

Generalized Multiagent Learning with Performance Bound

Bikramjit Banerjee & Jing Peng
Department of Electrical Engr. & Computer Science
Tulane University



Outline

- Why Multiagent Systems?
- Reinforcement Learning
- Game Theoretic Solution Concepts
- Our contribution and comparison
- Demo

Why Multiagent Systems?

Multiagent Systems provide the following benefits to real problem solving domains:

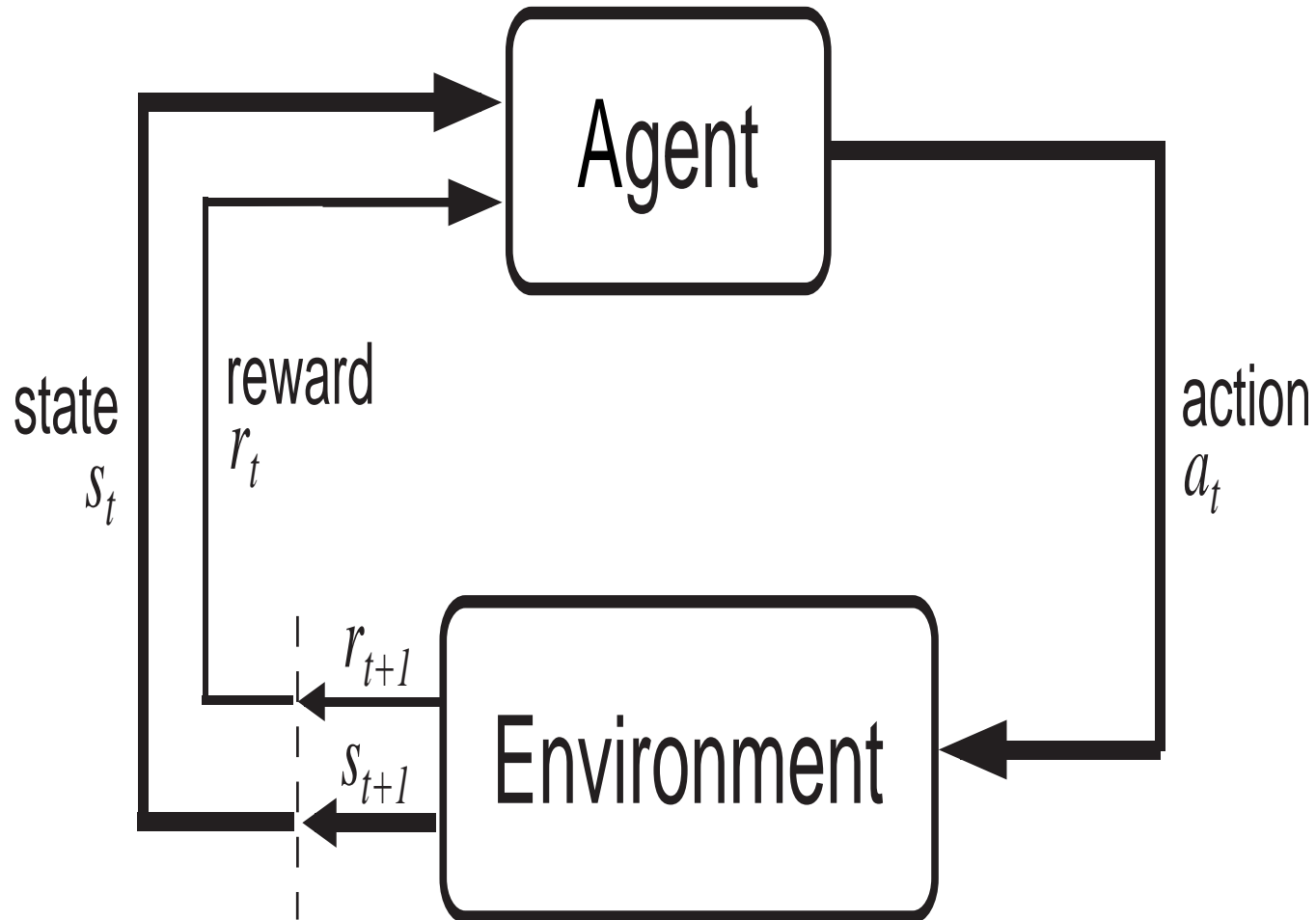
- Concurrency \Rightarrow speed
- Redundancy \Rightarrow fault tolerance

Why Multiagent Systems?

Distributed Problem Solving (DPS) in complex domains can suffer from the following drawbacks that MAS can address :

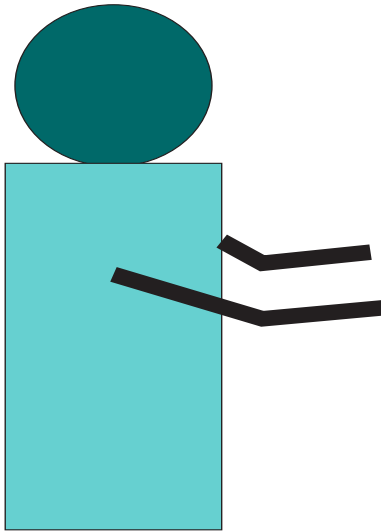
- Knowledge of the entire system may not be available precluding a central controller ⇒ **Distributed Knowledge.**
- System may be too complex to perceive by a central controller ⇒ **Distributed Control.**
- All (future) states of the system may not be known beforehand or too cumbersome to hardwire exhaustively ⇒ **Learn appropriate behavior (Multiagent Learning).**
- Certain subgoals may necessitate dynamic rearrangement ⇒ **MAS with social capabilities like negotiation, competition, cooperation and coalition building.**

The Reinforcement Learning Scenario

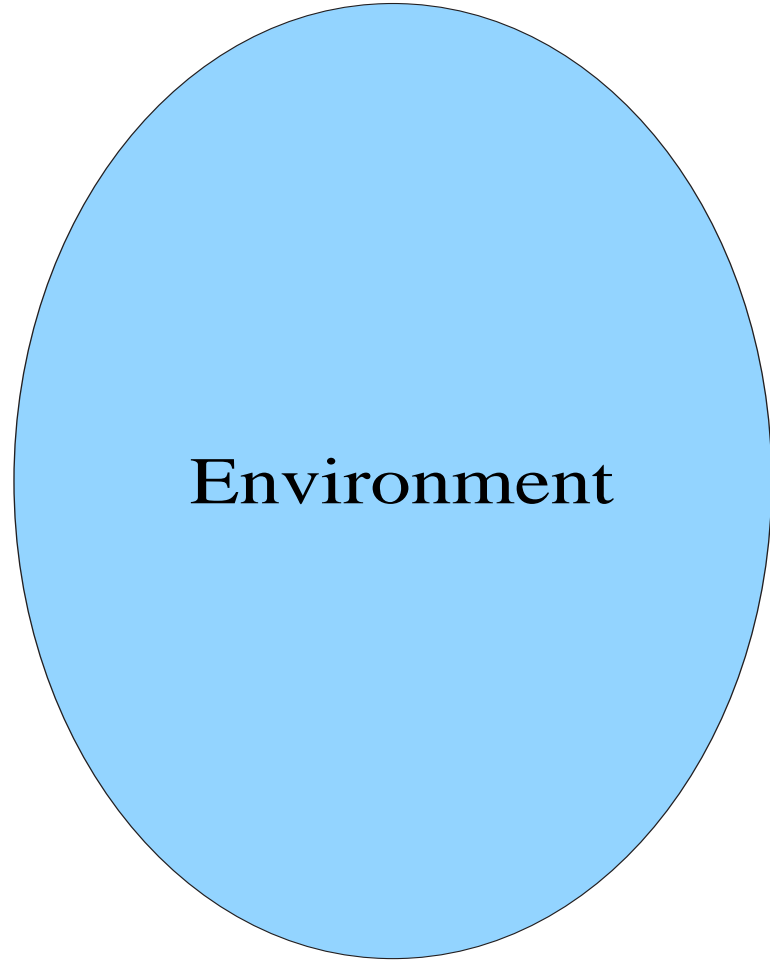


The Reinforcement Learning Scenario

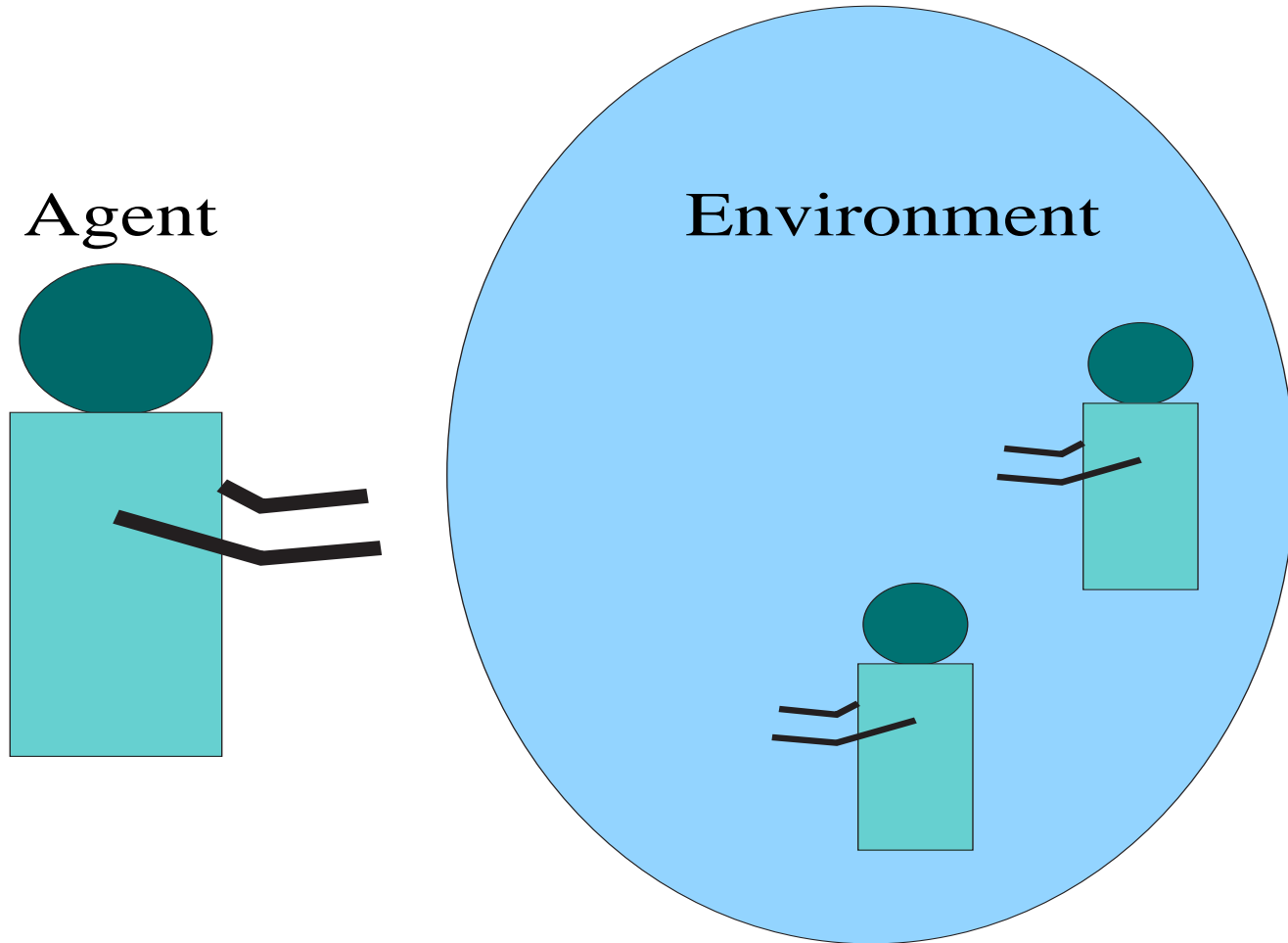
Agent



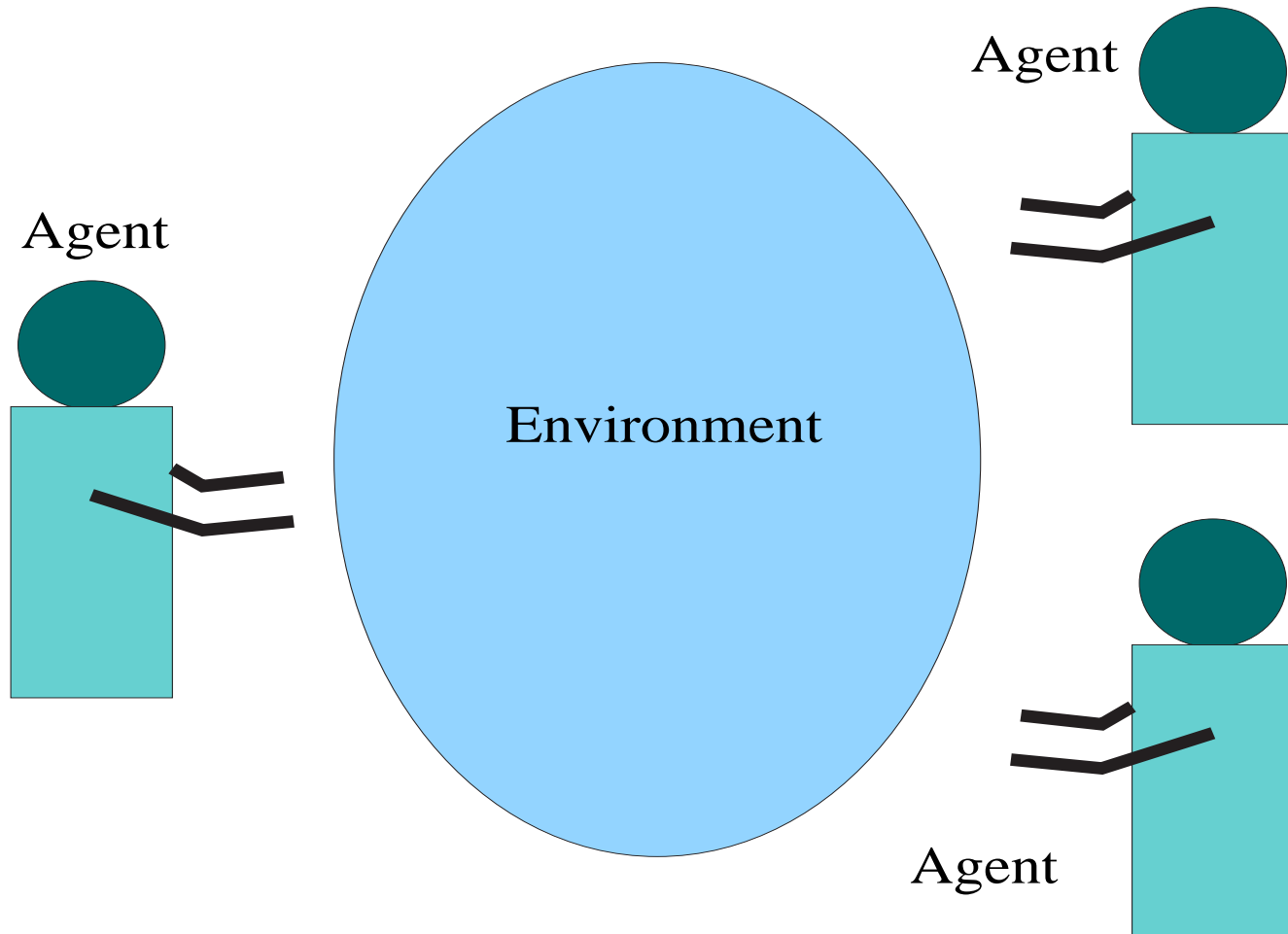
Environment



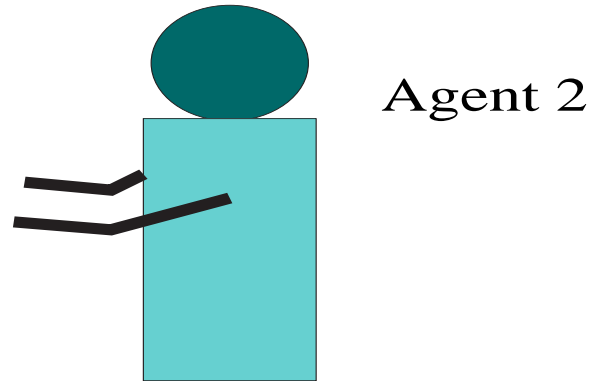
The Concurrent Reinforcement Learning Scenario



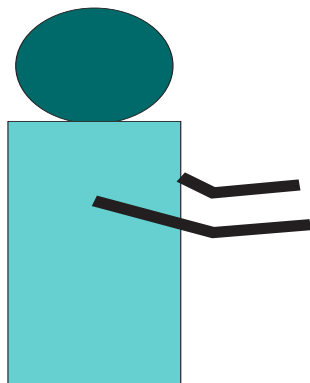
The Equivalent Scenario









The Equivalent Scenario (With only 2 agents)



Agent 1



| | Action 1 | Action 2 | | Action n |
|----------|---|--|------|---|
| Action 1 |  |  | |  |
| Action 2 |  |  | |  |
| ⋮ | | | | |
| ⋮ | | | | |
| ⋮ | | | | |
| ⋮ | | | | |
| Action n | | | | |

Bimatrix Games: Example

Rock-Scissors-Paper game

- **Rock** breaks **Scissors**
- **Scissors** cuts **Paper**
- **Paper** wraps **Rock**

Bimatrix Games: Example

Rock-Scissors-Paper game

- **Rock** breaks **Scissors**
- **Scissors** cuts **Paper**
- **Paper** wraps **Rock**

Row Player's Payoffs:

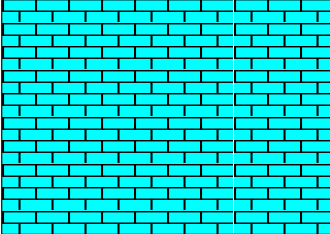



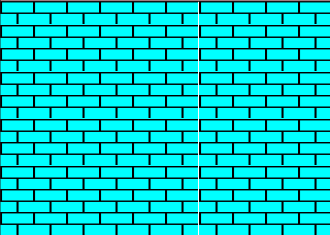



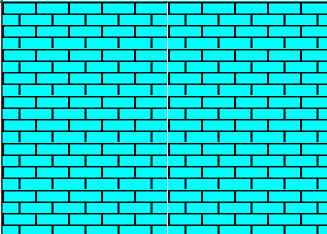
| ACTIONS | ROCK (R) | SCISSOR (S) | PAPER (P) |
|-------------|----------|-------------|-----------|
| ROCK (R) | 0 | 1 | -1 |
| SCISSOR (S) | -1 | 0 | 1 |
| PAPER (P) | 1 | -1 | 0 |

Column Player's Payoffs:

| ACTIONS | ROCK (R) | SCISSOR (S) | PAPER (P) |
|-------------|----------|-------------|-----------|
| ROCK (R) | 0 | -1 | 1 |
| SCISSOR (S) | 1 | 0 | -1 |
| PAPER (P) | -1 | 1 | 0 |

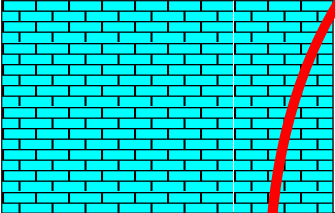



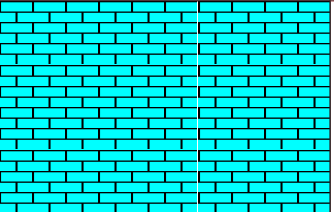



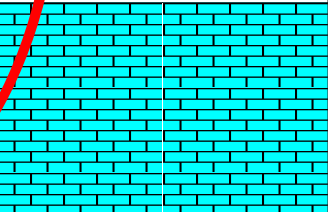
Solution Concepts: Best Response

What is my **best** strategy against the opponent's **current** strategy?

| | Rock | Scissors | Paper |
|-----------------|---|---|---|
| Rock |  |  |  |
| Scissors |  |  |  |
| Paper |  |  |  |

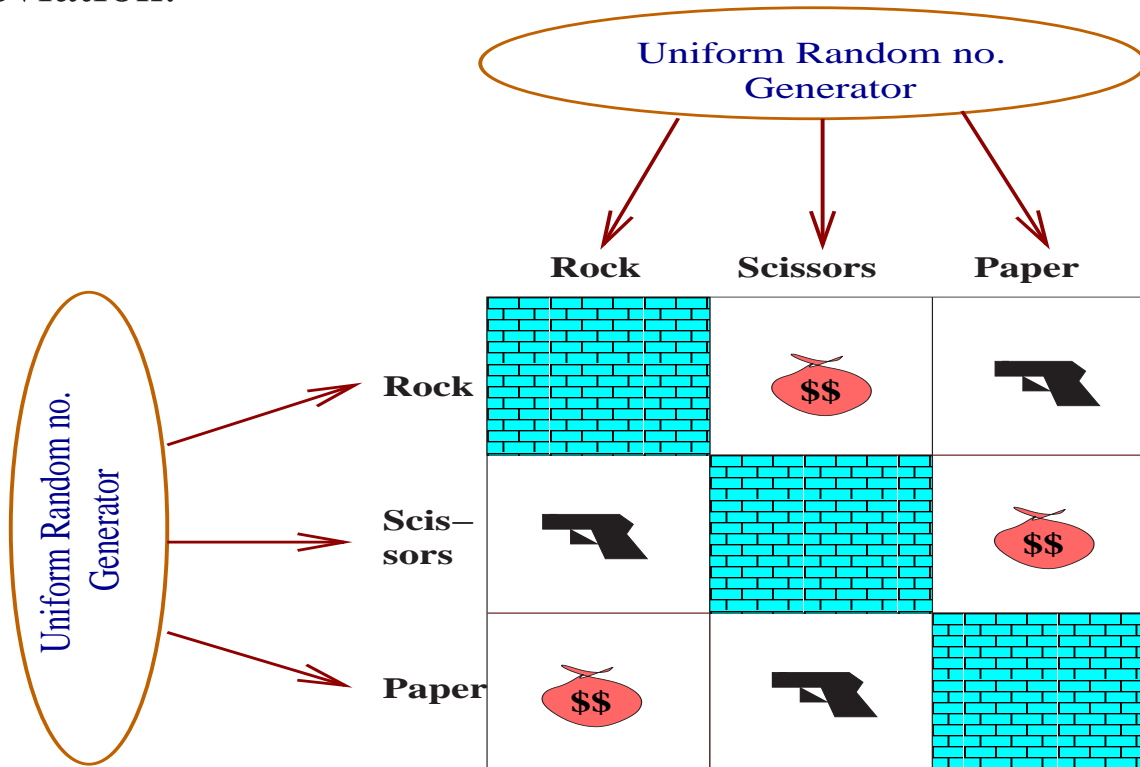
Solution Concepts: Best Response

What is my **best** strategy against the opponent's **current** strategy?

| | Rock | Scissors | Paper |
|----------|---|---|---|
| Rock |  |  |  |
| Scissors |  |  |  |
| Paper |  |  |  |

Solution Concepts: Nash Equilibrium

Mutual Best Responses, i.e., none has incentive for **unilateral** deviation.



Regret

| | R | S | P |
|---|----|----|----|
| R | 0 | 1 | -1 |
| S | -1 | 0 | 1 |
| P | 1 | -1 | 0 |

Sequence of 10 actions:

R S S P R P S R R R

R P R R S P S R P P

W W W

L L L

T T T

Net payoff = 0

Regret

| | R | S | P |
|---|----|----|----|
| R | 0 | 1 | -1 |
| S | -1 | 0 | 1 |
| P | 1 | -1 | 0 |

Sequence of 10 actions:

R S S P R P S R R R

R P R R S P S R P P

Instead play following sequence :

P R R S P S R P P P

Always Wins !

Regret

| | R | S | P |
|---|----|----|----|
| R | 0 | 1 | -1 |
| S | -1 | 0 | 1 |
| P | 1 | -1 | 0 |

Sequence of 10 actions:

R S S P R P S R R R

R P R R S P S R P P

Instead play following sequence :

P R R S P S R P P P

~~Always Wins !~~

Impossible !!

Regret

| | R | S | P |
|---|----|----|----|
| R | 0 | 1 | -1 |
| S | -1 | 0 | 1 |
| P | 1 | -1 | 0 |

Sequence of 10 actions:

R S S P R P S R R R

R P R R S P S R P P

Instead play following sequence :

R R R R R R R R R R

W W W
L L
T T T

Regret for not playing R's : 1-0

Regret

| | R | S | P |
|---|----|----|----|
| R | 0 | 1 | -1 |
| S | -1 | 0 | 1 |
| P | 1 | -1 | 0 |

Sequence of 10 actions:

R S S P R P S R R R

R P R R S P S R P P

Instead play following sequence :

S S S S S S S S S S

L W W
 L L L L L
 T T T

Regret for not playing R's : 1-0

Regret for not playing S's : -3-0

Regret

| | R | S | P |
|---|----|----|----|
| R | 0 | 1 | -1 |
| S | -1 | 0 | 1 |
| P | 1 | -1 | 0 |

Sequence of 10 actions:

R S S P R P S R R R

R P R R S P S R P P

Instead play following sequence :

P P P P P P P P P P

W W W W W
 L L L
 T T

Regret for not playing R's : $1-0$

Regret for not playing S's : $-3-0$

Regret for not playing P's : $2-0$

Regret

| | R | S | P |
|---|----|----|----|
| R | 0 | 1 | -1 |
| S | -1 | 0 | 1 |
| P | 1 | -1 | 0 |

Sequence of 10 actions:

R S S P R P S R R R

R P R R S P S R P P

Instead play following sequence :

P P P P P P P P P P

W W W W W
 L L L
 T T

Regret for not playing R's : 1-0
 Regret for not playing S's : -3-0
 Regret for not playing P's : 2-0



Max = 2 = External Regret

Average External Regret = 2/10

Performance Bounded Learning against unknown Opponents

Our Contribution

A new MAL algorithm (**ReDVaLeR**) that

- converges to stationary best response against eventually stationary opponents.
- converges to Nash Equilibrium in self-play.
- asymptotically attains no-regret payoff against all types of opponents.

without knowing the type of the opponents.

Comparison with other algorithms

| IGA | WoLF-IGA | AWESOME | R-MAX | ReDVaLeR |
|-------------------|-------------------------------|-------------------------------------|-------------------|-------------------------------|
| 1. Game | | Size | | |
| 2×2 | 2×2 | $m \times n$ | $m \times n$ (SG) | $m \times n$ |
| 2. Game | | Payoffs | | |
| Bounded | Bounded | Bounded | Bounded, 0-sum | Bounded, +ve |
| 3. Learner | | knows | | |
| own payoffs | own payoffs + own eq. pol. | own payoffs + entire eq. profile | max payoff | own payoffs + own eq. pol. |
| 4. Learner | | can observe | | |
| policy | policy | actions | actions | policies |

Conclusion

Latest Improvement

Need not compare with sequence of fixed actions. Can improve to comparing with sequences like **RRRR PPPPP SSSSSSSSS RRRRRR SSSS RRRR...** Achieves higher payoff bound by no-regret learning.

Future Improvements

- May be possible to achieve **negative** regrets if opponents' strategies never converge.
- To ensure above property while maintaining previous properties of ReDVaLeR.