Learning, Equilibria, Limitations, and Robots*

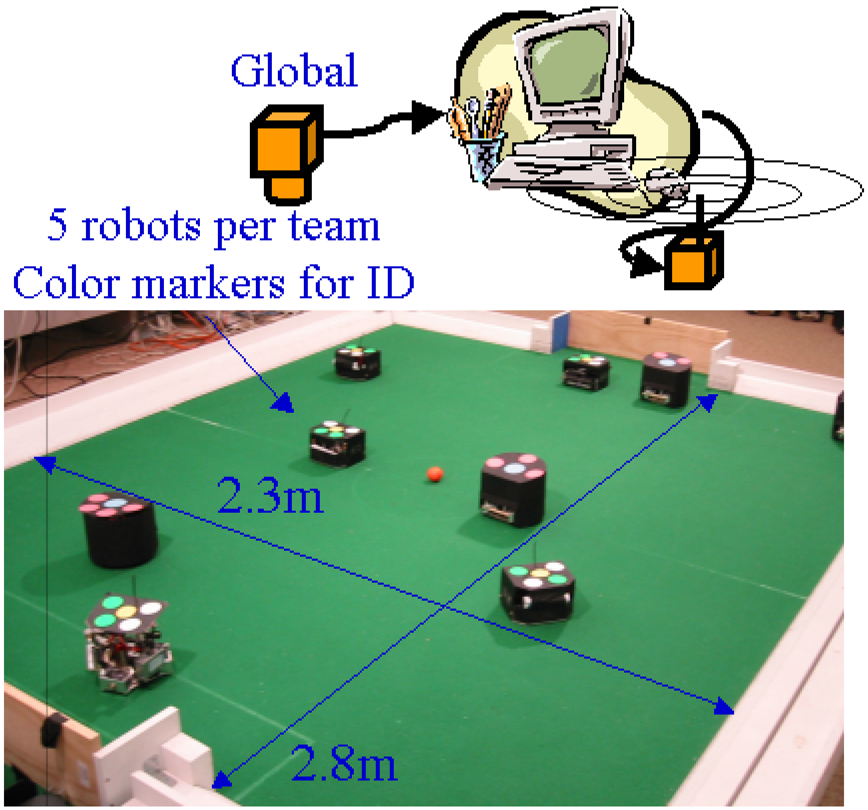
Michael Bowling
Computer Science Department
Carnegie Mellon University

*Joint work with Manuela Veloso

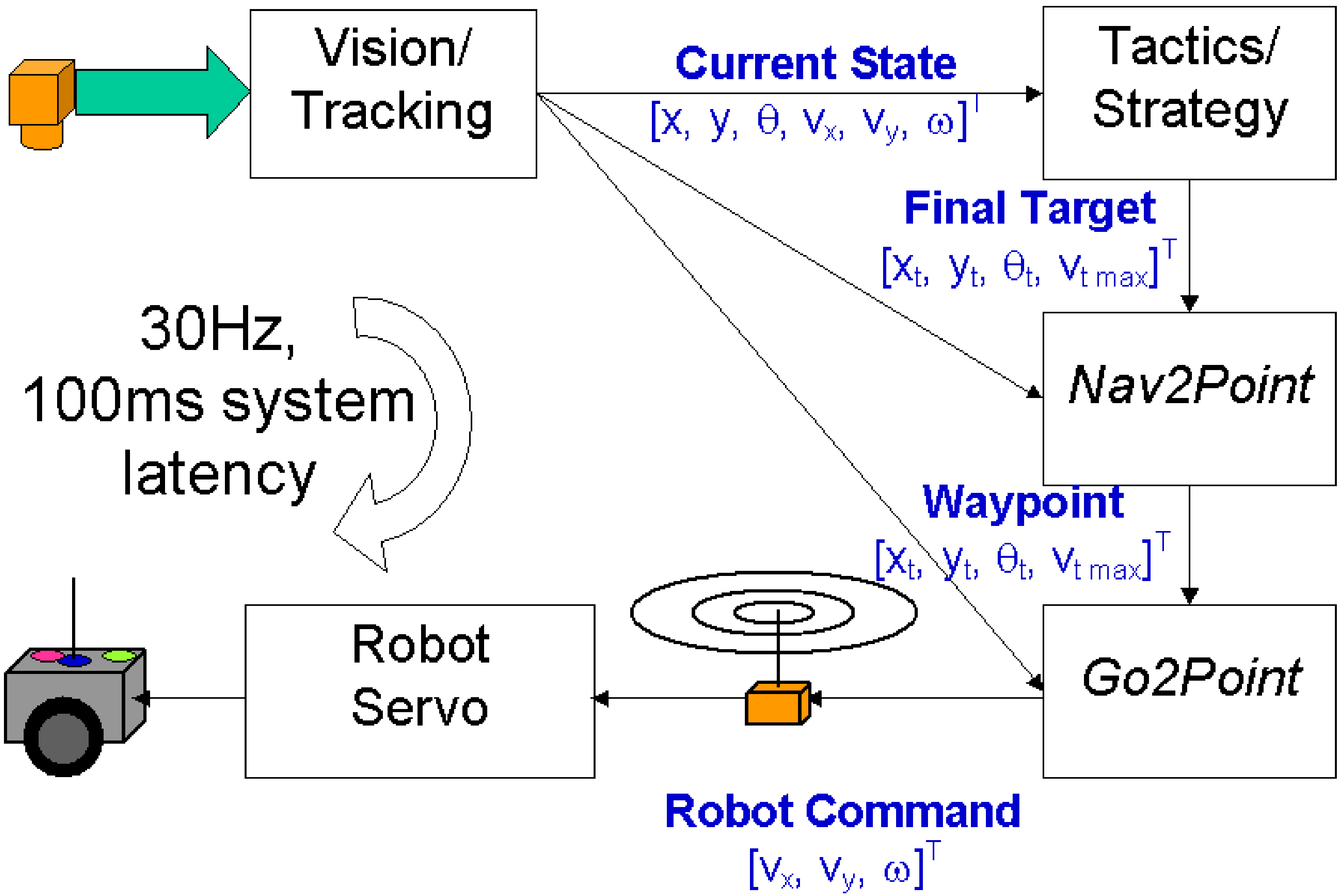
Talk Outline

- Robots
  - A two robot, adversarial, concurrent learning problem.
  - The challenges for multiagent learning.
- Limitations and Equilibria
- Limitations and Learning

The Domain – CMDragons – 1

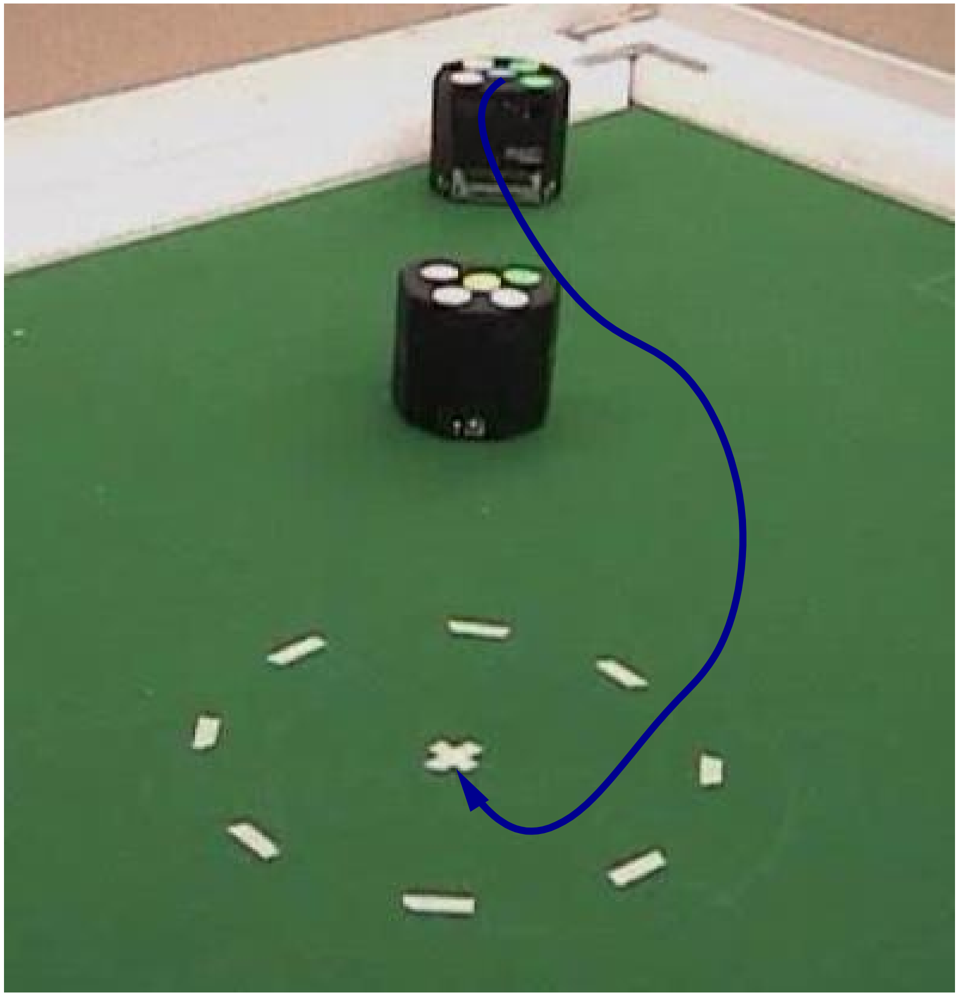


The Domain – CMDragons – 2

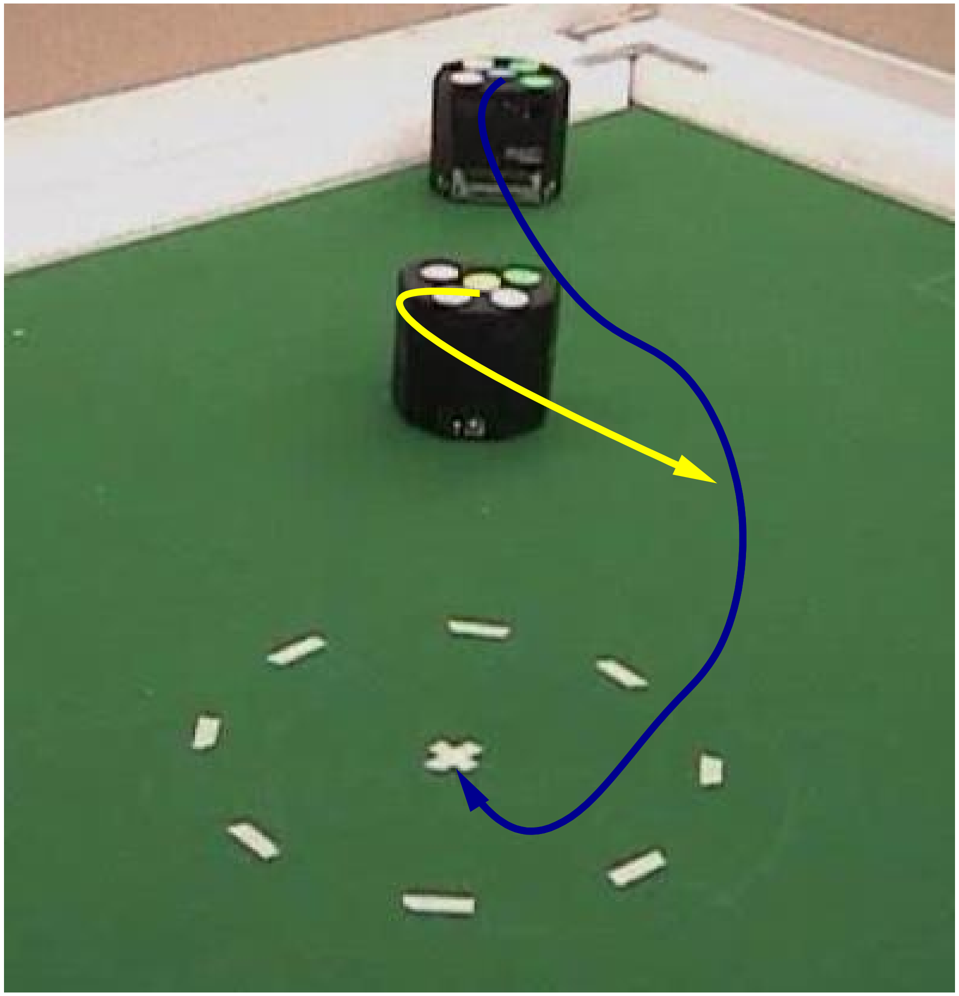


The Task – Breakthrough

The Task – Breakthrough



The Task = Breakthrough



The Challenges

The Challenges

- Challenge #1: Continuous State and Action Spaces
  - Value function approximation, parameterized policies, state and temporal abstractions.
  - Limits agent behavior, sacrificing optimality.

The Challenges

- Challenge #1: Continuous State and Action Spaces
  - Value function approximation, parameterized policies, state and temporal abstractions.
  - Limits agent behavior, sacrificing optimality.
- Challenge #2: Fixed Behavioral Components
  - Don't learn motion control or obstacle avoidance.
  - Limits agent behavior, sacrificing optimality.

The Challenges

- Challenge #1: Continuous State and Action Spaces
  - Value function approximation, parameterized policies, state and temporal abstractions.
  - Limits agent behavior, sacrificing optimality.

- Challenge #2: Fixed Behavioral Components
  - Don't learn motion control or obstacle avoidance.
  - Limits agent behavior, sacrificing optimality.

- Challenge #3: Latency
  - Can predict our own state through latency, not others.
  - Asymmetric partial observability.
  - Limits agent behavior, sacrificing optimality.

The Challenges – 1

- Challenge #1: Continuous State and Action Spaces
- Challenge #2: Fixed Behavioral Components
- Challenge #3: Latency

All of these challenges involve agent **limitations.**
. . . their *own* and *other´s*.

Talk Outline

- Robots
  - A two robot, adversarial, concurrent learning problem.
  - The challenges for multiagent learning.
- Limitations and Equilibria
- Limitations and Learning

Limitations Restrict Behavior

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

  Any subset of stochastic policies, $\pi : \mathcal{S} \rightarrow PD(\mathcal{A}_i)$.

- 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.

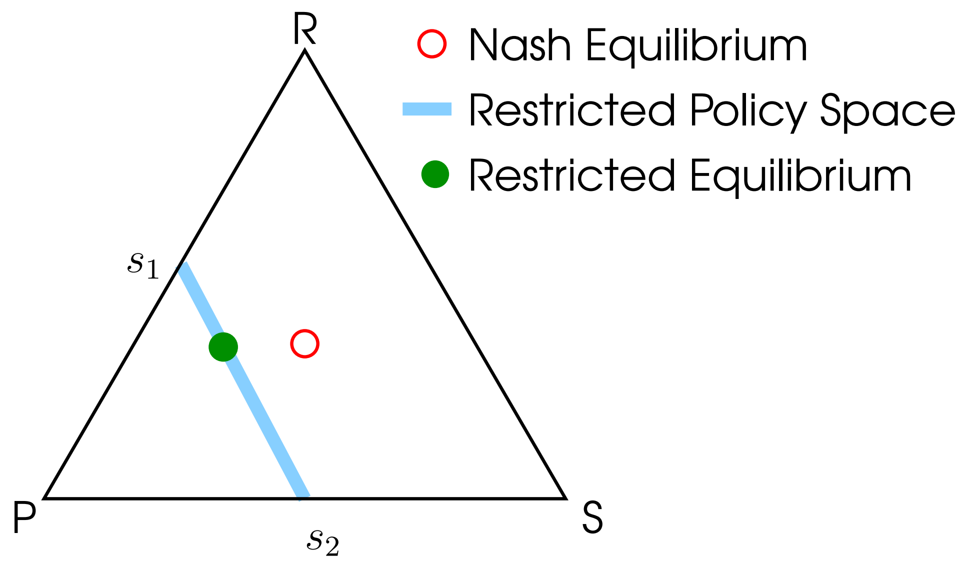
- Restricted Equilibrium — $\pi_{i=1\ldots n}$

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

  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?

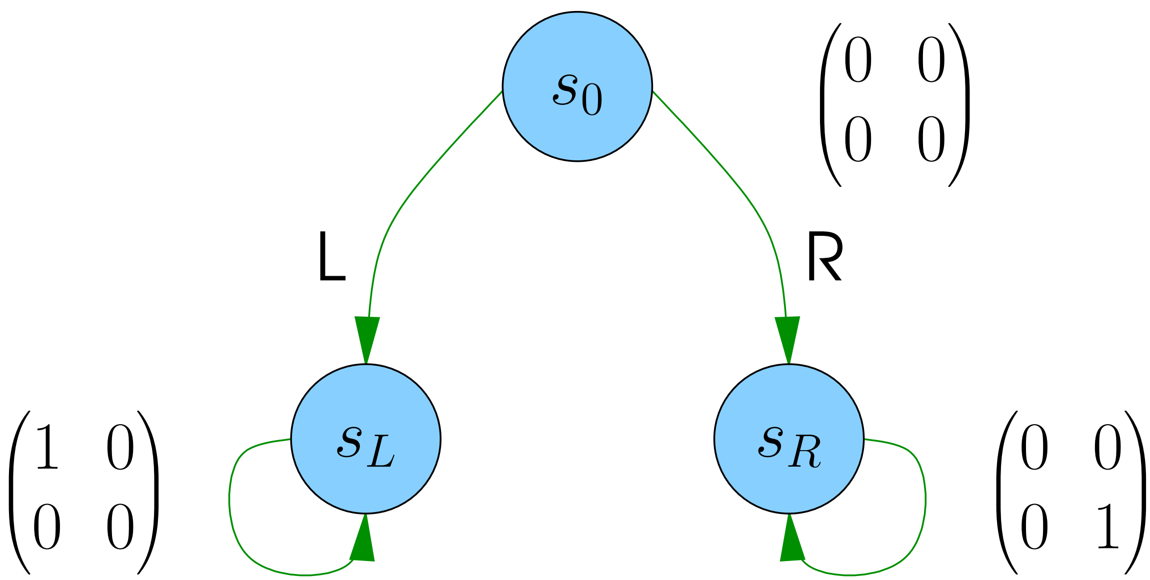
Do Restricted Equilibria Exist? – 1



	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? – 2

- Two-player, zero-sum stochastic game (Marty's Game 2).[1]



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

This game has no restricted equilibria!

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

Do Restricted Equilibria Exist? – 3

- 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 ...

Do Restricted Equilibria Exist? – 3

- 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

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

The Challenges – 2

- Challenge #1: Continuous State and Action Spaces
- Challenge #2: Fixed Behavioral Components
- Challenge #3: Latency

None of these are nice enough to guarantee the existence of equilibria.

Talk Outline

- Robots
  - A two robot, adversarial, concurrent learning problem.
  - The challenges for multiagent learning.
- Limitations and Equilibria
- Limitations and Learning

Three Ideas = One Algorithm

- Idea #1: Policy Gradient Ascent
- Idea #2: WoLF Variable Learning Rate

GrāWoLF— Gradient-based WoLF

- Idea #3: Tile Coding

Idea #1

- Policy Gradient Ascent (Sutton et al., 2000)
  - Policy improvement with parameterized policies.
  - Takes steps in direction of the gradient of the value.

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

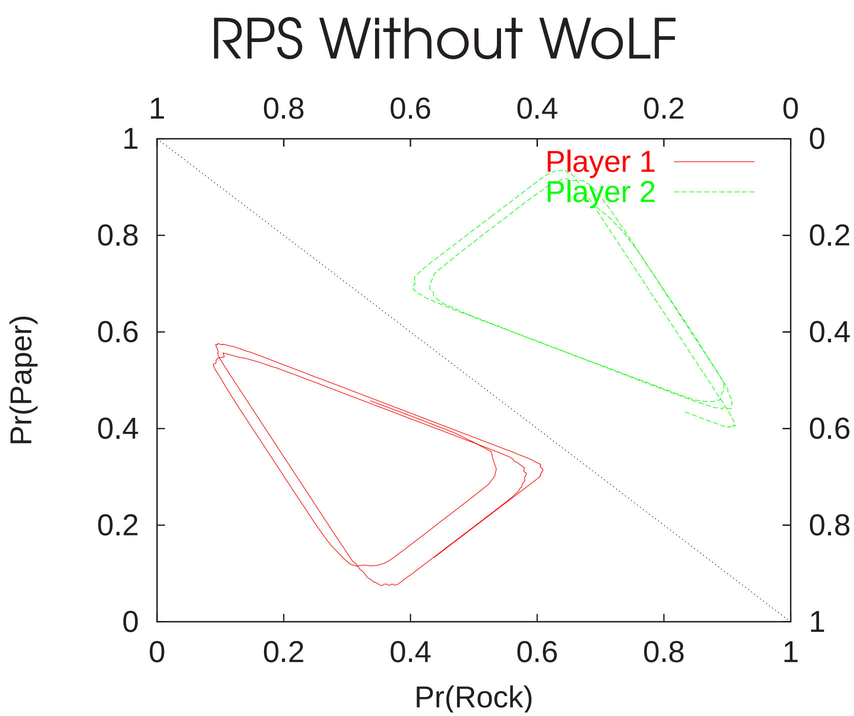
$$\theta_{k+1} = \theta_k + \alpha_k \sum_a \phi_{sa}\pi(s,a) f_k(s,a)$$

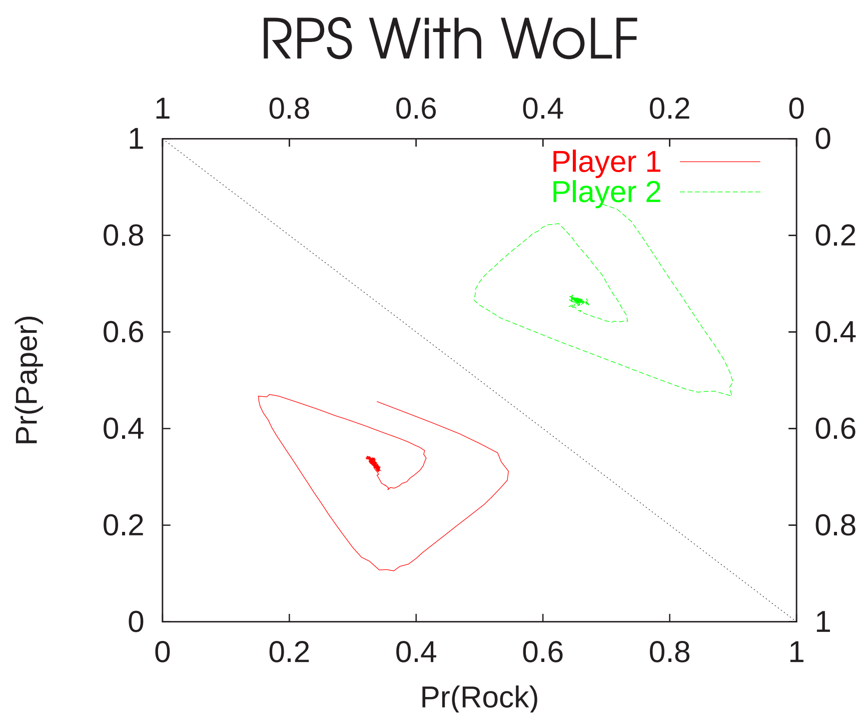
  - f_k is an approximation of the advantage function.

$$\begin{aligned} f_k(s,a) &\approx Q(s,a) - V^{\pi}(s) \\ &\approx Q(s,a) - \sum_b \pi(s,b) Q(s,b) \end{aligned}$$

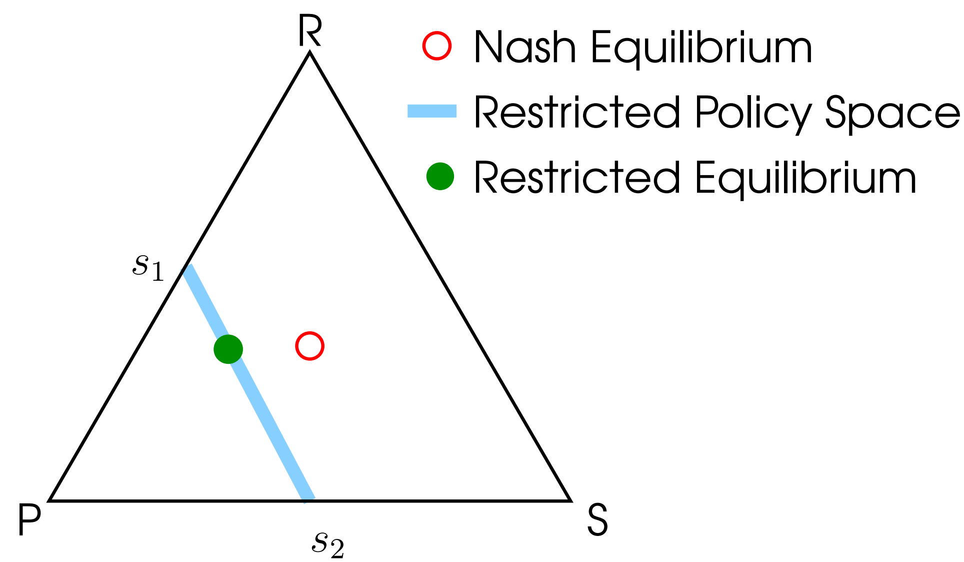
Idea #2

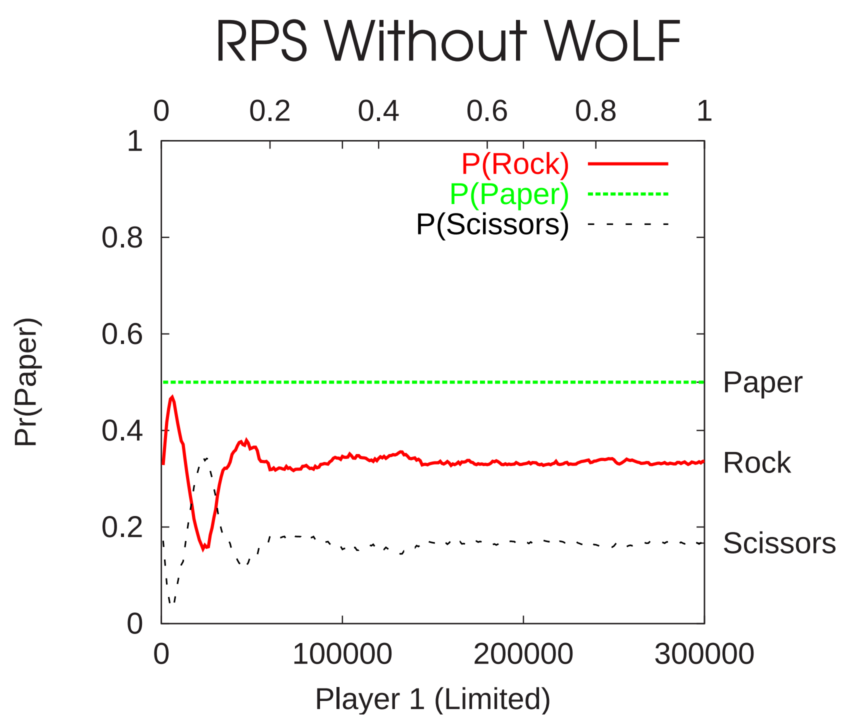
- Win or Learn Fast (WoLF) (Bowling & Veloso, 2002)
  - Variable learning rate accounts for other agents.
    - Learn fast when losing.
    - Cautious when winning, since agents may adapt.
  - Theoretical and empirical evidence of convergence.

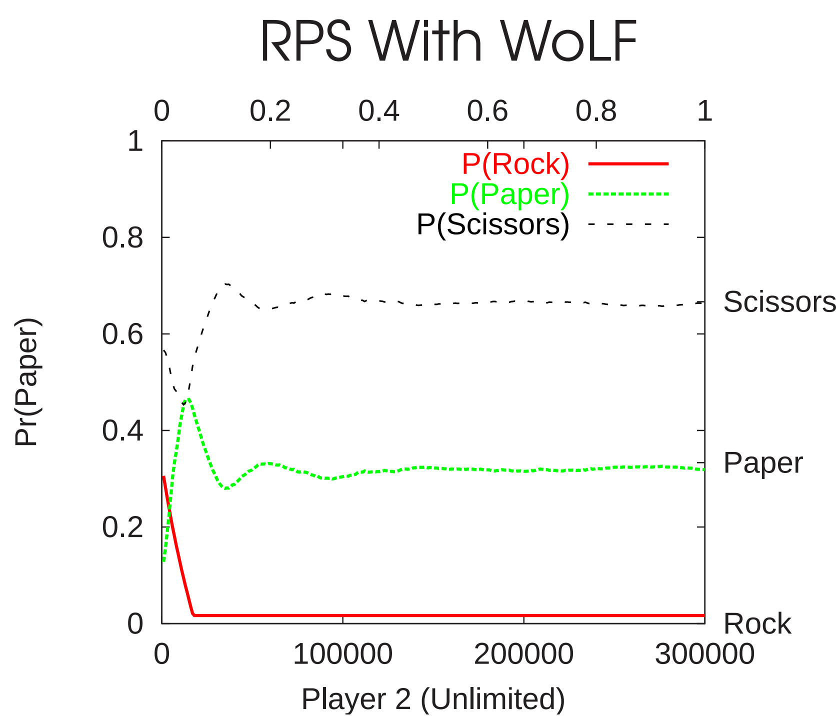




Idea #2 = 2

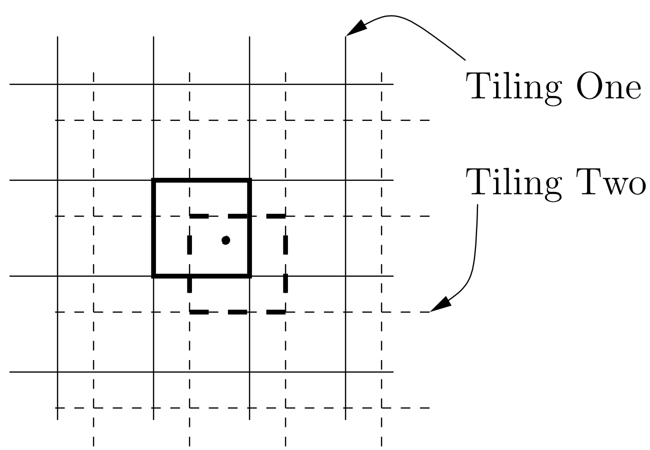

R
P
S
s_1
s_2
Nash Equilibrium
Restricted Policy Space
Restricted Equilibrium


RPS Without WoLF
P(Rock)
P(Paper)
P(Scissors)
Paper
Rock
Scissors
Pr(Paper)
Player 1 (Limited)


RPS With WoLF
P(Rock)
P(Paper)
P(Scissors)
Scissors
Paper
Rock
Pr(Paper)
Player 2 (Unlimited)

Idea #3

- Tile Coding (a.k.a. CMACS) (Sutton & Barto 1998)
  - Space covered by overlapping and offset tilings.
  - Maps continuous (or discrete) spaces to a vector of boolean values.
  - Provides discretization and generalization.



The Task

The Task – Goofspiel

- A.k.a. “The Game of Pure Strategy“

The Task – Goofspiel

- A.k.a. “The Game of Pure Strategy”
- Each player plays a full suit of cards.
- Each player uses their cards (without replacement) to bid on cards from another suit.

The Task – Goofspiel

- A.k.a. "The Game of Pure Strategy"
- Each player plays a full suit of cards.
- Each player uses their cards (without replacement) to bid on cards from another suit.

n	$\lvert S\rvert$	$\lvert S \times A\rvert$	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

- The game is very large.
- Deterministic policies are very bad.
- The random policy isn't too bad.

The Task – Goofspiel – 2

My Hand	1	3	4	5	6	8	11	13
Quartiles	*		*		*	*		*

Opp Hand	4	5	8	9	10	11	12	13
Quartiles	*		*		*	*		*

Deck	1	2	3	5	9	10	11	12
Quartiles	*		*		*	*		*

Card 11
Action 3

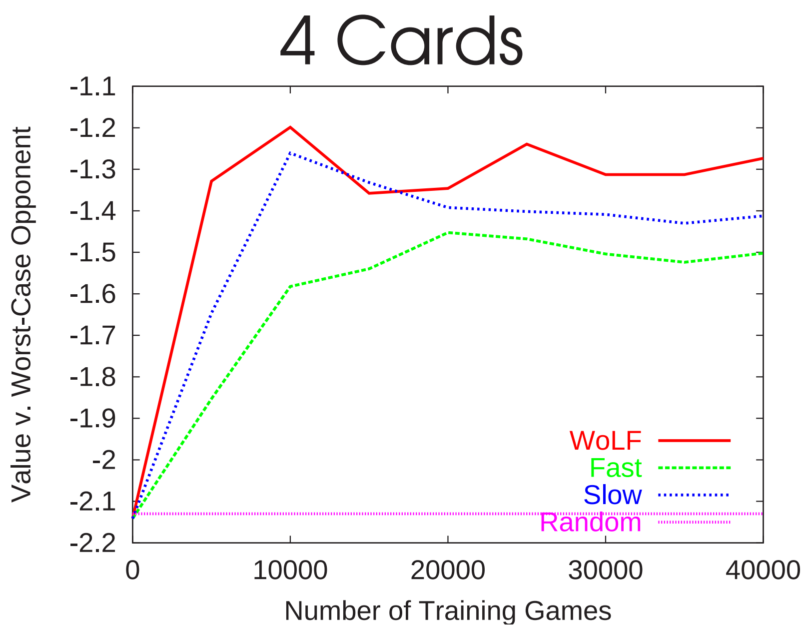
$$\langle 1,4,6,8,13\rangle\,,\quad \langle 4,8,10,11,13\rangle\,,\quad \langle 1,3,9,10,12\rangle\,,\quad 11,3$$

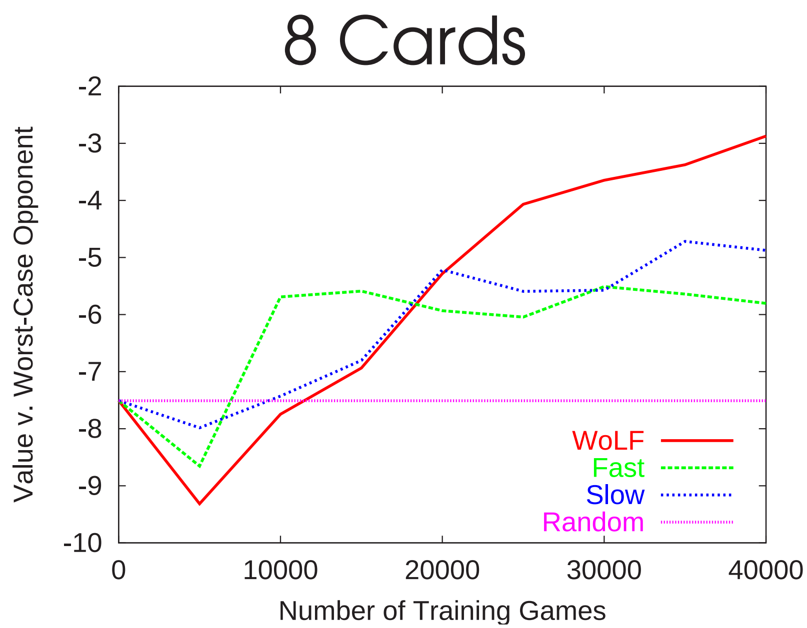
$$\Big\Downarrow \text{(Tile Coding)}$$

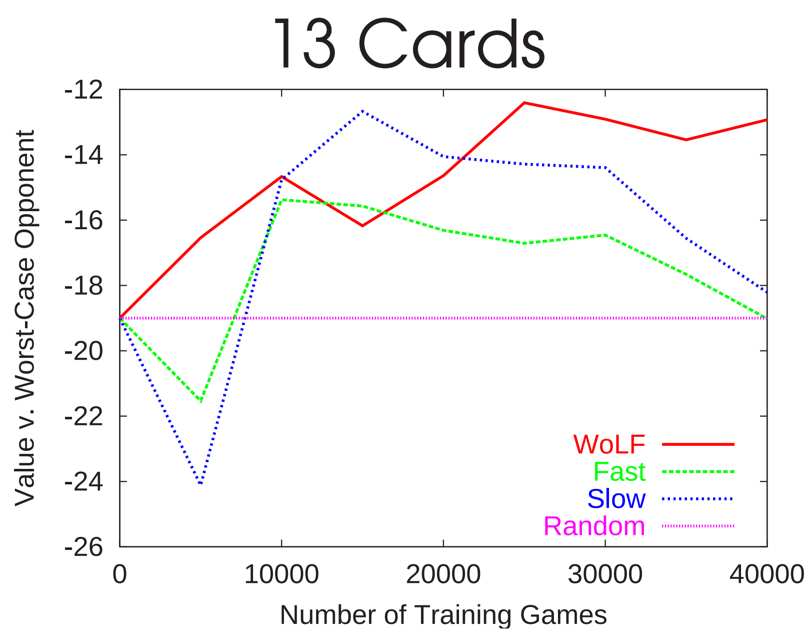
$$\textsc{Tiles} \in \{0,1\}^{10^6}$$

- Gradient ascent on this parameterization.
- WoLF variable learning rate on the gradient step size.

Results – Worst-Case

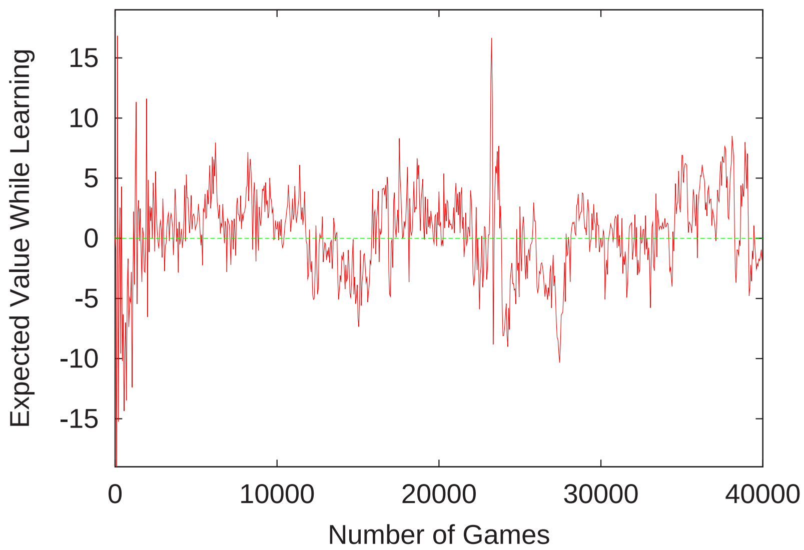




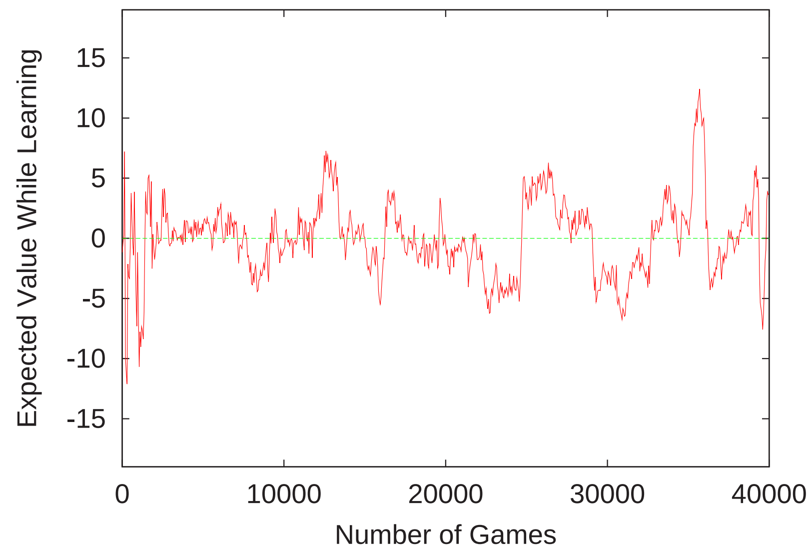


Results – While Learning

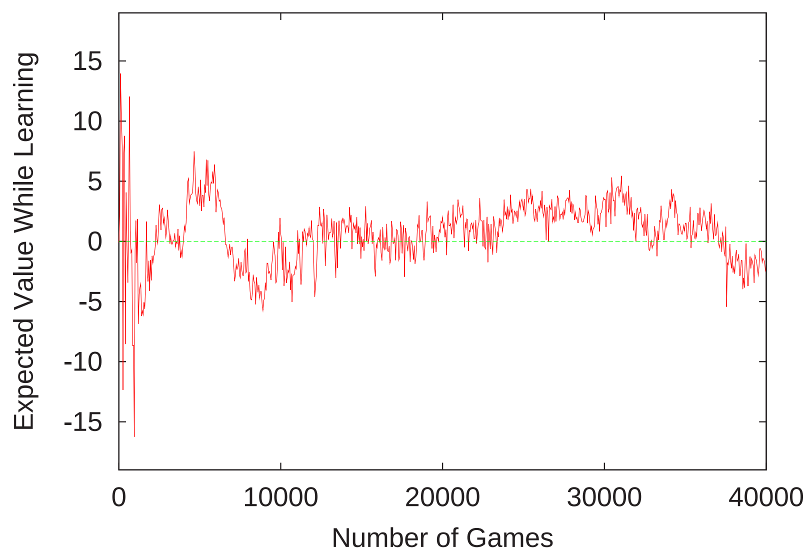
Fast



Slow

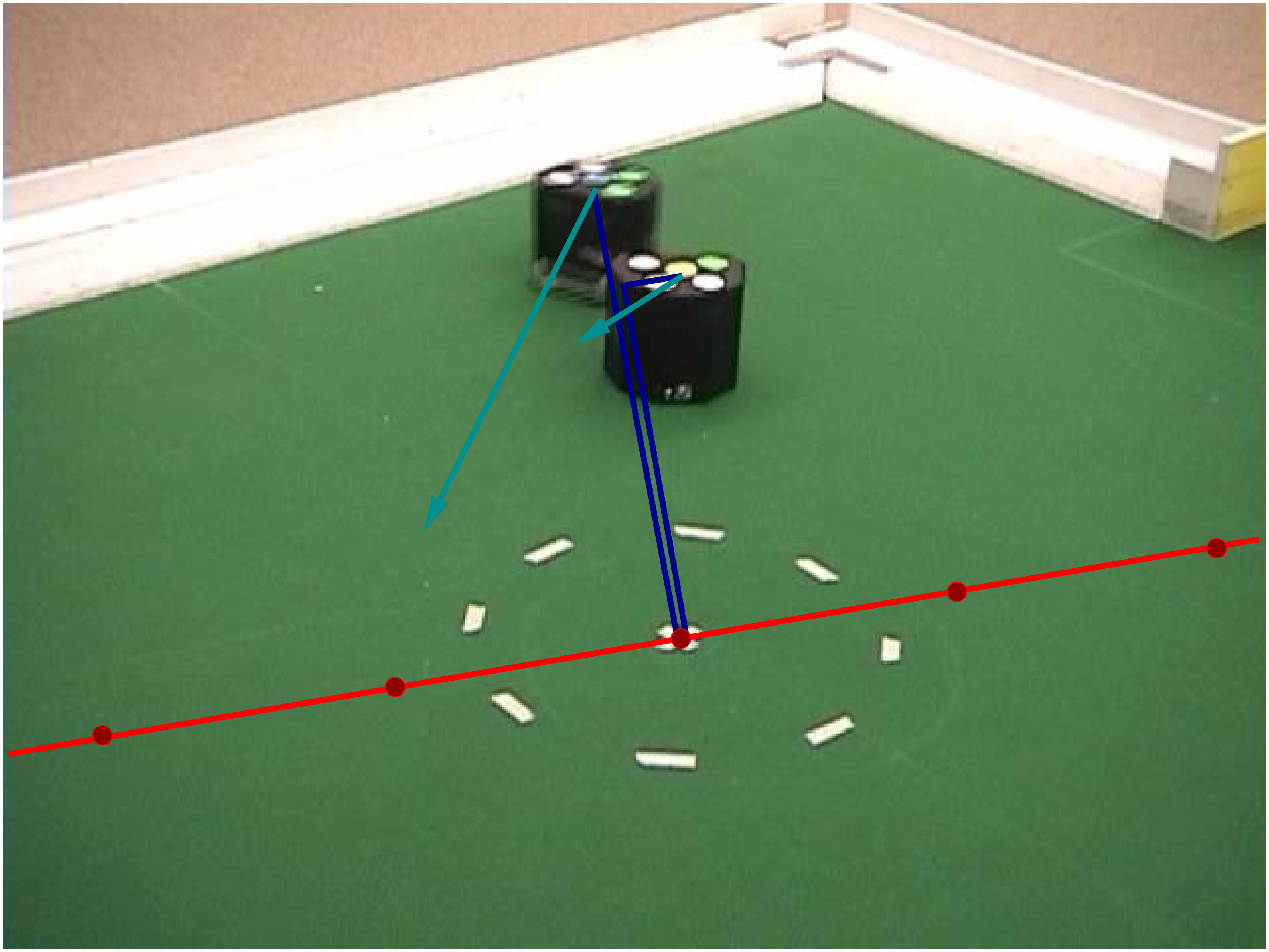


WoLF



The Task - Breakthrough

The Task – Breakthrough – 2



WARNING!

Results = Breakthrough

WARNING!

- These results are preliminary.... some are only hours old.
- They involve a single run of learning in a highly stochastic learning environment.
- More experiments in progress.

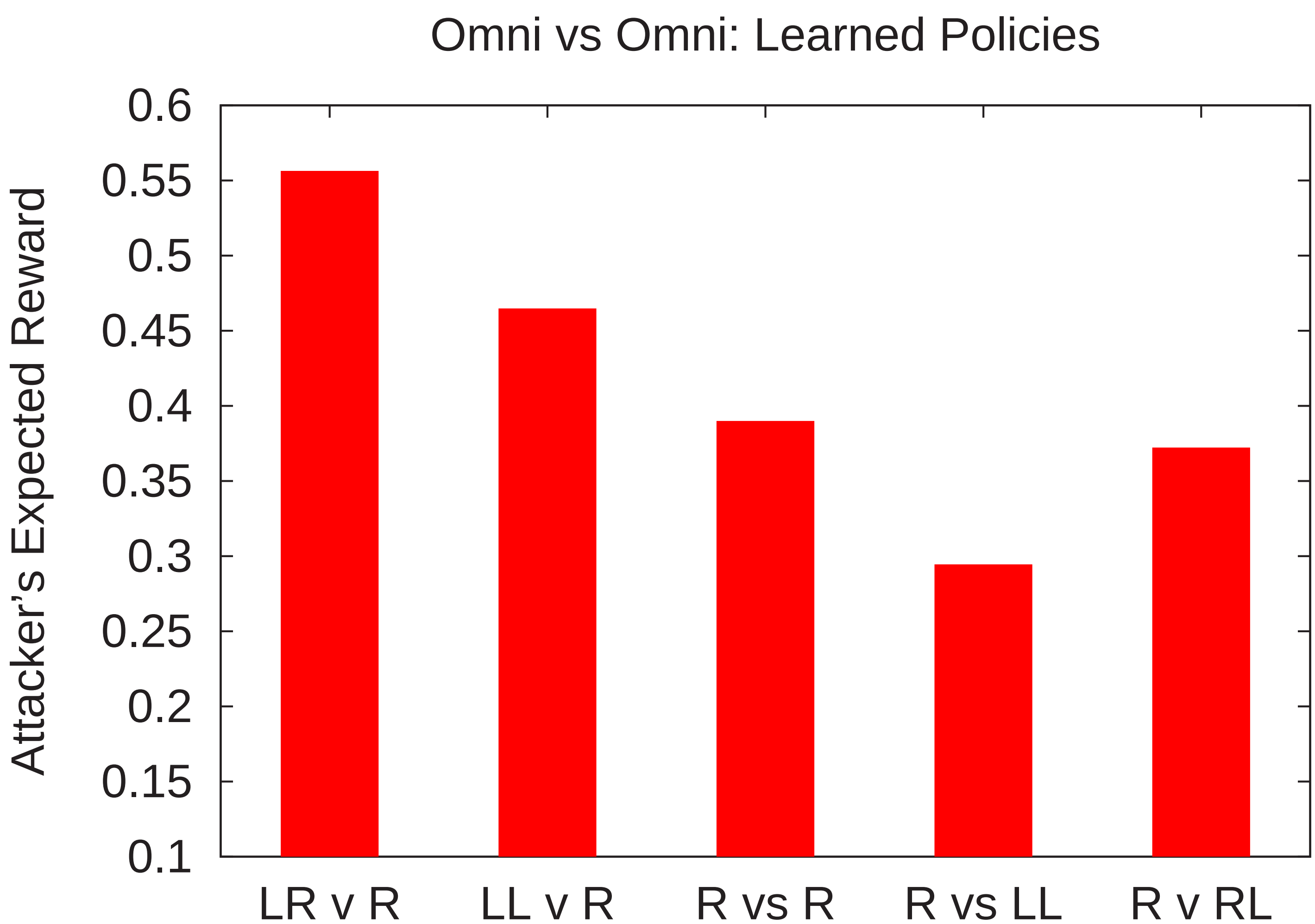
Results – "To the videotape..."

Playback of learned policies in simulation and on the robots.

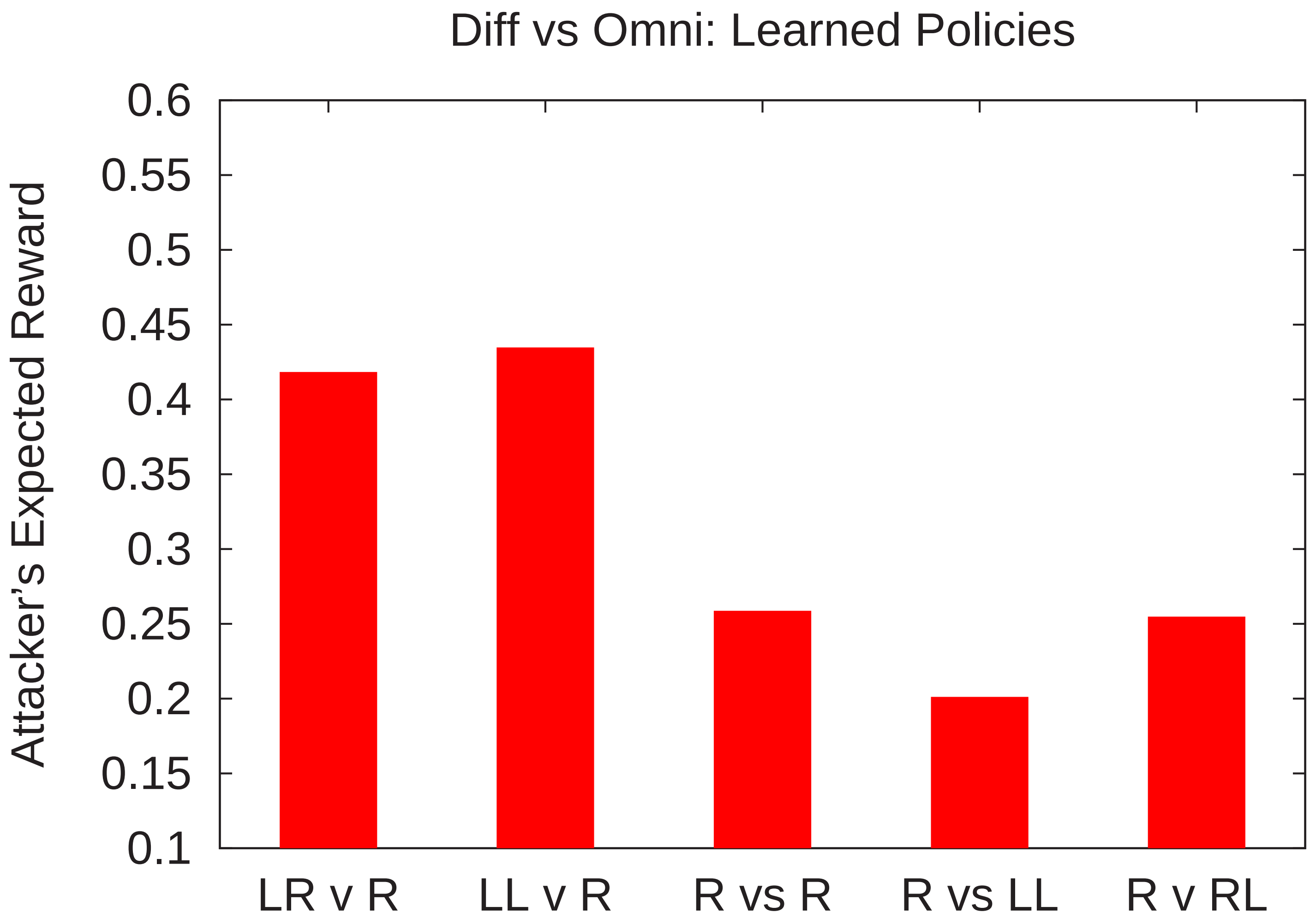
The robot video can be downloaded from. . .

http://www.cs.cmu.edu/~mhb/research/

Results – 3

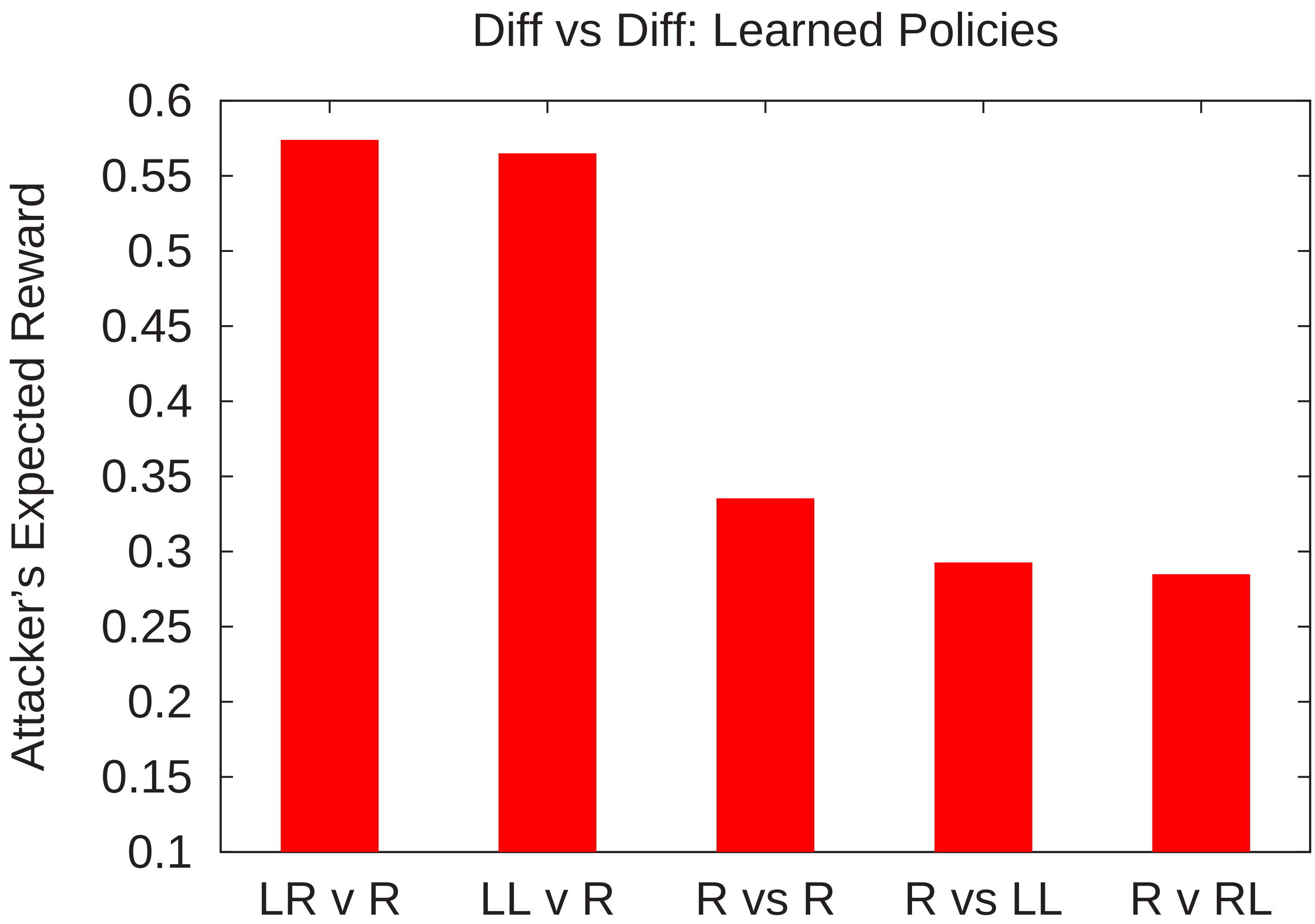

Omni vs Omni: Learned Policies
Attacker's Expected Reward
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
LR v R
LL v R
R vs R
R vs LL
R v RL

Results - 4



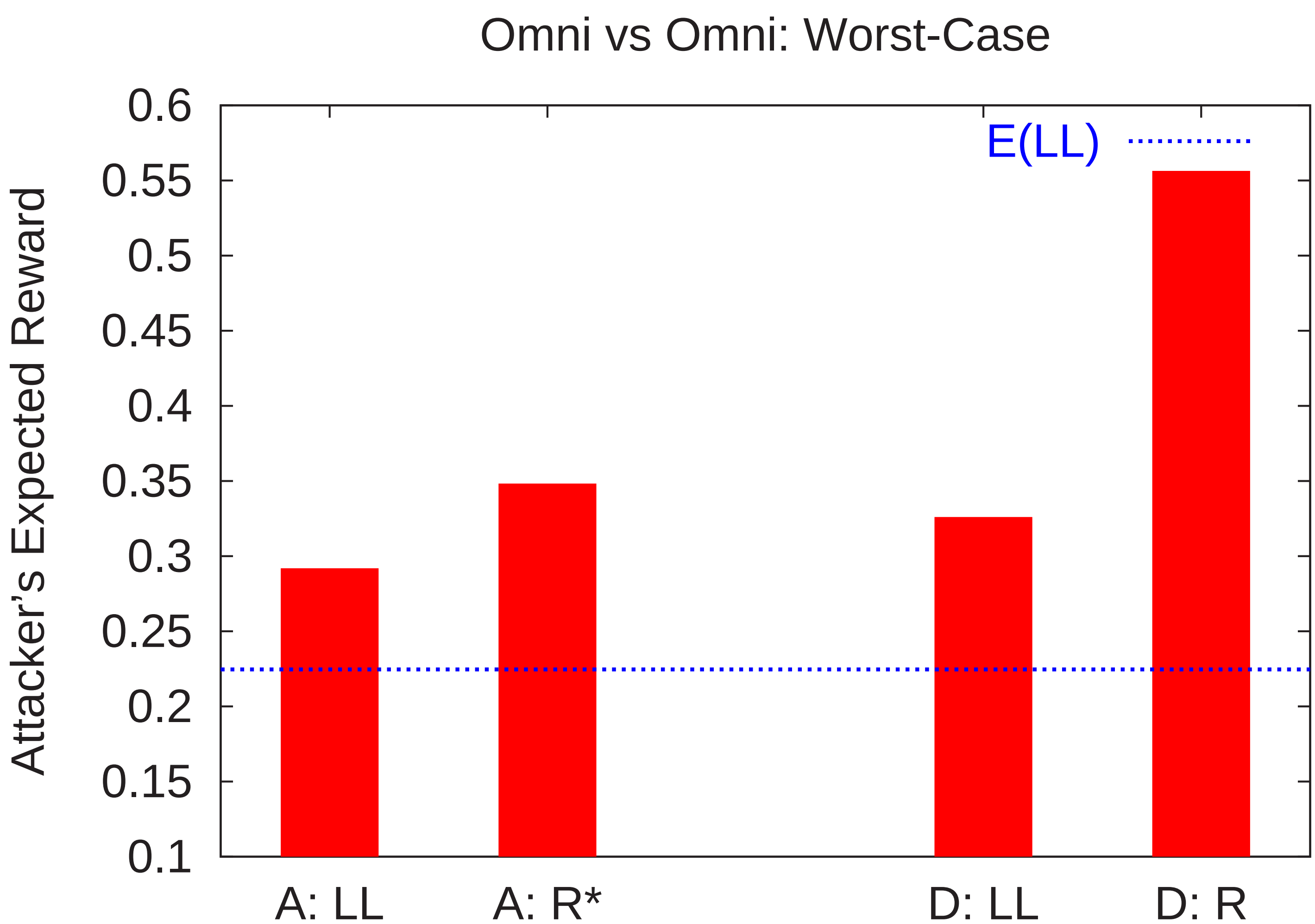
Diff vs Omni: Learned Policies
Attacker's Expected Reward
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
LR v R
LL v R
R vs R
R vs LL
R v RL

Results - 5



Diff vs Diff: Learned Policies
Attacker's Expected Reward
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
LR v R
LL v R
R vs R
R vs LL
R v RL

Results – 6


Omni vs Omni: Worst-Case
Attacker's Expected Reward
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
E(LL)
A: LL
A: R*
D: LL
D: R

Results = Breakthrough

WARNING!

- These results are preliminary.... some are only hours old.
- They involve a single run of learning in a highly stochastic learning environment.
- More experiments in progress.

Big Picture

- How do we scale our (collective) algorithms to large problems with limited agents?

Big Picture

- How do we scale our (collective) algorithms to large problems with limited agents?
  - Equilibrium learning may be in trouble.
  - Lagoudakis and Parr's approximation and minimax. (NIPS '02)
  - Correlated equilibria?

Big Picture

- How do we scale our (collective) algorithms to large problems with limited agents?
  - Equilibrium learning may be in trouble.
  - Lagoudakis and Parr's approximation and minimax. (NIPS '02)
  - Correlated equilibria?
- What is the objective?

Big Picture

- How do we scale our (collective) algorithms to large problems with limited agents?
  - Equilibrium learning may be in trouble.
  - Lagoudakis and Parr's approximation and minimax. (NIPS '02)
  - Correlated equilibria?

- What is the objective?
  - Performance during learning.
  - Generality of learned policies.
    * How can I be exploited?
    * What if everyone played this policy?