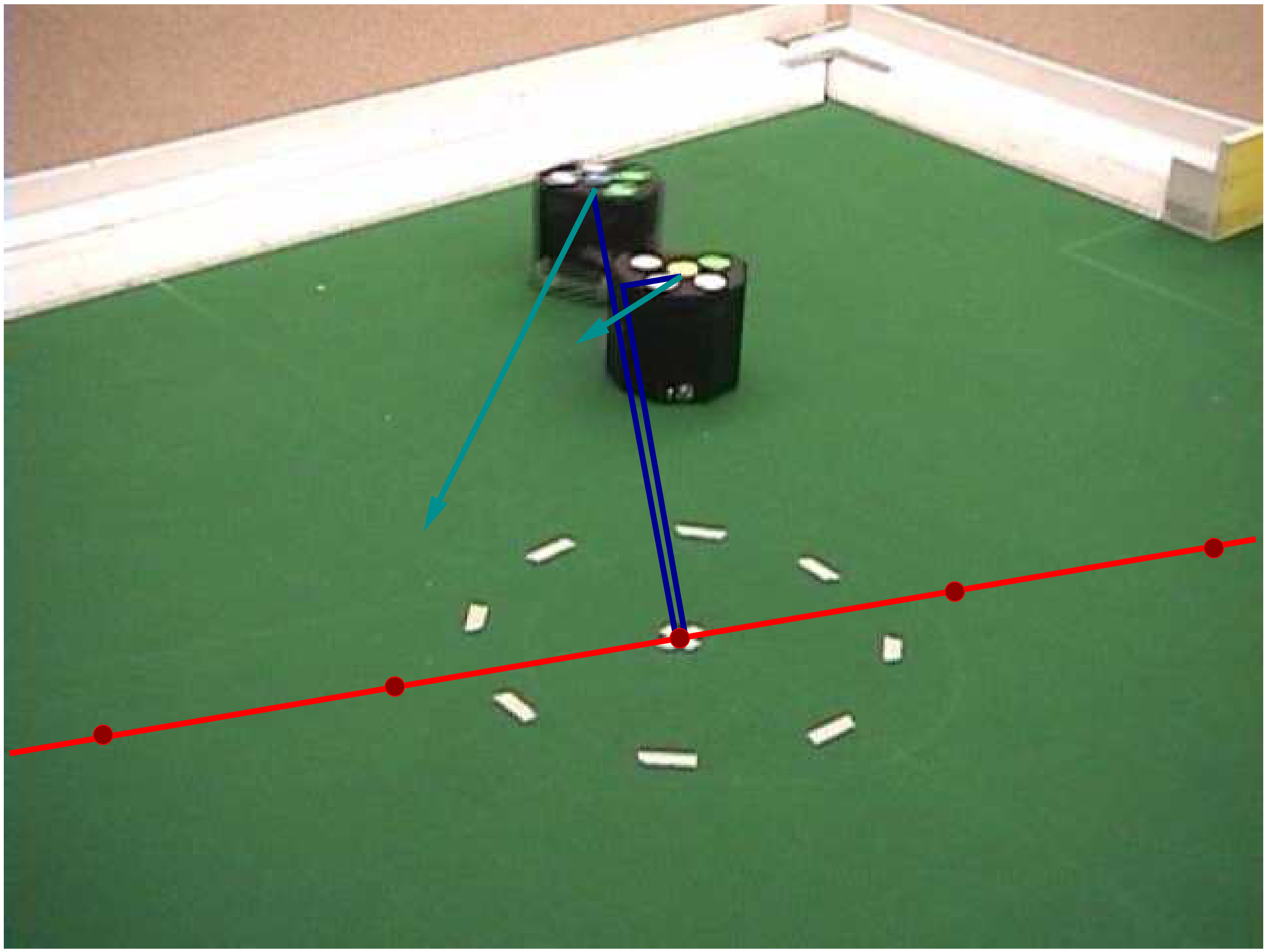
$$\pi_\theta(s,a) = \frac{e^{\phi_{sa} \cdot \theta}}{\sum_{b \in \mathcal{A}_i} e^{\phi_{sb} \cdot \theta}}$$

$$\theta \leftarrow \theta + \delta \sum_a \phi_{sa} \pi_\theta(s,a) \left(Q^\pi(s,a) - V^\pi(s)\right)$$

- Idea #2: WoLF

$$\delta = \begin{cases} \delta_w & \text{if } V^{\pi_\theta} > V^{\pi_{\bar{\theta}}} \\ \delta_l & \text{otherwise} \end{cases} \qquad \text{where } \delta_w < \delta_l$$

Applying GraWoLF – Keepout



Applying GraWoLF – Goofspiel

My Hand	1	3	4	5	6	8	11	13
Opp Hand	4	5	8	9	10	11	12	13
Deck	1	2	3	5	9	10	11	12

Applying GraWoLF – Goofspiel

My Hand	1	3	4	5	6	8	11	13
Opp Hand	4	5	8	9	10	11	12	13
Deck	1	2	3	5	9	10	11	12
Quartiles	$\times$		$\times$		$\times$	$\times$		$\times$

Card 11
Action 3

$$\left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \begin{array}{c} \langle 1, 4, 6, 8, 13 \rangle, \\ \langle 4, 8, 10, 11, 13 \rangle, \\ \langle 1, 3, 9, 10, 12 \rangle, \\ 11, 3 \end{array}$$

Applying GraWoLF – Goofspiel

My Hand	1	3	4	5	6	8	11	13
Opp Hand	4	5	8	9	10	11	12	13
Deck	1	2	3	5	9	10	11	12
Quartiles	×		×		×	×		×
Card	11							
Action	3							

$$\left.\right\} \begin{array}{c} \langle 1, 4, 6, 8, 13 \rangle, \\ \langle 4, 8, 10, 11, 13 \rangle, \\ \langle 1, 3, 9, 10, 12 \rangle, \\ 11, 3 \end{array}$$

$$\Downarrow \text{ (Tile Coding)}$$

$$\phi_{sa} \in \{0, 1\}^{10^6}$$

Multiagent Learning Evaluation

- This is an important part of the ongoing research.

Multiagent Learning Evaluation

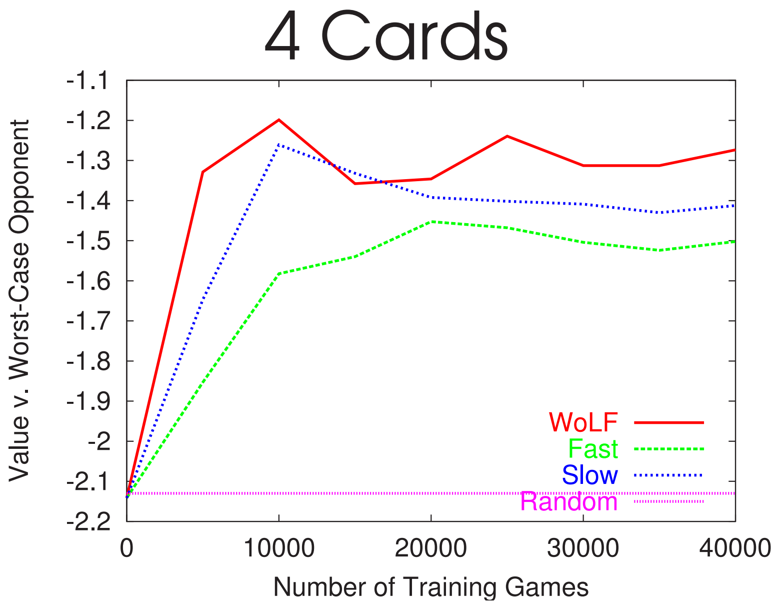
- This is an important part of the ongoing research.
- No optimal policy. No equilibrium for convergence.

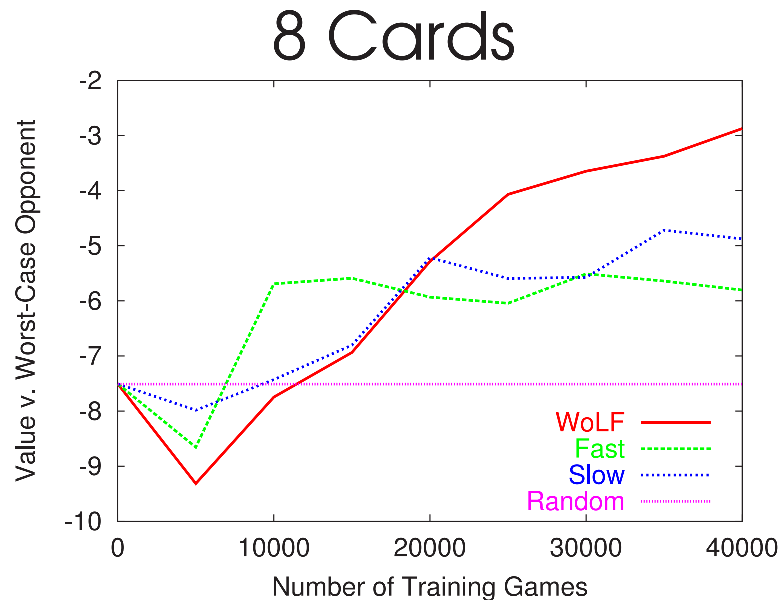
Multiagent Learning Evaluation

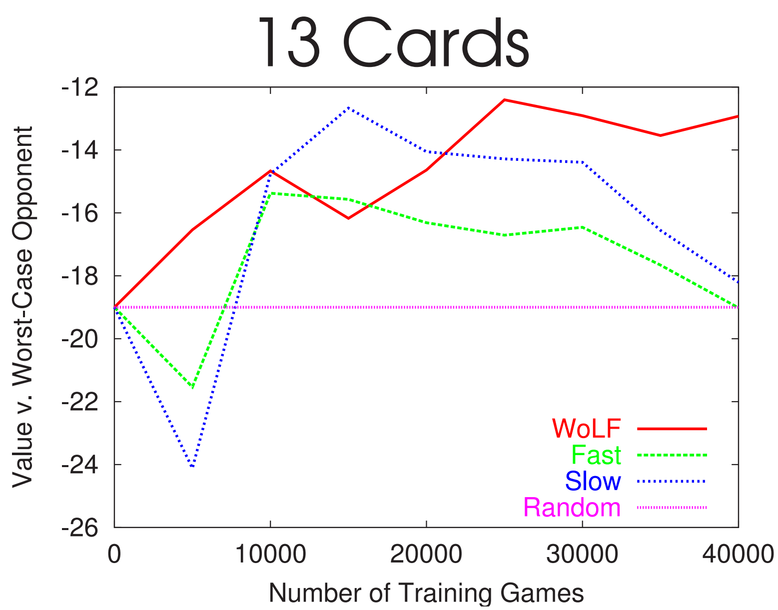
- This is an important part of the ongoing research.
- No optimal policy. No equilibrium for convergence.
- Measure the policy's worst-case value.
  - For a given policy, train a "challenger".
  - Measures distance to the equilibrium.
  - Measures robustness of the learned policy.

Goofspiel

Goofspiel

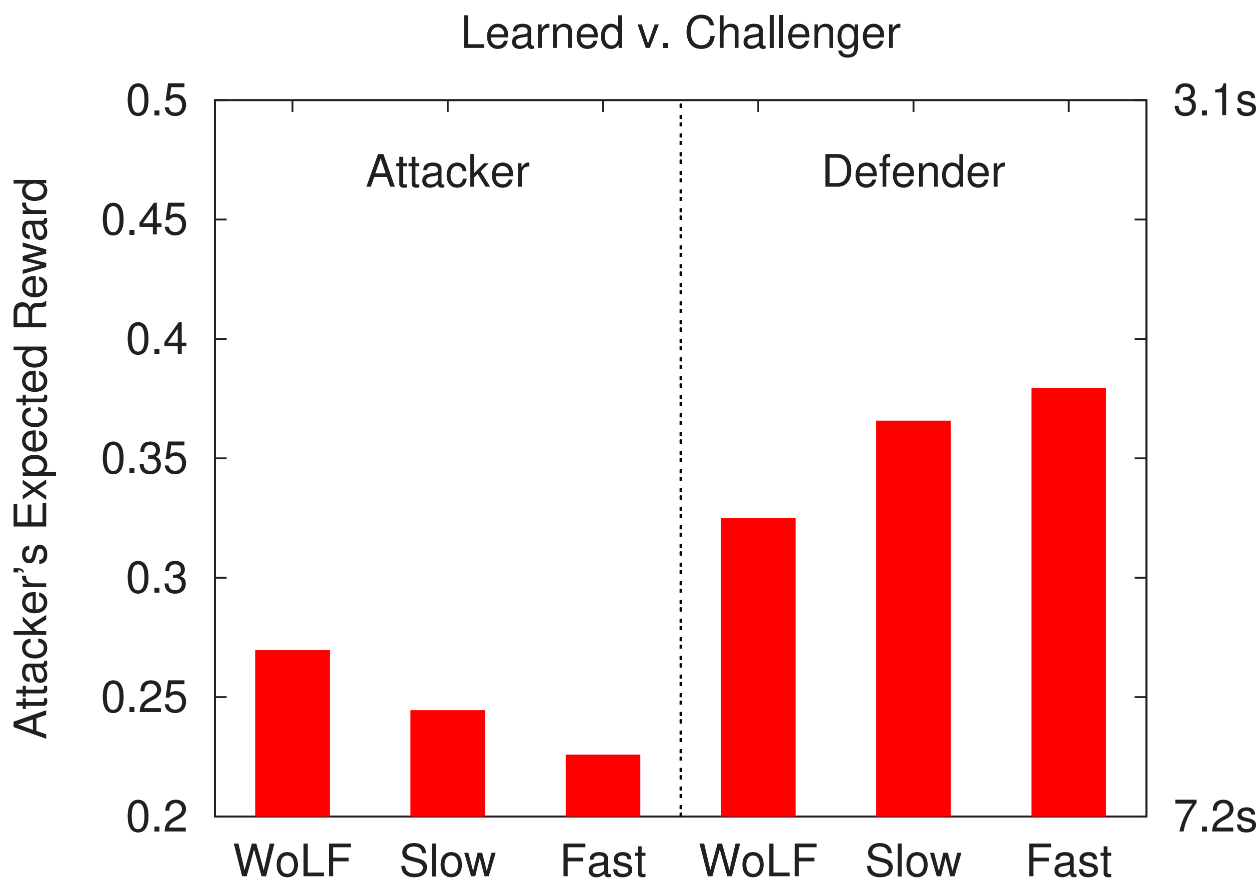

4 Cards
Value v. Worst-Case Opponent
Number of Training Games
-1.1
-1.2
-1.3
-1.4
-1.5
-1.6
-1.7
-1.8
-1.9
-2
-2.1
-2.2
0
10000
20000
30000
40000
WoLF
Fast
Slow
Random


8 Cards
Value v. Worst-Case Opponent
Number of Training Games
-2
-3
-4
-5
-6
-7
-8
-9
-10
0
10000
20000
30000
40000
WoLF
Fast
Slow
Random


13 Cards
Value v. Worst-Case Opponent
Number of Training Games
-12
-14
-16
-18
-20
-22
-24
-26
0
10000
20000
30000
40000
WoLF
Fast
Slow
Random

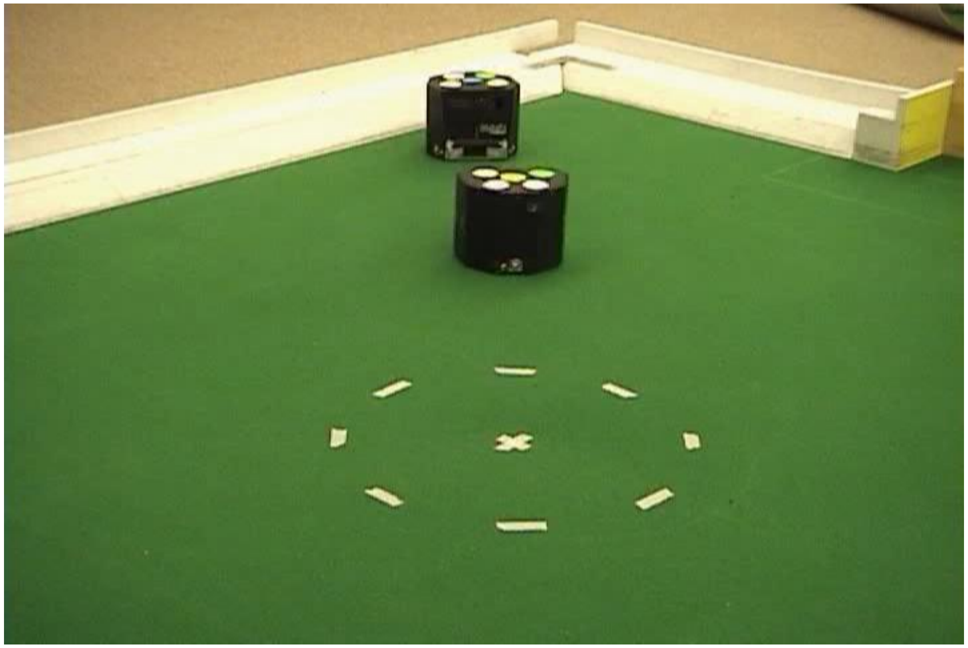
Keepout – Simulation

- 4000 Trials of Simultaneous Learning in Simulation.



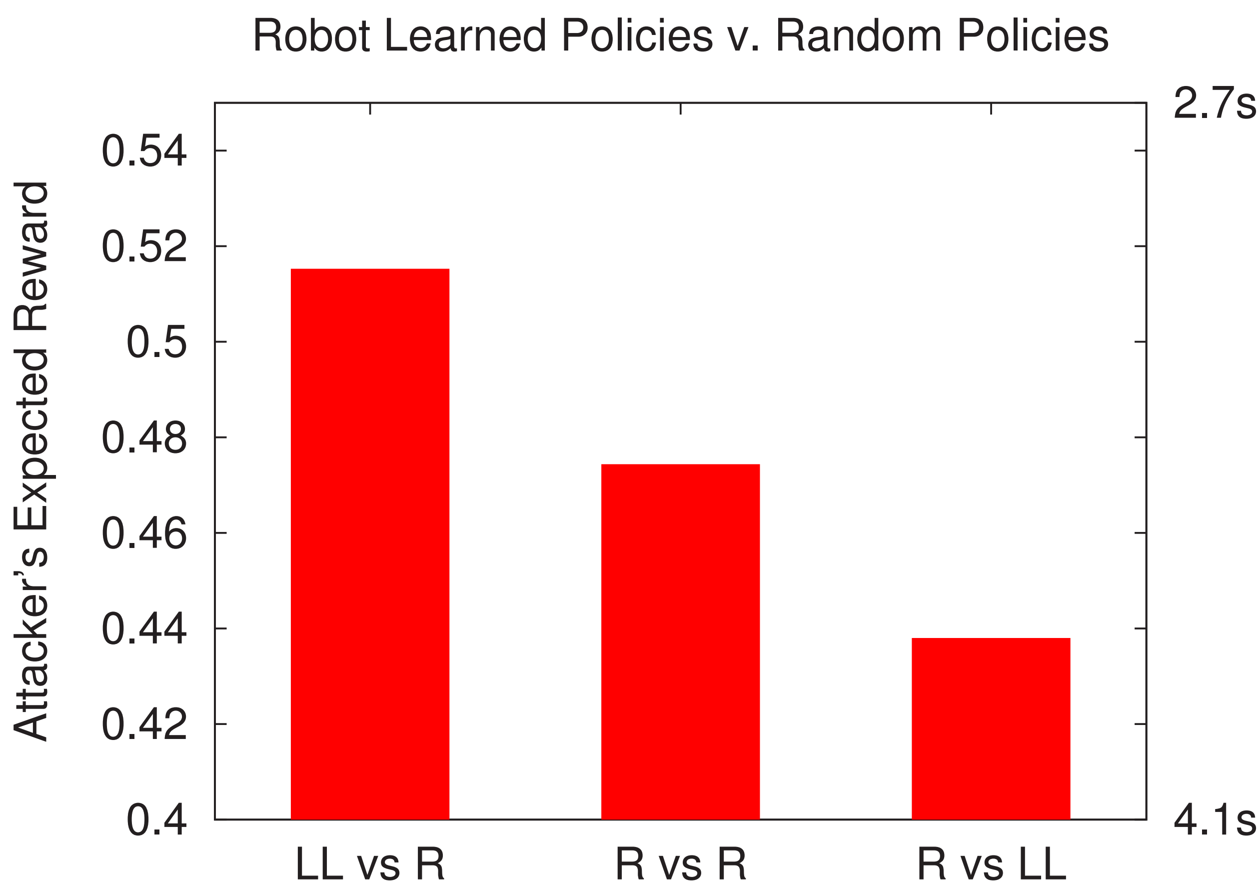
Keepout – Robots

- 2000 Trials of Simultaneous Learning in Simulation.
- 2000 Trials of Simultaneous Learning on Robots.



Keepout – Robots – Versus Random

- 2000 Trials of Simultaneous Learning in Simulation.
- 2000 Trials of Simultaneous Learning on Robots.
- 500 Trials of Evaluation on Robots



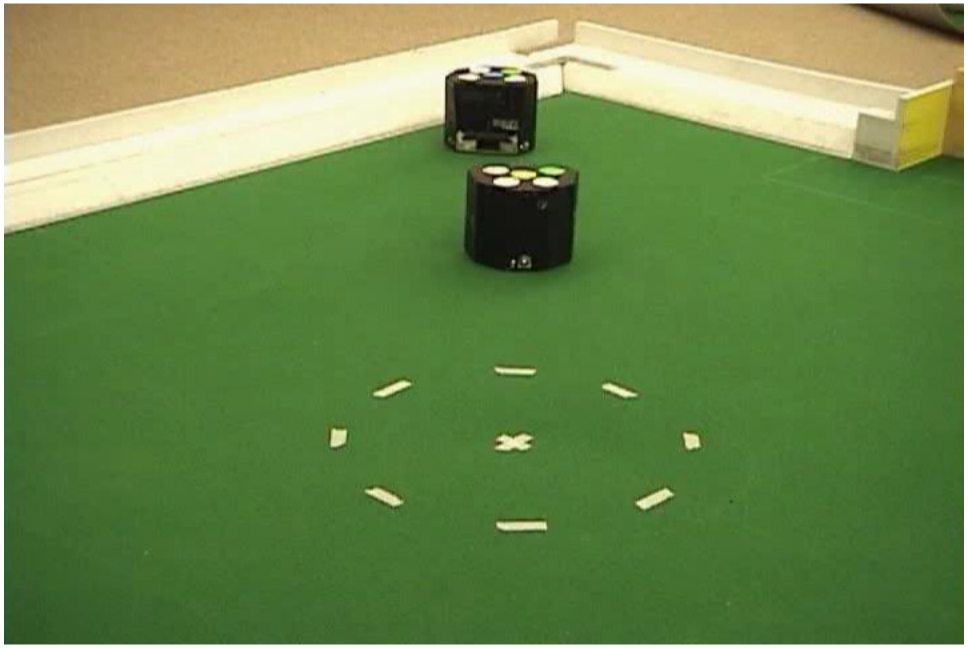
Summary

- Multiagent learning is important and challenging.
- WoLF makes rational learners converge.
  - Theoretical results for a class of matrix games.
  - Empirical results on "small" stochastic games.
  - WoLF can also learn restricted equilibria.
- GraWoLF is a scalable multiagent learning algorithm.
  - Combines approximation and WoLF.
  - Empirical results in Goofspiel and Keepout.

Future Work

- Further Theoretical Analysis of WoLF
  - WoLF dynamics outside matrix games.
  - How does WoLF relate to regret-minimizing algorithms.
- Asymmetric Learning
  - Can we systematically exploit "weaker" algorithms?
  - Can we guarantee an algorithm cannot be exploited?
  - Human–Agent Interaction.
- Multiagent Learning Evaluation
  - Learn general policies for a range of opponents, or
  - Learn policies specific to a particular opponent.
  - Reusing data between different opponents.

Questions



Three-Player Matching Pennies

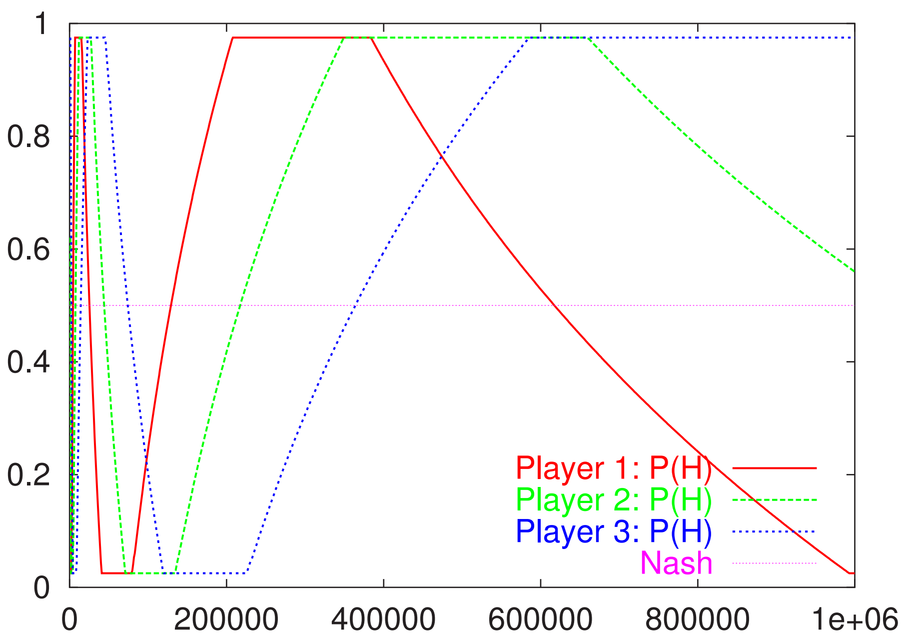
- Three players. Each simultaneously picks an action: *Heads*, or *Tails*.
- The rules:

Player One	wins by matching	Player Two,
Player Two	wins by matching	Player Three,
Player Three	wins by *not* matching	Player One.

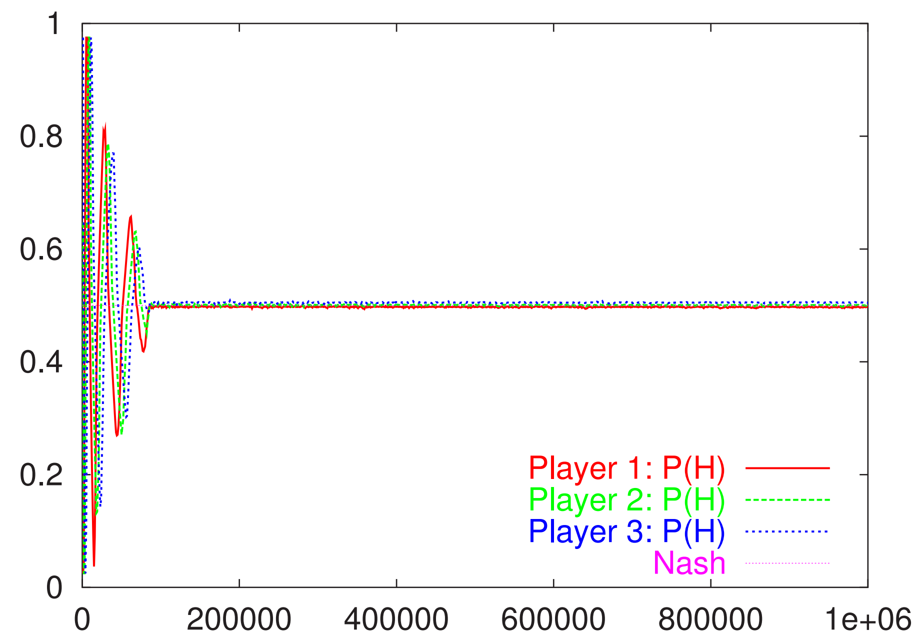
Results – Three Player Matching Pennies

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

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



$$\frac{\delta_l}{\delta_w} = 1$$

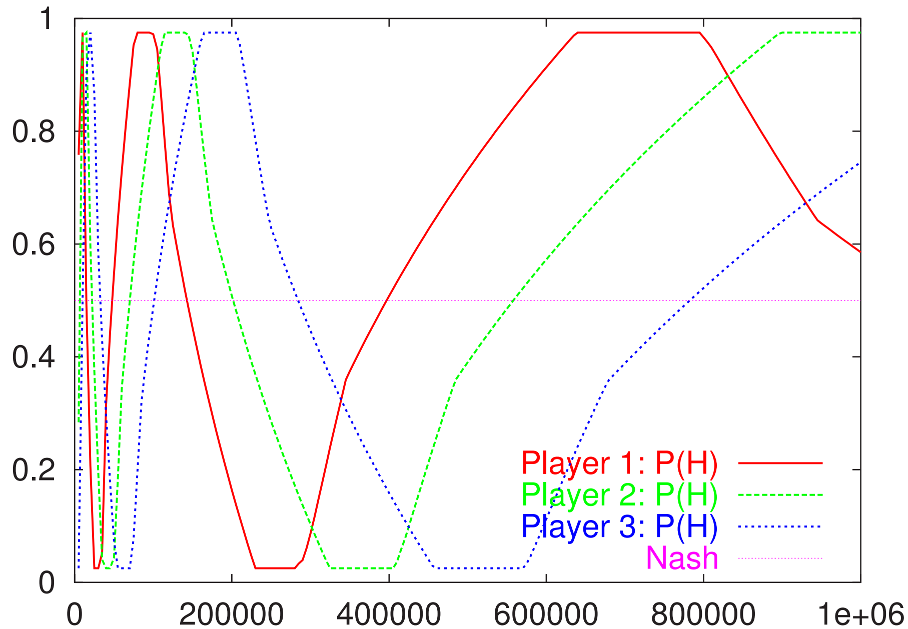


$$\frac{\delta_l}{\delta_w} = 3$$

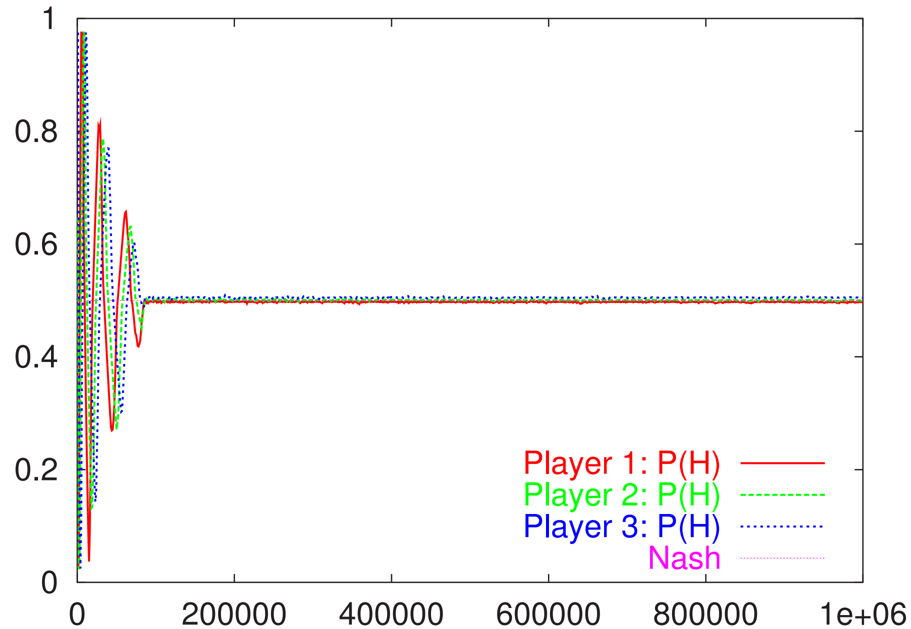
Results – Three Player Matching Pennies – 2

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

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



$$\frac{\delta_l}{\delta_w} = 2$$



$$\frac{\delta_l}{\delta_w} = 3$$

Limitations

- Anything that prevents an agent from acting optimally.

Physical Limitations	Rational Limitations
Broken Actuators	Reward Shaping
Poor Control	Abstraction/Subproblems
Hardwired Behavior	Parameterized Policy
State Aliasing	Exploration
Poor Communication	Bounded Memory
Latency	Function Approximation

- Limitations restrict behavior.

Limitations Restrict Behavior

- Restricted Policy Space — $\overline{\Pi}_i \subseteq \Pi_i$

  Any subset of stochastic policies.

Limitations Restrict Behavior

- Restricted Policy Space — $\overline{\Pi}_i \subseteq \Pi_i$

  Any subset of stochastic policies.

- Restricted Best-Response — $\overline{\mathrm{BR}}_i(\pi_{-i})$

  The set of all policies from $\overline{\Pi}_i$ that are optimal given the policies of the other players.

- Restricted Equilibrium — $\pi_{i=1\ldots n}$

$$\pi_i \in \overline{\mathrm{BR}}_i(\pi_{-i})$$

  A strategy for each player, where no player *can* and *wants* to deviate given the other players continue to play the equilibrium.

Do Restricted Equilibria Exist?

Do Restricted Equilibria Exist?

- No.

  Rock-Paper-Scissors with only deterministic policies.

Do Restricted Equilibria Exist?

- No.

  Rock-Paper-Scissors with only deterministic policies.

- Yes.

  If π^* is a Nash equilibrium and $\forall i \, \pi_i^* \in \overline{\Pi}_i$ then π^* is a restricted equilibrium.

Do Restricted Equilibria Exist?

- No.

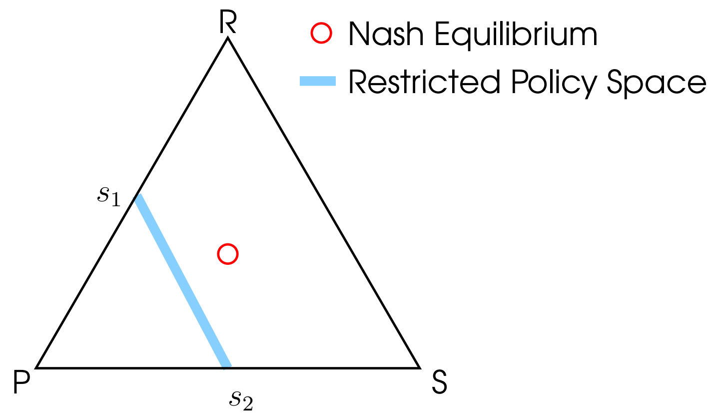
  Rock-Paper-Scissors with only deterministic policies.

- Yes.

  If π^* is a Nash equilibrium and $\forall i \, \pi_i^* \in \overline{\Pi}_i$ then π^* is a restricted equilibrium.

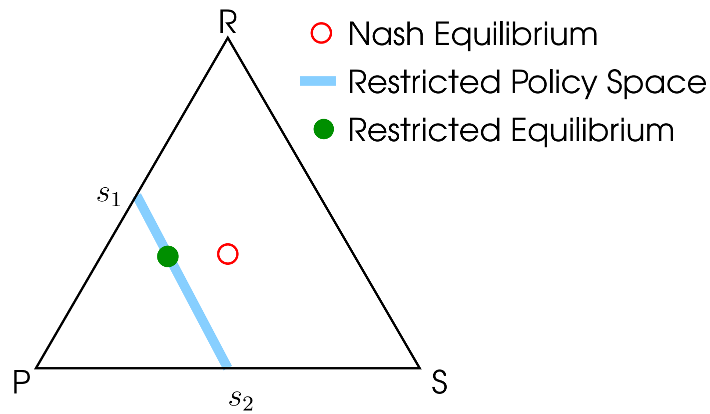
- Not everything is so trivial.

Do Restricted Equilibria Exist? – 2



	Explicit Game
Payoffs	$\begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}$
Equilibrium	$\left\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right\rangle, \left\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right\rangle$

Do Restricted Equilibria Exist? – 3



	Explicit Game	Implicit Game
Payoffs	$\begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} -\frac{1}{2} & 0 \\ \frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} \end{pmatrix}$
Equilibrium	$\left\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right\rangle, \left\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right\rangle$	$\left\langle 0, \frac{1}{3}, \frac{2}{3} \right\rangle, \left\langle \frac{2}{3}, \frac{1}{3} \right\rangle$
Restricted Equilibrium	$\left\langle 0, \frac{1}{3}, \frac{2}{3} \right\rangle, \left\langle \frac{1}{3}, \frac{1}{2}, \frac{1}{6} \right\rangle$	

Do Restricted Equilibria Exist? – 4

- In matrix games, if $\overline{\Pi}_i$ is convex, then there exists a restricted equilibrium..

  Proof. Uses Rosen's theorem for concave games.

Do Restricted Equilibria Exist? – 4

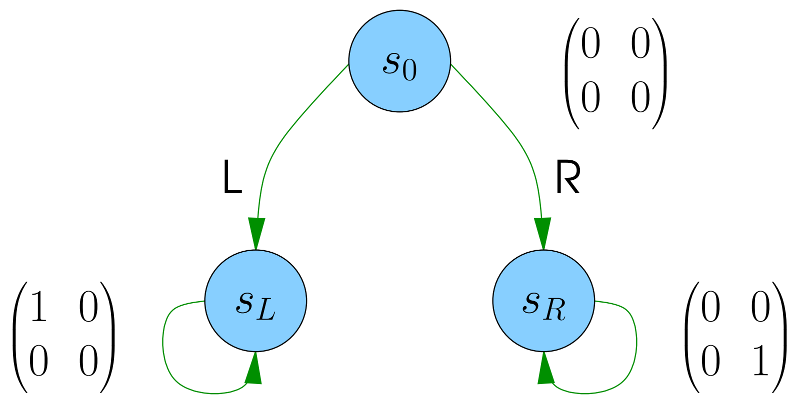
- In matrix games, if $\overline{\Pi}_i$ is convex, then there exists a restricted equilibrium..

  Proof. Uses Rosen's theorem for concave games.

- This is not generally true for stochastic games.

Do Restricted Equilibria Exist? – 5

- Two-player, zero-sum stochastic game[1]



- Players restricted to policies that play the same distribution over actions in all states.

- No restricted equilibria!

[1] This counterexample is brought to you by Martin Zinkevich.

Do Restricted Equilibria Exist? – 6

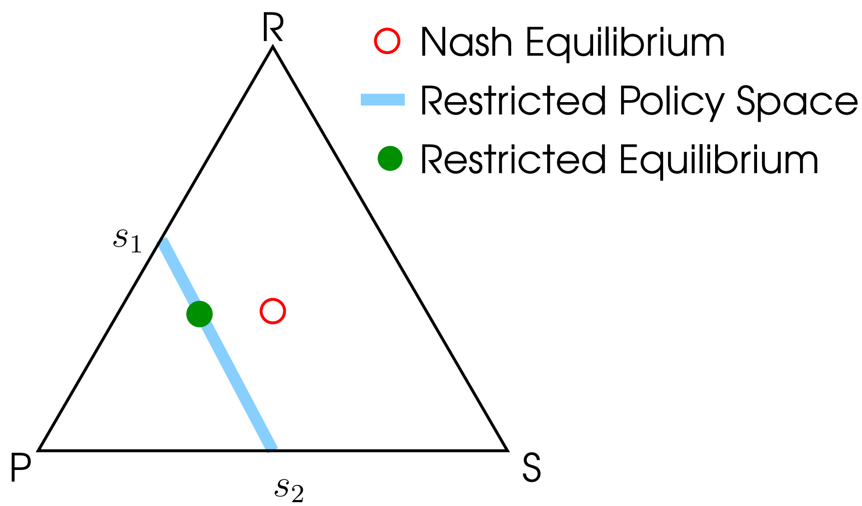
- In matrix games, if $\overline{\Pi}_i$ is convex, then ...
- If $\overline{\Pi}_i$ is statewise convex, then ...
- In no-control stochastic games, if convex $\overline{\Pi}_i$, then ...
- In single-controller stochastic games, if $\overline{\Pi}_1$ is statewise convex, and $\overline{\Pi}_{i\neq 1}$ is convex, then ...
- In team games ...

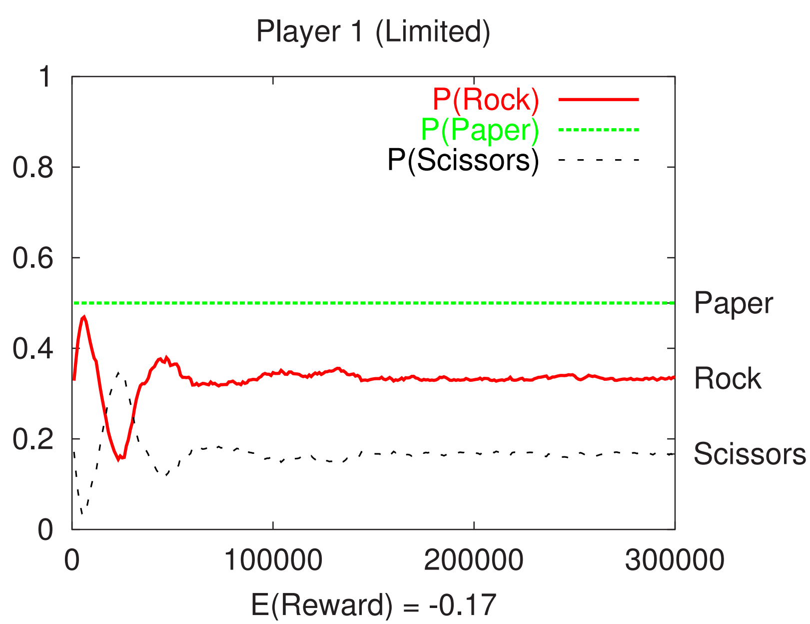
...there exists a restricted equilibrium.

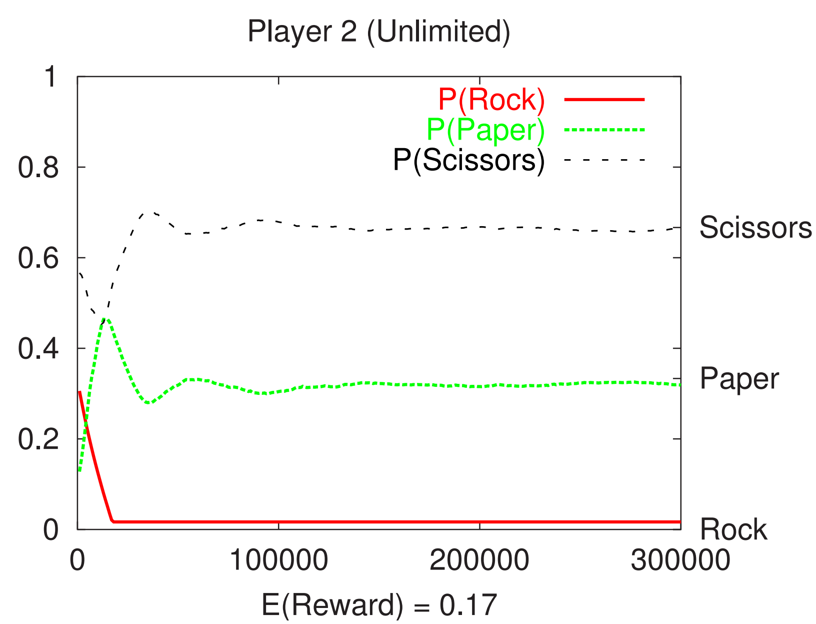
Proofs. Uses Kakutani's fixed point theorem after showing

$$\forall \pi_{-i} \quad \overline{\mathrm{BR}}_i(\pi_{-i}) \text{ is convex.}$$

Limitations and Learning

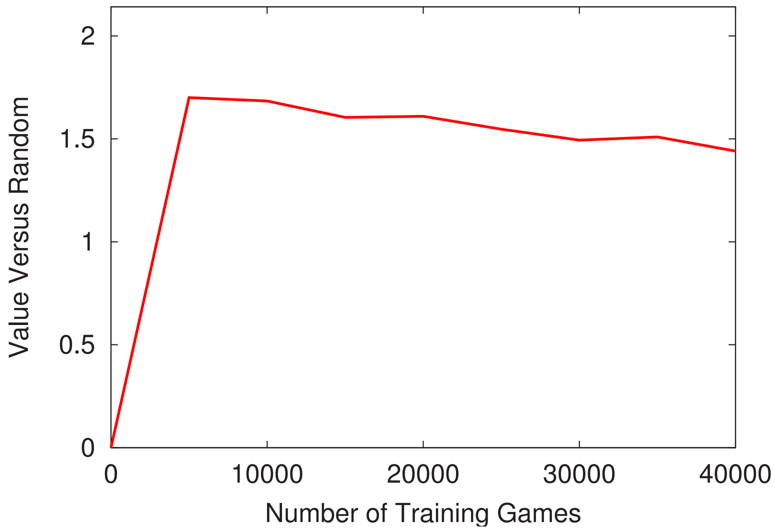




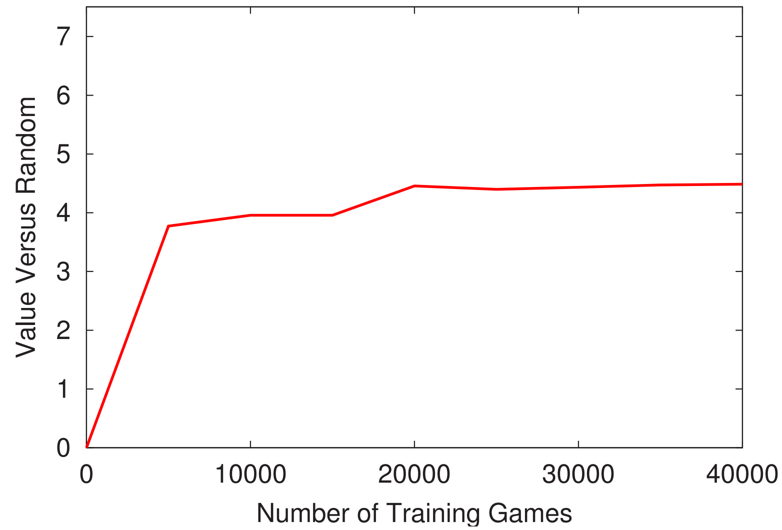


Goofspiel – Versus Random

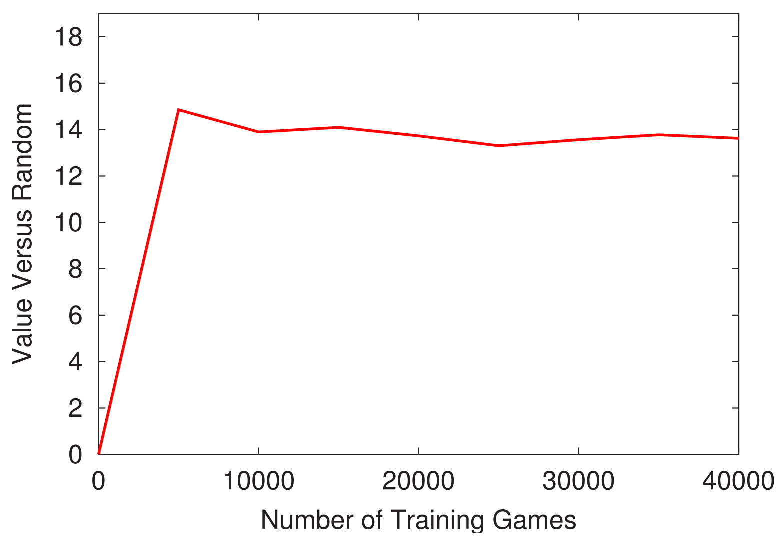
4 Cards



8 Cards



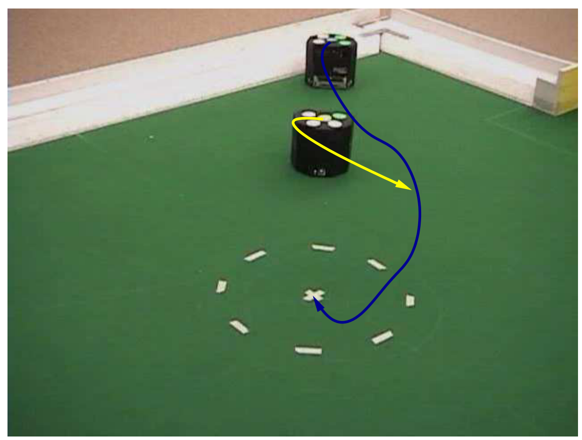
13 Cards



Keepout – Simulation – Versus Random

The Key . . .

R	Random policy
LL	Policy learned against learning opponent
LR	Policy learned against random opponent



Keepout – Simulation – Versus Random