Rational Learning of Mixed Equilibiria in Stochastic Games*

Michael Bowling

UAI Workshop: Beyond MDPs
June 30, 2000

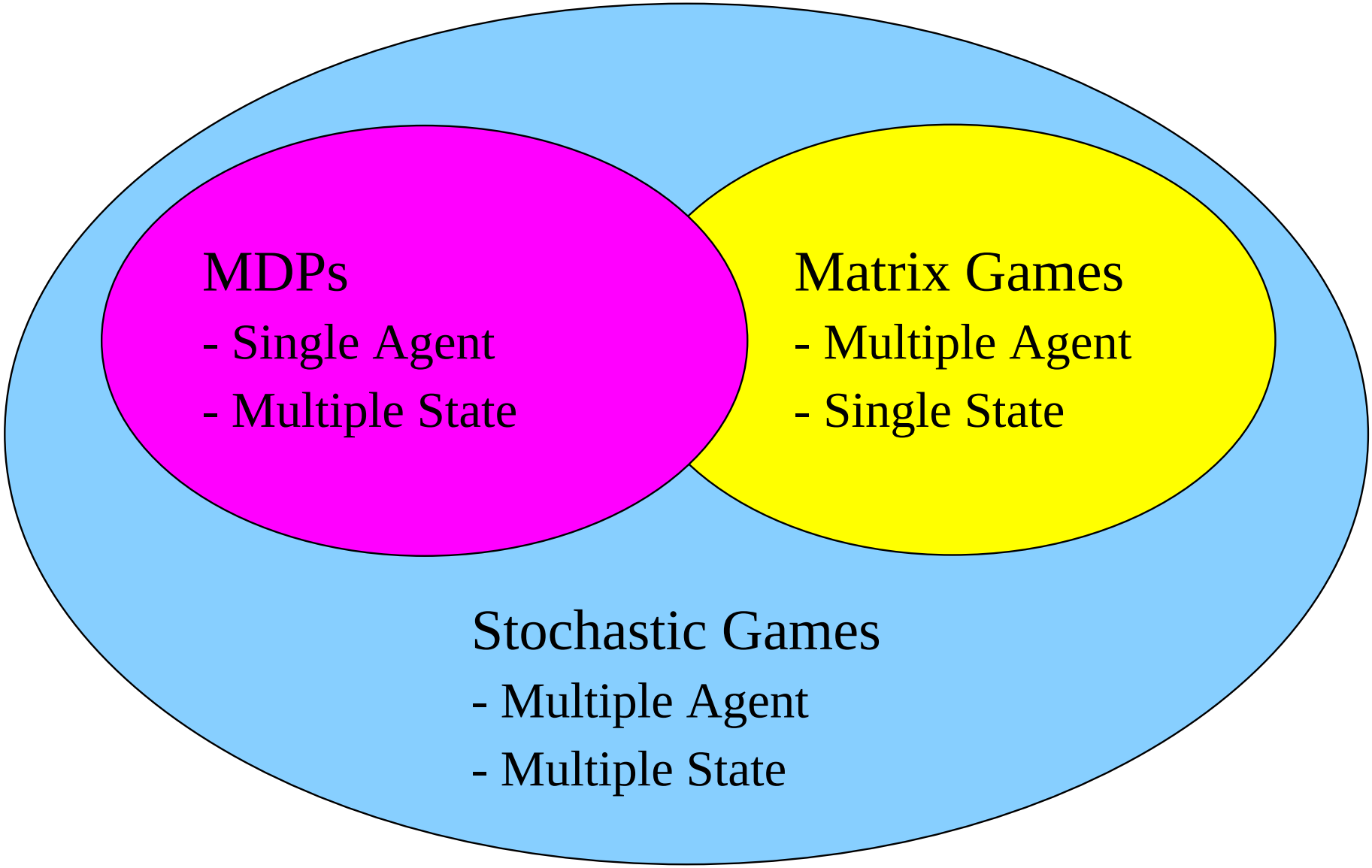
*Joint work with Manuela Veloso

Overview

- Stochastic Game Framework
- Existing Techniques ...

  ... and Their Shortcomings
- A New Algorithm
- Experimental Results

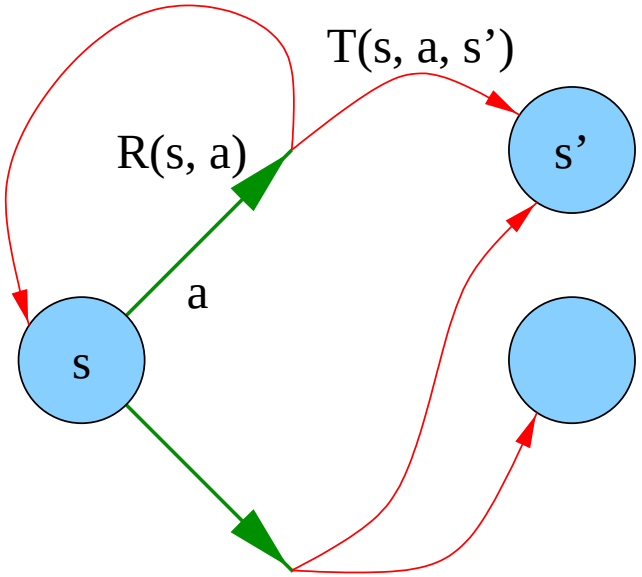
Stochastic Game Framework



Markov Decision Processes

A *Markov decision process* (MDP) is a tuple, $(\mathcal{S}, \mathcal{A}, T, R)$, where,

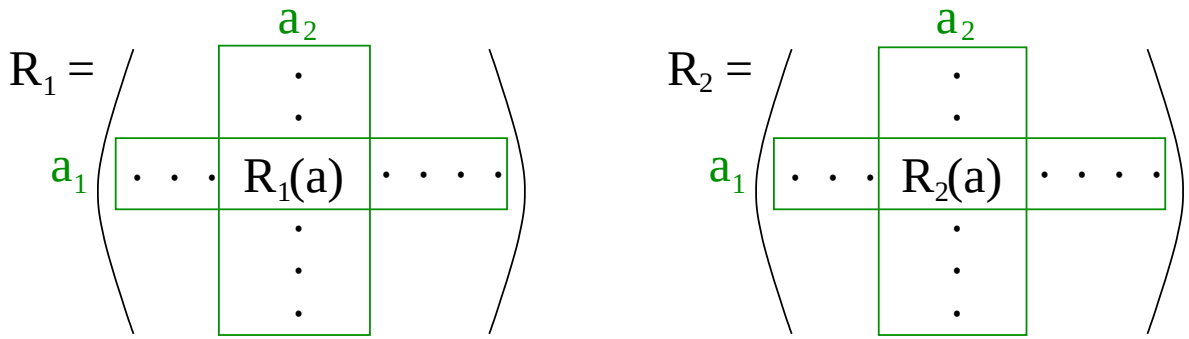
- $\mathcal{S}$ is the set of states,
- $\mathcal{A}$ is the set of actions,
- T is a transition function $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$,
- R is a reward function $\mathcal{S} \times \mathcal{A} \rightarrow \Re$.



Matrix Games

A *matrix game* is a tuple $(n, \mathcal{A}_{1...n}, R_{1...n})$, where,

- n is the number of players,
- $\mathcal{A}_i$ is the set of actions available to player i
  - $\mathcal{A}$ is the joint action space $\mathcal{A}_1 \times \ldots \times \mathcal{A}_n$,
- R_i is player i's payoff function $\mathcal{A} \rightarrow \Re$.



Matrix Games – Example

Rock-Paper-Scissors

- Two players. Each simultaneously picks an action:
  Rock, *Paper*, or *Scissors*.
- The rules:

Rock	beats	*Scissors*
Scissors	beats	*Paper*
Paper	beats	*Rock*

- Represent game as two matrices, one for each player:

$$R_1 = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix} \qquad R_2 = -R_1 = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}$$

Matrix Games – Best Response

- No optimal opponent independent strategies.
- Mixed (i.e. stochastic) strategies does not help.
- Opponent dependent strategies,

  Definition 1 *For a game, define the* best-response *function for player* i, $\mathrm{BR}_i(\sigma_{-i})$, *to be the set of all, possibly mixed, strategies that are optimal given the other player(s) play the possibly mixed joint strategy* σ_{-i}.

Matrix Games – Equilibria

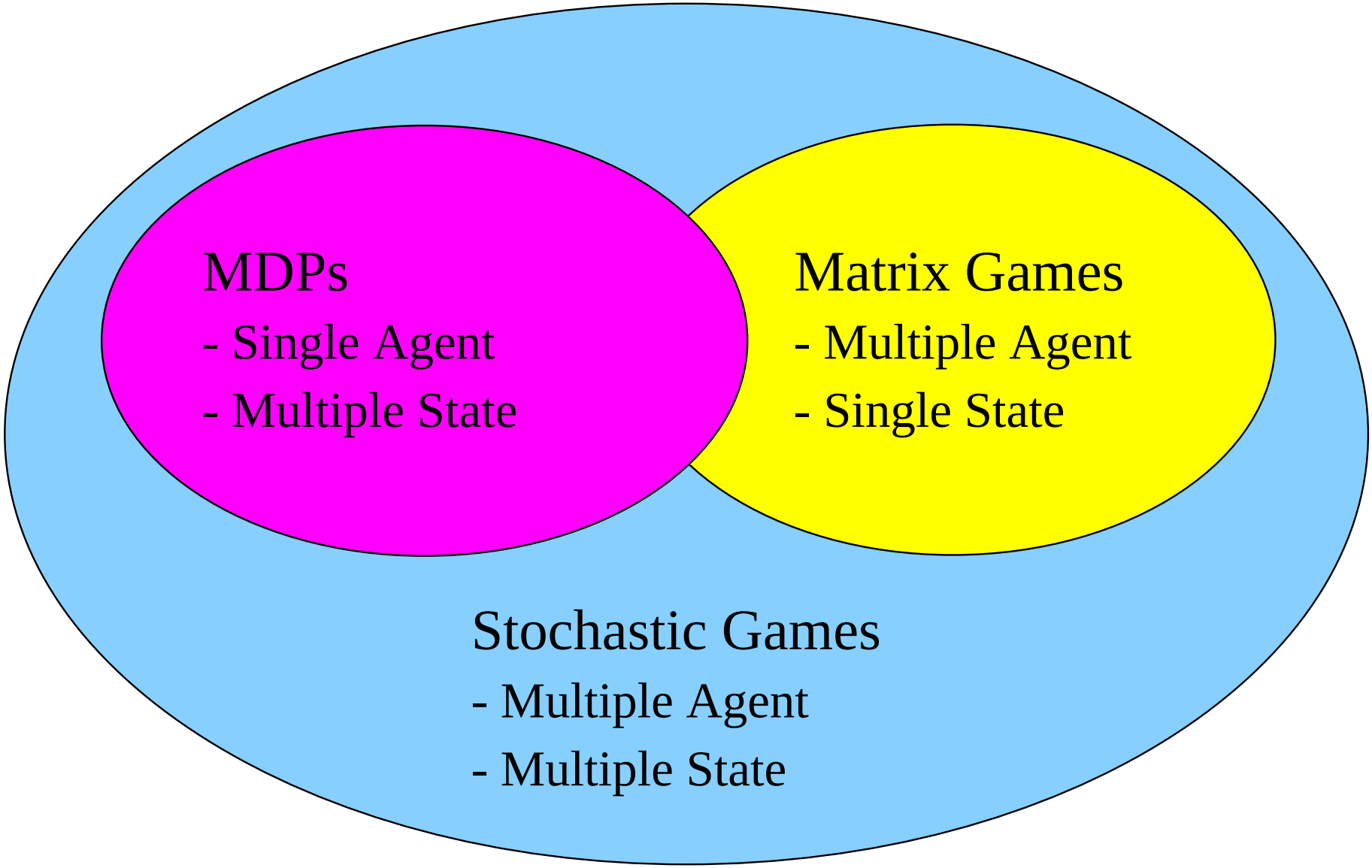
- Best-response equilibrium [Nash, 1950],

  Definition 2 *A* Nash equilibrium *is a collection of strategies (possibly mixed) for all players,* σ_i*, with,*

  $$\sigma_i \in \mathrm{BR}_i(\sigma_{-i}).$$

- An equilibrium in *Rock-Paper-Scissors* consists of both players randomizing evenly among all its actions.

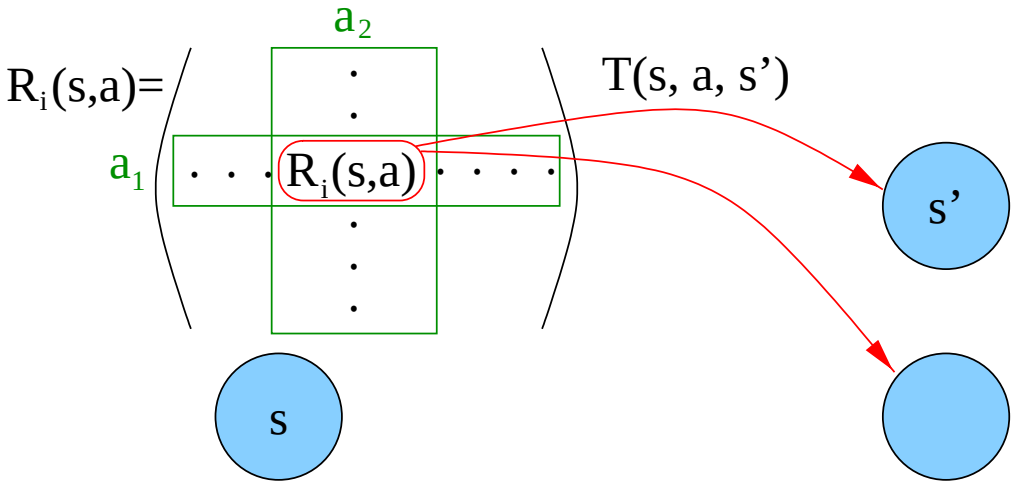
Stochastic Game Framework



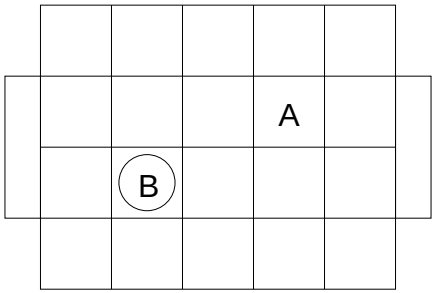
Stochastic Games

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

- 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



- Players: Two
- States: Players' positions and possession of the ball (780).
- Actions: N, S, E, W, Hold (5).
- Transitions:
  - Actions are selected simultaneously but executed in a random order.
  - If a player moves to another player's square, the stationary play gets possession of the ball.
- Rewards: Reward is only received when the ball is moved into one of the goals.

[Littman, 1994]

Solving Stochastic Games

MG	+	MDP	=	Game Theory	RL
LP		TD(0)		Shapley	MiniMax-Q
LP		TD(1)		Pollatschek and Avi-Itzhak	–
LP		TD(λ)		Van der Wal	–
QP		TD(0)		–	Hu and Wellman
FP		TD(0)		Fictitious Play	JALs / Opponent-Modeling

LP: linear programming QP: quadratic programming FP: fictitious play

Minimax-Q

1. Initialize $Q(s \in \mathcal{S}, a \in \mathcal{A})$ arbitrarily.
2. Repeat,
   (a) From state s select action a_i that solves the matrix game $[Q(s,a)_{a \in \mathcal{A}}]$, with some exploration.
   (b) Observing joint-action a, reward r, and next state s',

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma V(s')),$$

   where,

$$V(s) = \text{Value}\left([Q(s,a)_{a \in \mathcal{A}}] \right).$$

[Littman, 1994]

- In zero-sum games, learns equilibrium almost independent of the actions selected by the opponent.

Joint-Action Learners

1. Initialize $Q(s \in \mathcal{S}, a \in \mathcal{A})$ arbitrarily.
2. Repeat,
   (a) From state s select action a_i that maximizes,

$$\sum_{a_{-i}} \frac{C(s, a_{-i})}{n(s)} Q(s, \langle a_i, a_{-i} \rangle)$$

   (b) Observing other agents' actions a_{-i}, reward r, and next state s',

$$\begin{aligned} C(s, a_{-i}) &\leftarrow C(s, a_{-i}) + 1 \\ n(s) &\leftarrow n(s) + 1 \\ Q(s, \langle a_i, a_{-i} \rangle) &\leftarrow (1 - \alpha) Q(s, \langle a_i, a_{-i} \rangle) + \alpha (r + \gamma V(s')) \end{aligned}$$

   where,

$$V(s) = \max_{a_i} \sum_{a_{-i}} \frac{C(s, a_{-i})}{n(s)} Q(s, \langle a_i, a_{-i} \rangle).$$

[Claus & Boutilier, 1998; Uther & Veloso, 1997]

Joint-Action Learners

- Finds equilibrium (when playing another JAL) in:
  - Fully collaborative games [Claus & Boutilier, 1998],
  - Iterated dominance solvable games [Fudenberg & Levine, 1998],
  - Fully competitive games [Uther & Veloso, 1997].
- Plays deterministically (i.e. cannot play mixed policies).

Problems with Existing Algorithms

- Minimax-Q
  - Converges to an equilibrium, independent of the opponent's actions.
  - Will not converge to a best-response unless the opponent also plays the equilibrium solution.
    - Consider a player that almost always plays *Rock*.

- Q-Learning, JALs, etc.
  - Always seeks to maximize reward.
  - Does not converge to stationary policies if the opponent is also learning.
    - Cannot play mixed strategies.

Properties

Property 1 *(Rational) If the other players' strategies converge to stationary strategies then the player will converge to a strategy that is optimal given their strategies.*

Property 2 *(Convergent) Given that the other players are following behaviors from a class of behaviors, $\mathcal{B}$, all the players will converge to stationary strategies.*

Algorithm	Rational	Convergent
Minimax-Q	No	Yes
JAL	Yes	No

- If all players are rational and they converge to stationary strategies, they must have converged to an equilibrium.
- If all players are both rational and convergent, then they are guaranteed to converge to an equilibrium.

A New Algorithm – Policy Hill-Climbing

1. Let α and δ be learning rates. Initialize,

$$Q(s,a) \leftarrow 0, \qquad \pi(s,a) \leftarrow \frac{1}{|\mathcal{A}_i|}.$$

2. Repeat,

   (a) From state s select action a according to mixed strategy $\pi(s)$ with some exploration.

   (b) Observing reward r and next state s',

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a')\right).$$

   (c) Update $\pi(s,a)$ and constrain it to a legal probability distribution,

$$\pi(s,a) \leftarrow \pi(s,a) + \begin{cases} \delta & \text{if } a = \text{argmax}_{a'} Q(s,a') \\ \frac{-\delta}{|A_i|-1} & \text{Otherwise} \end{cases}.$$

- PHC is rational, but still not convergent.

A New Algorithm – Adjusted Policy Hill-Climbing

- APHC preserves rationality, while encouraging convergence.
  - Makes a change only to the algorithm's learning rate.
  - "Learn faster while losing, slower while winning."

1. Let α, $\delta_l > \delta_w$ be learning rates. Initialize,

$$Q(s,a) \leftarrow 0, \qquad \pi(s,a) \leftarrow \frac{1}{|\mathcal{A}_i|},$$

2. Repeat,

(a,b) Same as PHC.

(c) Maintain running estimate of average policy, $\bar{\pi}$.

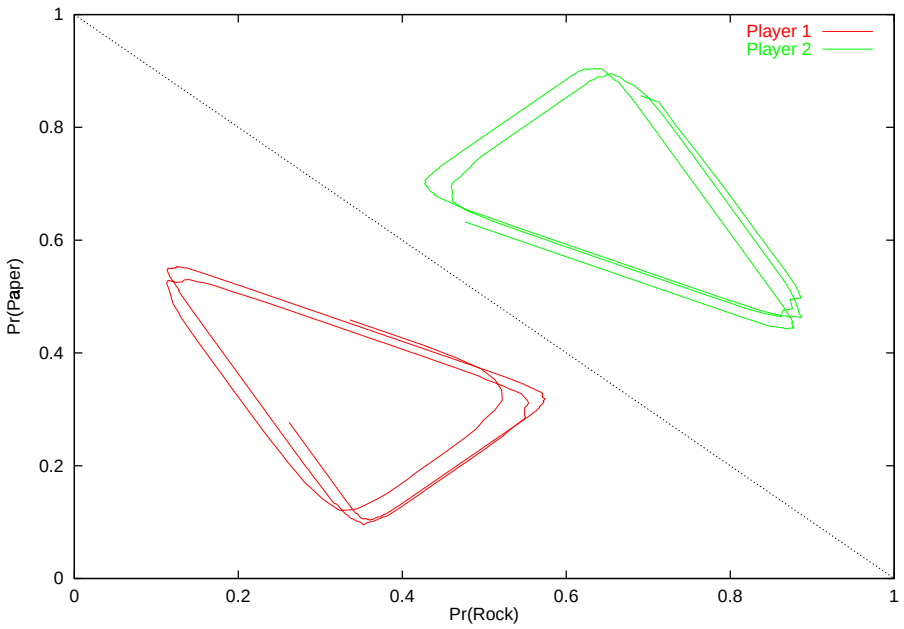
(d) Update $\pi(s,a)$ and constrain it to a legal probability distribution,

$$\pi(s,a) \leftarrow \pi(s,a) + \begin{cases} \delta & \text{if } a = \text{argmax}_{a'} Q(s,a') \\ \frac{-\delta}{|A_i|-1} & \text{Otherwise} \end{cases},$$

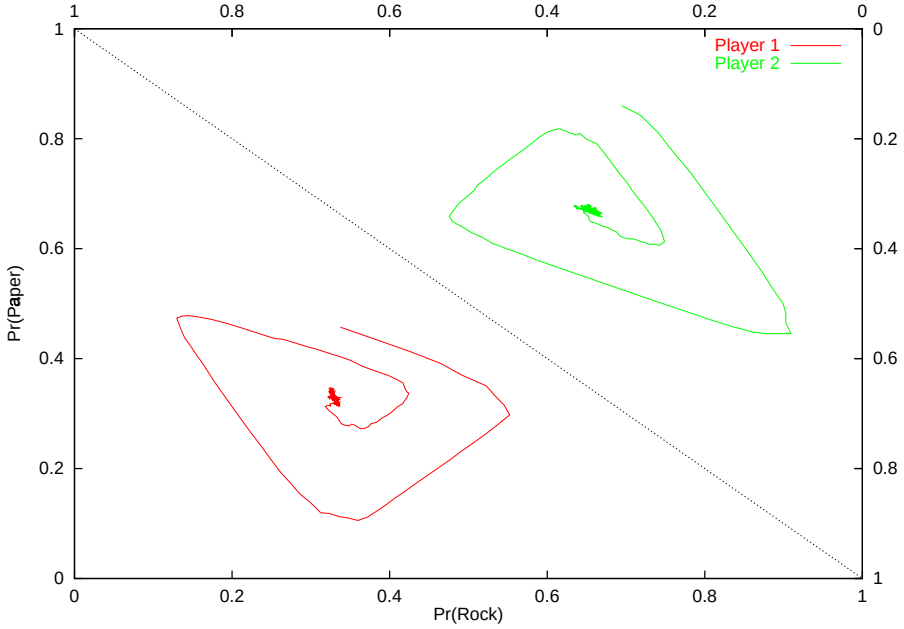
where,

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

Results – Rock-Paper-Scissors

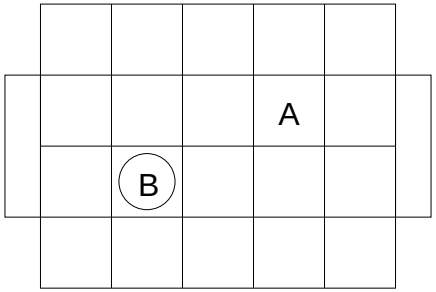


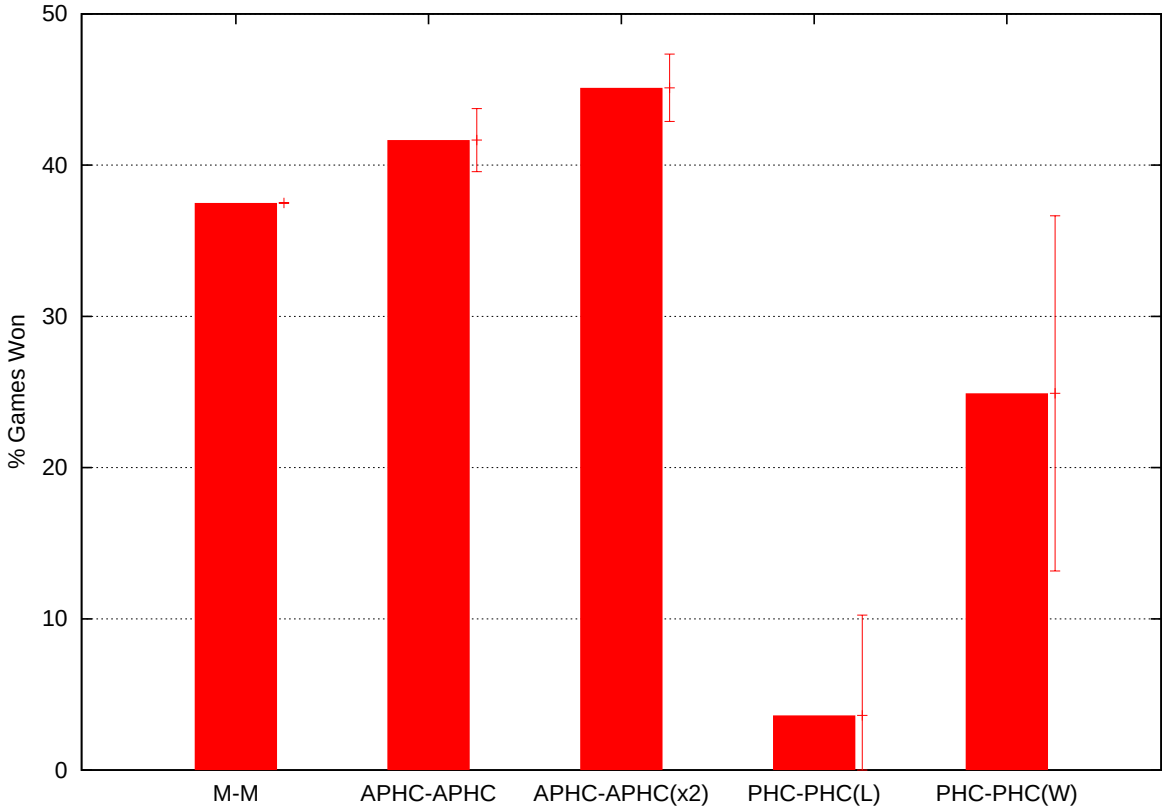
PHC



APHC

Results – Soccer





Discussion

- Why convergence?
  - Non-stationary policies are hard to evaluate.
  - Complications with assigning delayed reward.
- Why rationality?
  - Multiple equilibria.
  - Opponent may not be playing optimally.

- What's next?
  - More experimental results on more interesting problems.
  - Family of learning algorithms.
  - Theoretical analysis of convergence.
  - Learning in the presence of agents with limitations.

`http://www.cs.cmu.edu/~mhb/publications/`