Multiagent Learning in Games

Amy Greenwald

Brown University

with

David Gondek, Keith Hall, Amir Jafari,
Michael Littman, Casey Marks, John Wicks, Martin Zinkevich

American Association of Artificial Intelligence

July 11, 2005

Key Problem

What is the outcome of multiagent learning in games?

Key Problem

What is the outcome of multiagent learning in games?

Candidate Solutions

Game-theoretic equilibria

- Minimax equilibria [von Neumann 1944]
- Nash equilibria [Nash 1951]
- Correlated equilibria [Aumann 1974]

Key Problem

What is the outcome of multiagent learning in games?

Candidate Solutions

Game-theoretic equilibria

- Minimax equilibria [von Neumann 1944]
- Nash equilibria [Nash 1951]
- Correlated equilibria [Aumann 1974]
- Cyclic equilibria [ZGL 2005]
- Φ-equilibria [GJ 2003]

Convergence is a Slippery Slope

I. Multiagent value iteration (Q-learning) in Markov games

- convergence to cyclic equilibrium policies [ZGL 2005]

II. No-regret learning in repeated games [Foster & Vohra 1997]

- convergence to a set of game-theoretic equilibria [GJ 2003]

III. Adaptive learning in repeated games [Young 1993]

- stochastic stability and equilibrium selection [WG 2005]

Game Theory: A Crash Course

General-Sum Games (e.g., Prisoners' Dilemma)

- Correlated Equilibrium
- Nash Equilibrium

Zero-Sum Games (e.g., Rock-Paper-Scissors)

- Minimax Equilibrium

An Example

Chicken

	l	r
T	6,6	2,7
B	7,2	0,0

CE

	l	r
T	1/2	1/4
B	1/4	0

$$\pi_{Tl} + \pi_{Tr} + \pi_{Bl} + \pi_{Br} = 1 \quad (1)$$

$$\pi_{Tl}, \pi_{Tr}, \pi_{Bl}, \pi_{Br} \geq 0 \quad (2)$$

$$6\pi_{l|T} + 2\pi_{r|T} \geq 7\pi_{l|T} + 0\pi_{r|T} \quad (3)$$

$$7\pi_{l|B} + 0\pi_{r|B} \geq 6\pi_{l|B} + 2\pi_{r|B} \quad (4)$$

$$6\pi_{T|l} + 2\pi_{B|l} \geq 7\pi_{T|l} + 0\pi_{B|l} \quad (5)$$

$$7\pi_{T|r} + 0\pi_{B|r} \geq 6\pi_{T|r} + 2\pi_{B|r} \quad (6)$$

Linear Program

Chicken

	l	r
T	6,6	2,7
B	7,2	0,0

CE

	l	r
T	1/2	1/4
B	1/4	0

$$\max\ 12\pi_{Tl} + 9\pi_{Tr} + 9\pi_{Bl} + 0\pi_{Br} \tag{7}$$

subject to

$$\pi_{Tl} + \pi_{Tr} + \pi_{Bl} + \pi_{Br} = 1 \tag{8}$$

$$\pi_{Tl}, \pi_{Tr}, \pi_{Bl}, \pi_{Br} \geq 0 \tag{9}$$

$$6\pi_{Tl} + 2\pi_{Tr} \geq 7\pi_{Tl} + 0\pi_{Tr} \tag{10}$$

$$7\pi_{Bl} + 0\pi_{Br} \geq 6\pi_{Bl} + 2\pi_{Br} \tag{11}$$

$$6\pi_{Tl} + 2\pi_{Bl} \geq 7\pi_{Tl} + 0\pi_{Bl} \tag{12}$$

$$7\pi_{Tr} + 0\pi_{Br} \geq 6\pi_{Tr} + 2\pi_{Br} \tag{13}$$

One-Shot Games

General-Sum Games

- N is a set of players
- A_i is player i's action set
- $R_i : A \rightarrow \mathbb{R}$ is player i's reward function, where $A = \prod_{i \in N} A_i$

Zero-Sum Games

- $\sum_i R_i(\vec{a}) = 0$, for all $\vec{a} \in A$

Equilibria

Notation

Write $\vec{a} = (a_i, \vec{a}_{-i}) \in A$ for $a_i \in A_i$ and $\vec{a}_{-i} \in A_{-i} = \prod_{j \neq i} A_j$ and $\Pi = \Delta(A)$

Definition

An action profile $\pi^* \in \Pi$ is a correlated equilibrium if for all $i \in N$, $a_i, a_i' \in A_i$, if $\pi(a_i) > 0$,

$$\sum_{\vec{a}_{-i} \in A_{-i}} \pi(\vec{a}_{-i} \mid a_i)\, R_i(a_i, \vec{a}_{-i}) \quad \geq \quad \sum_{\vec{a}_{-i} \in A_{-i}} \pi(\vec{a}_{-i} \mid a_i)\, R_i(a_i', \vec{a}_{-i}) \tag{14}$$

A Nash equilibrium is an independent correlated equilibrium.

A minimax equilibrium is a Nash equilibrium in a zero-sum game.

I. Multiagent Value Iteration in Markov Games

Theory

Multiagent value iteration does not necessarily converge to stationary equilibrium policies in general-sum Markov games.

Experiments

Multiagent value iteration converges to cyclic equilibrium policies

- randomly generated Markov games
- Grid Game 1 [Hu and Wellman 1998]
- Shopbots and Pricebots [G and Kephart 1999]

Markov Decision Processes (MDPs)

Decision Process

- S is a set of states
- A is a set of actions
- $R : S \times A \to \mathbb{R}$ is a reward function
- $P[s_{t+1} \mid s_t, a_t, \ldots, s_0, a_0]$ is a probabilistic transition function that describes transitions between states, conditioned on past states and actions

MDP = Decision Process + Markov Property:

$$P[s_{t+1} \mid s_t, a_t, \ldots, s_0, a_0] = P[s_{t+1} \mid s_t, a_t]$$

$\forall t,\ \forall s_0, \ldots, s_t \in S,\ \forall a_0, \ldots, a_t \in A$

Bellman's Equations

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s'} P[s' \mid s,a] V^*(s') \tag{15}$$

$$V^*(s) = \max_{a \in A} Q^*(s,a) \tag{16}$$

Value Iteration

VI(MDP, γ)

Inputs	discount factor γ
Output	optimal state-value function V^*
	optimal action-value function Q^*
Initialize	V arbitrarily

```
REPEAT
      for all s ∈ S
           for all a ∈ A
              Q(s,a) = R(s,a) + γ Σ_{s'} P[s' | s,a] V(s')
           V(s) = max_a Q(s,a)
FOREVER
```

Markov Games

Stochastic Game

- N is a set of players
- S is a set of states
- A_i is the ith player's set of actions
- $R_i(s, \vec{a})$ is the ith player's reward at state s given action vector $\vec{a}$
- $P[s_{t+1} \mid s_t, \vec{a}_t, \ldots, s_0, \vec{a}_0]$ is a probabilistic transition function that describes transitions between states, conditioned on past states and actions

Markov Game = Stochastic Game + Markov Property:

$$P[s_{t+1} \mid s_t, \vec{a}_t, \ldots, s_0, \vec{a}_0] = P[s_{t+1} \mid s_t, \vec{a}_t]$$

$\forall t$, $\forall s_0, \ldots, s_t \in S$, $\forall \vec{a}_0, \ldots, \vec{a}_t \in A$

Bellman's Analogue

$$Q_i^*(s, \vec{a}) = R_i(s, \vec{a}) + \gamma \sum_{s'} P[s' \mid s, \vec{a}] V_i^*(s') \quad (17)$$

$$V_i^*(s) = \sum_{\vec{a} \in A} \pi^*(s, \vec{a}) Q_i^*(s, \vec{a}) \quad (18)$$

Foe–VI $\pi^*(s) = (\sigma_1^*, \sigma_2^*)$, a minimax equilibrium policy
[Shapley 1953, Littman 1994]

Friend–VI $\pi^*(s) = e_{\vec{a}^*}$ where $\vec{a}^* \in \arg\max_{\vec{a} \in A} Q_i^*(s, \vec{a})$
[Littman 2001]

Nash–VI $\pi^*(s) \in \text{Nash}(Q_1^*(s), \ldots, Q_n^*(s))$
[Hu and Wellman 1998]

CE–VI $\pi^*(s) \in \text{CE}(Q_1^*(s), \ldots, Q_n^*(s))$
[GH 2003]

Multiagent Value Iteration

MULTI–VI(MGame, γ, f)

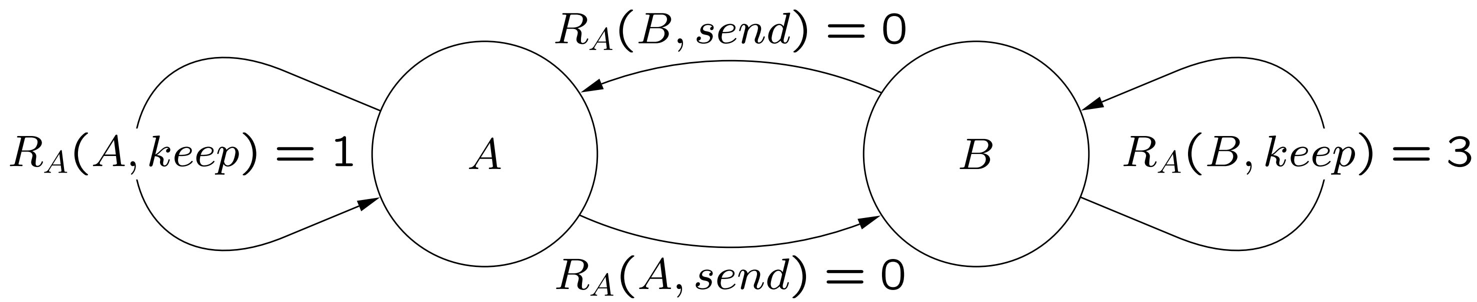
Inputs	discount factor γ
	selection mechanism f
Output	equilibrium state-value function V^*
	equilibrium action-value function Q^*
	equilibrium policy π^*
Initialize	V arbitrarily

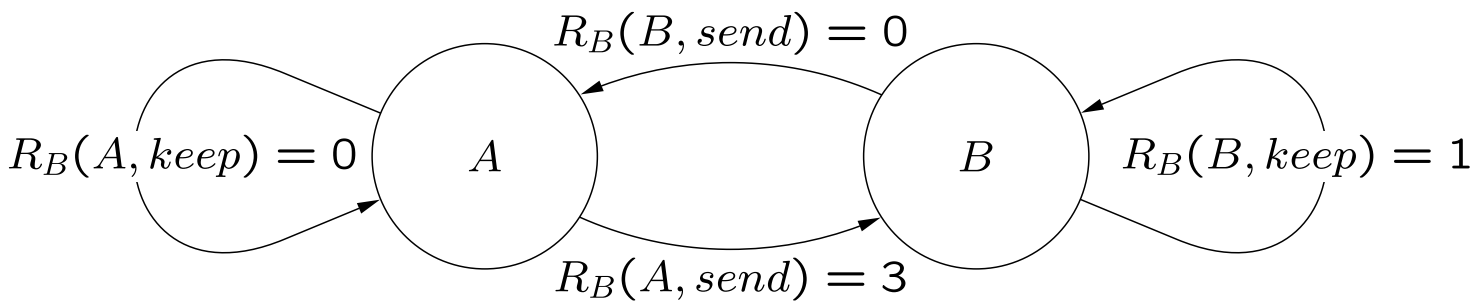
REPEAT
    for all $s \in S$
        for all $\vec{a} \in A$
            for all $i \in N$
                $Q_i(s, \vec{a}) = R_i(s, \vec{a}) + \gamma \sum_{s'} P[s' \mid s, \vec{a}] V_i(s')$
        $\pi(s) \in f(Q_1(s), \ldots, Q_n(s))$
        for all $i \in N$
            $V_i(s) = \sum_{\vec{a} \in A} \pi(s, \vec{a}) Q_i(s, \vec{a})$
FOREVER

Friend–or–Foe–VI *always* converges [Littman 2001]

Nash–VI and CE–VI converge *to equilibrium policies* in zero-sum & common-interest Markov games [GHZ 2005]

NoSDE Game: Rewards

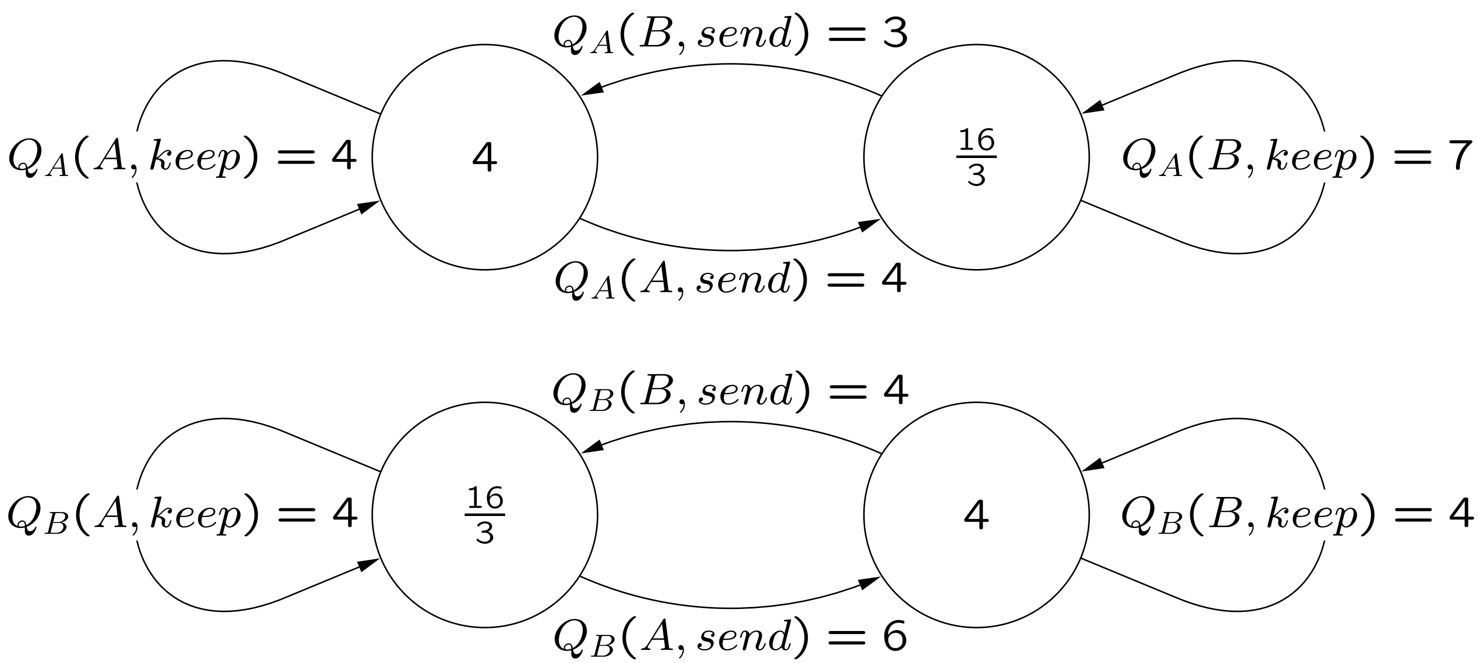




Observation [ZGL 2005]

This game has no stationary deterministic equilibrium policy when $\gamma = \frac{3}{4}$.

NoSDE Game: Q-Values and Values



Theorem [ZGL 2005]

Every NoSDE game has a unique (probabilistic) stationary equilibrium policy.

Cyclic Correlated Equilibria

A stationary policy is a function $\pi : S \to \Delta(A)$.

A cyclic policy ρ is a finite sequence of stationary policies.

$$Q_i^{\rho,t}(s,\vec{a}) = R_i(s,\vec{a}) + \gamma \sum_{s' \in S} P[s' \mid s,\vec{a}] V_i^{\rho,\tilde{t}+1}(s') \tag{19}$$

$$V_i^{\rho,t}(s) = \sum_{\vec{a} \in A} \rho_t(s,\vec{a}) Q_i^{\rho,t}(s,\vec{a}) \tag{20}$$

A cyclic policy of length k is a correlated equilibrium

if for all $i \in N$, $s \in S$, $a_i' \in A_i$, and $t \in \{1, \ldots, k\}$,

$$\sum_{\vec{a}_{-i} \in A_{-i}} \rho_t(s,\vec{a}_{-i} \mid a_i) Q_i^{\rho,t}(s,\vec{a}_{-i},a_i) \geq \sum_{\vec{a}_{-i} \in A_{-i}} \rho_t(s,\vec{a}_{-i} \mid a_i) Q_i^{\rho,t}(s,\vec{a}_{-i},a_i') \tag{21}$$

Positive Result

Theorem [ZGL 2005]
For every NoSDE game, given any natural equilibrium selection mechanism, there exists some $k > 1$ s.t. multiagent value iteration converges to a cyclic equilibrium policy of length k.

Negative Result

Corollary
Multiagent value iteration does not necessarily converge to stationary equilibrium policies in general-sum Markov games, regardless of the equilibrium selection mechanism.

Random Markov Games

$|N| = 2$

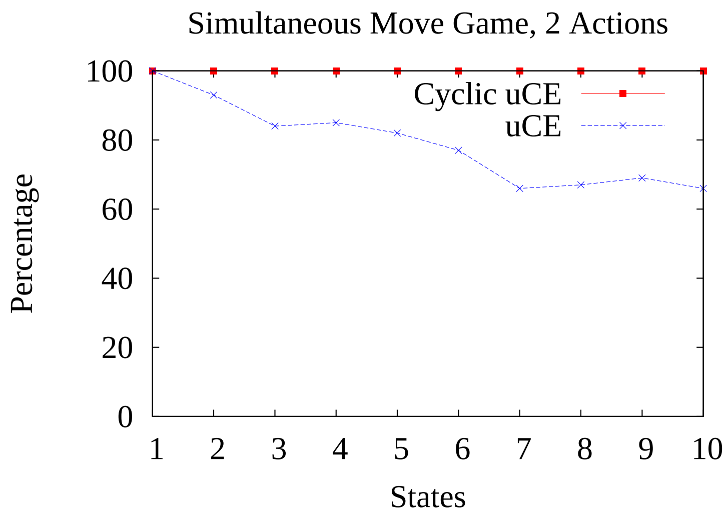
$|A| \in \{2, 3\}$

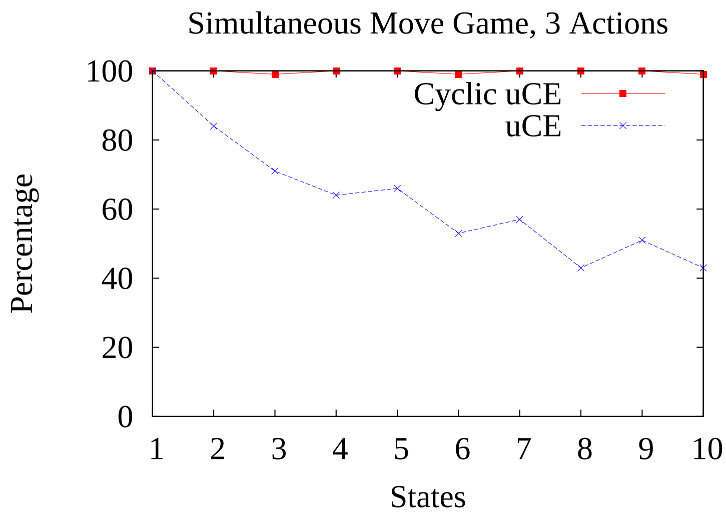
$|S| \in \{1, \ldots, 10\}$

Random Rewards $\in [0, 99]$

Random Deterministic Transitions

$\gamma = \frac{3}{4}$





I. Multiagent Value Iteration in Markov Games

Summary of Observations

- Multiagent value iteration converges empirically to not necessarily deterministic, not necessarily stationary, cyclic equilibrium policies in randomly generated Markov games and Grid Game 1.
  - eCE converges to a nonstationary nondeterministic cyclic equilibrium policy in Grid Game 1.

Open Questions

- Just as multiagent value iteration necessarily converges to stationary equilibrium policies in zero-sum Markov games, does multiagent value iteration necessarily converge to nonstationary cyclic equilibrium policies in general-sum Markov games?

II. No-Regret Learning in Repeated Games

Theorem

No-Φ-regret learning algorithms exist for a natural class of Φs.

Theorem

The empirical distribution of play of no-Φ-regret learning converges to the set of Φ-equilibria in repeated general-sum games.

- No-external-regret learning converges to the set of minimax equilibria in repeated zero-sum games. [e.g., Freund and Schapire 1996]

- No-internal-regret learning converges to the set of correlated equilibria in repeated general-sum games. [Foster and Vohra 1997]

Single Agent Learning Model

- set of actions $N = \{1, \ldots, n\}$
- for all times t,
  - mixed action vector $q^t \in Q = \{q \in \mathbb{R}^n | \sum_i q_i = 1 \ \& \ q_i \geq 0, \forall i\}$
  - pure action vector $a^t = e_i$ for some pure action i
  - reward vector $r^t = (r_1, \ldots, r_n) \in [0, 1]^n$

A learning algorithm $\mathcal{A}$ is a sequence of functions q^t : History$^{t-1} \rightarrow Q$, where a History is a sequence of action-reward pairs $(a^1, r^1), (a^2, r^2), \ldots$.

Transformations

$$\begin{aligned}\Phi_{\text{LINEAR}} &= \{\phi : Q \to Q\} \\ &= \text{the set of all linear transformations} \\ &= \text{the set of all row stochastic matrices}\end{aligned}$$

$\Phi_{\text{EXT}} = \{\phi^j \in \Phi_{\text{LINEAR}} \mid j \in N\}$, where $e_k \phi^j = e_j$

$\Phi_{\text{INT}} = \{\phi^{ij} \in \Phi_{\text{LINEAR}} \mid ij \in N\}$, where $e_k \phi^{ij} = \begin{cases} e_j & \text{if } k = i \\ e_k & \text{otherwise} \end{cases}$

Example

$$\phi^2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad \phi^{23} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$\langle q_1, q_2, q_3, q_4 \rangle \phi^2 = \langle 0, 1, 0, 0 \rangle$, for all $\langle q_1, q_2, q_3, q_4 \rangle \in Q$.

$\langle q_1, q_2, q_3, q_4 \rangle \phi^{23} = \langle q_1, 0, q_2 + q_3, q_4 \rangle$, for all $\langle q_1, q_2, q_3, q_4 \rangle \in Q$.

Regret Matching $(\Phi, g : \mathbb{R}^{\Phi} \rightarrow \mathbb{R}^{\Phi}_{+})$

for $t = 1, \ldots,$

1. play mixed strategy q^t
2. realize pure action a^t
3. observe rewards r^t
4. for all $\phi \in \Phi$
   - compute instantaneous regret
     * observed $\rho^t_\phi \equiv \rho_\phi(r^t, a^t) = r^t \cdot a^t \phi - r^t \cdot a^t$
     * expected $\rho^t_\phi \equiv \rho_\phi(r^t, q^t) = r^t \cdot q^t \phi - r^t \cdot q^t$
   - update cumulative regret vector $X^t_\phi = X^{t-1}_\phi + \rho^t_\phi$
5. compute $Y = g(X^t)$
6. compute $M = \frac{\sum_{\phi \in \Phi} \phi Y_\phi}{\sum_{\phi \in \Phi} Y_\phi}$
7. solve for a fixed point $q^{t+1} = q^{t+1} M$

Regret Matching Theorem

Blackwell's Approachability Theorem: A Generalization

For finite $\Phi \in \Phi_{\text{LINEAR}}$ and for appropriate choices of $g : \mathbb{R} \rightarrow \mathbb{R}_{+}^{\Phi}$, if $\rho(r, q) \cdot g(X) \leq 0$, then the negative orthant $\mathbb{R}_{-}^{\Phi}$ is approachable.

Regret Matching Theorem

For all $\Phi \in \Phi_{\text{LINEAR}}$ and for appropriate choices of g, Regret Matching (Φ, g) satisfies the generalized Blackwell condition: $\rho(r, q) \cdot g(X) \leq 0$.

Corollary

For all $\Phi \in \Phi_{\text{LINEAR}}$ and for appropriate choices of g, Regret Matching (Φ, g) is a no-Φ-regret algorithm.

Special Cases of Regret Matching

Foster and Vohra 1997 (Φ_{INT})

Hart and Mas-Colell 2000 (Φ_{EXT})

Choose $G(X) = \frac{1}{2}\sum_k (X_k^+)^2$ so that $g_k(X) = X_k^+$

Freund and Schapire 1995 (Φ_{EXT})

Cesa-Bianchi and Lugosi 2003 (Φ_{INT})

Choose $G(X) = \frac{1}{\eta}\ln\left(\sum_k e^{\eta X_k}\right)$ so that $g_k(X) = \frac{e^{\eta X_k}}{\sum_k e^{\eta X_k}}$

Multiagent Model

- a set of players N
- for all players i,
  - a set of pure actions A_i
  - a set of mixed actions Q_i
  - a reward function $r_i : A \to [0,1]$, where $A = \prod_i A_i$
  - an expected reward function $r_i : Q \to [0,1]$, where $Q = \Delta(A)$
    $r_i(q) = \sum_{a \in A} q(a) r_i(a)$ for $q \in Q$
  - a set Φ_i

Φ-Equilibrium

Definition

An mixed action profile $q^* \in Q$ is a Φ-equilibrium iff
$r_i(\ddot{\phi}_i(q^*)) \leq r_i(q^*)$, for all players i and for all $\phi_i \in \Phi_i$.

Examples

Correlated Equilibrium: $\Phi_i = \Phi_{\text{INT}}$, for all players i

Generalized Minimax Equilibrium: $\Phi_i = \Phi_{\text{EXT}}$, for all players i

Theorem

The empirical distribution of play of no-Φ-regret learning converges to the set of Φ-equilibria in repeated general-sum games.

Zero-Sum Games

Matching Pennies

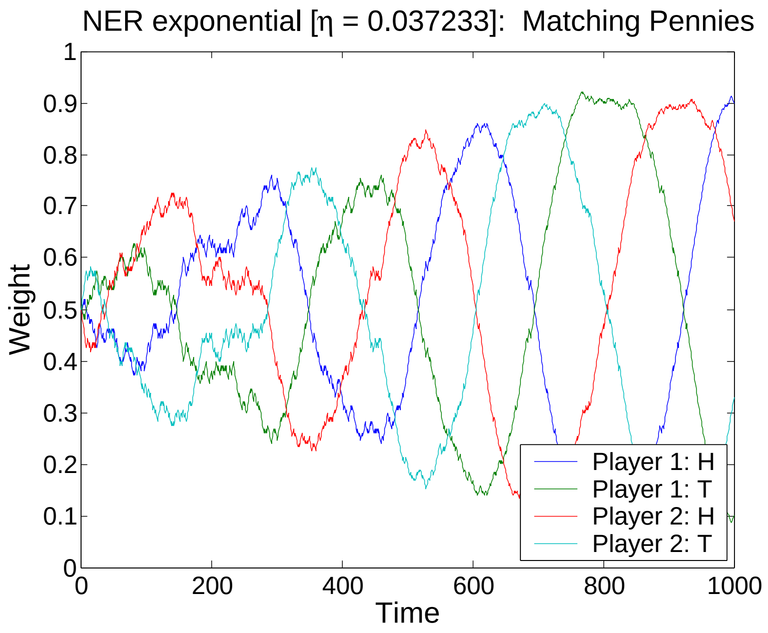
	h	t
H	$-1, 1$	$1, -1$
T	$1, -1$	$-1, 1$

Rock-Paper-Scissors

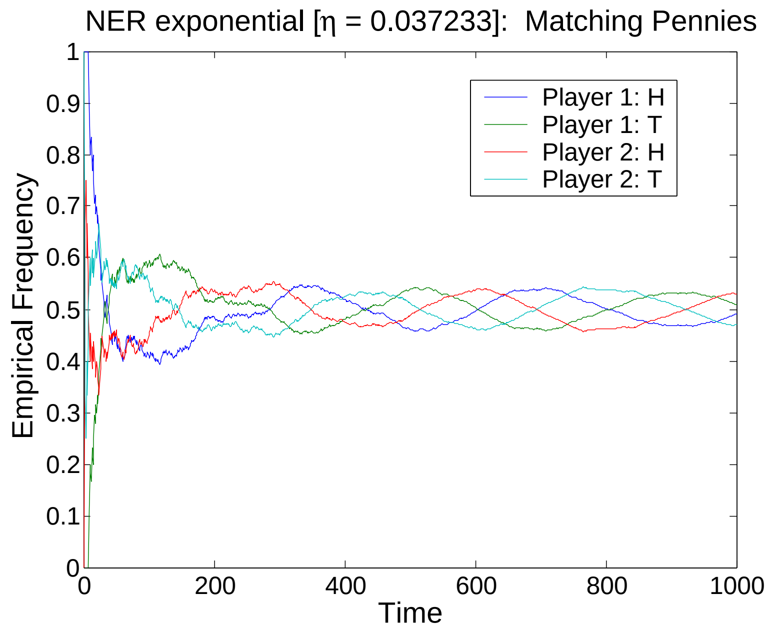
	r	p	s
R	$0, 0$	$-1, 1$	$1, -1$
P	$1, -1$	$0, 0$	$-1, 1$
S	$-1, 1$	$1, -1$	$0, 0$

Matching Pennies

Weights

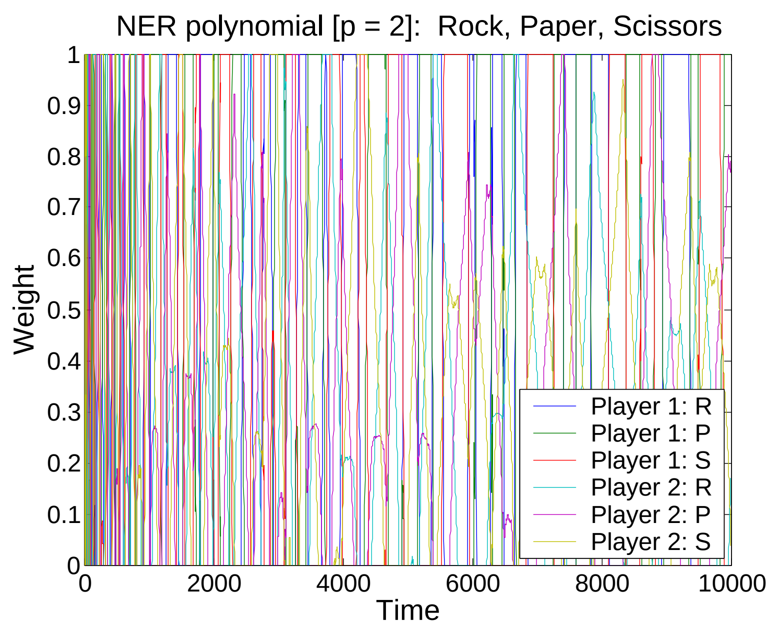


Frequencies

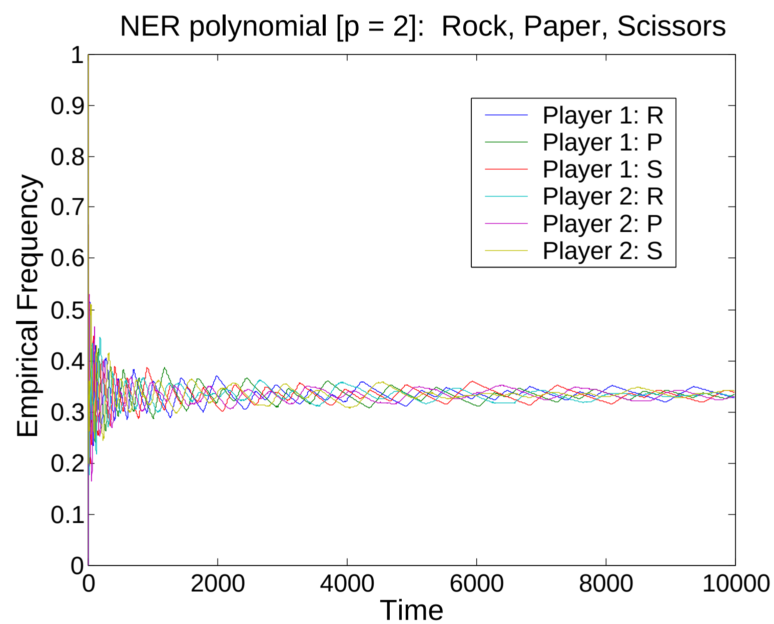


Rock-Paper-Scissors

Weights



Frequencies



General-Sum Games

Shapley Game

	l	c	r
T	$0, 0$	$1, 0$	$0, 1$
M	$0, 1$	$0, 0$	$1, 0$
B	$1, 0$	$0, 1$	$0, 0$

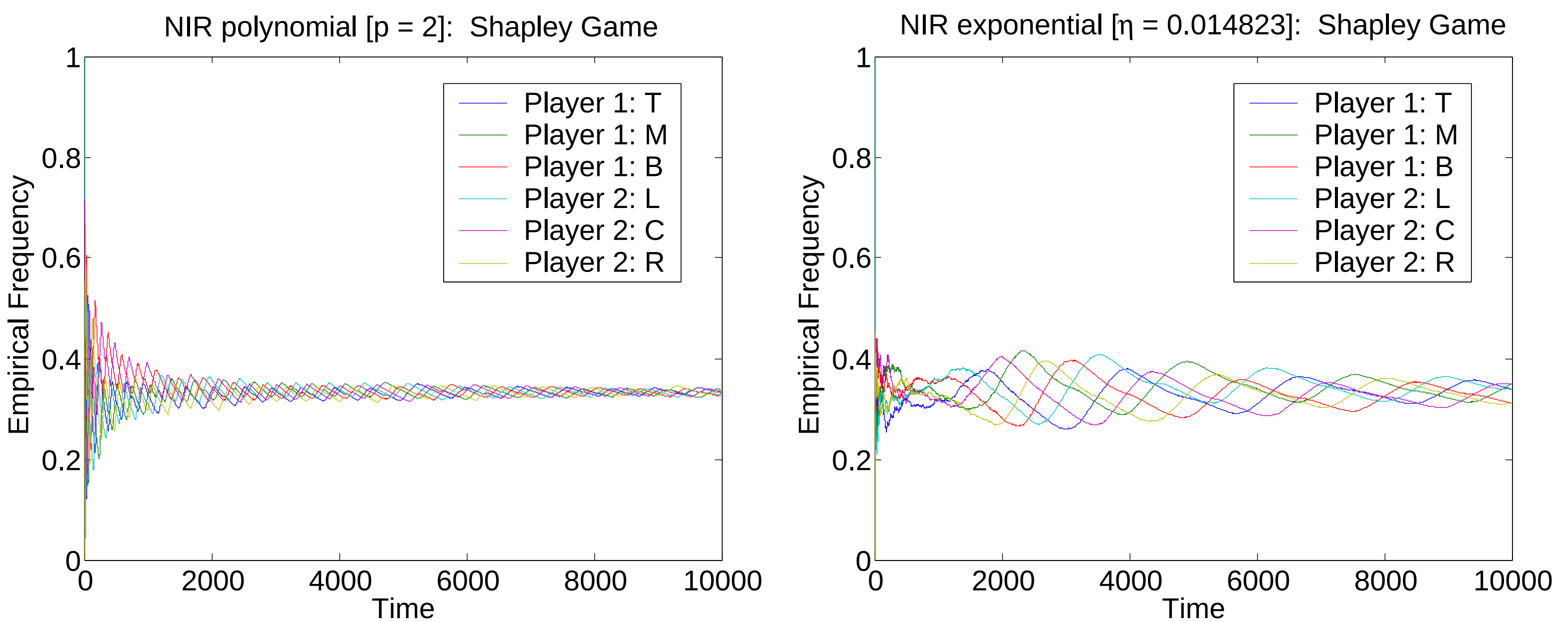
Correlated Equilibrium

	l	c	r
T	0	$1/6$	$1/6$
M	$1/6$	0	$1/6$
B	$1/6$	$1/6$	0

	l	c	r
T	2ϵ	$1/6 - \epsilon$	$1/6 - \epsilon$
M	$1/6 - \epsilon$	2ϵ	$1/6 - \epsilon$
B	$1/6 - \epsilon$	$1/6 - \epsilon$	2ϵ

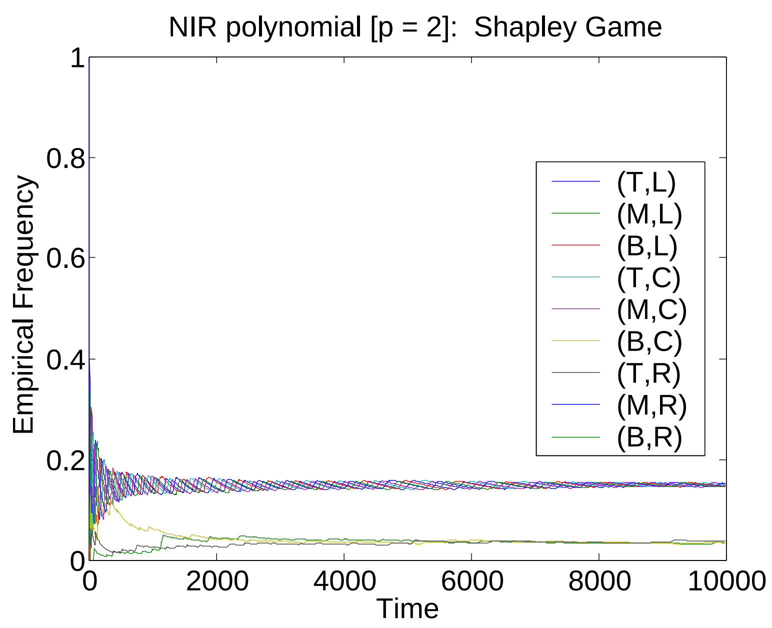
Shapley Game: No Internal Regret Learning

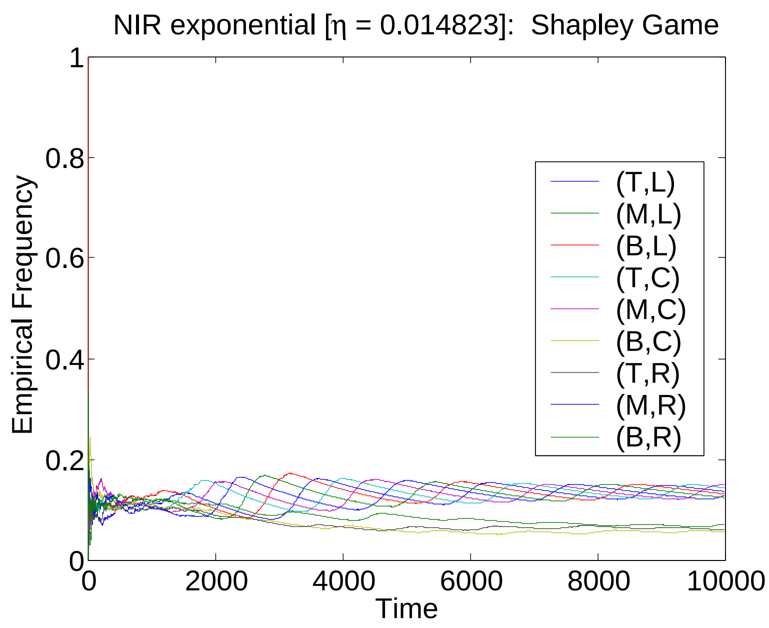
Frequencies



Shapley Game: No Internal Regret Learning

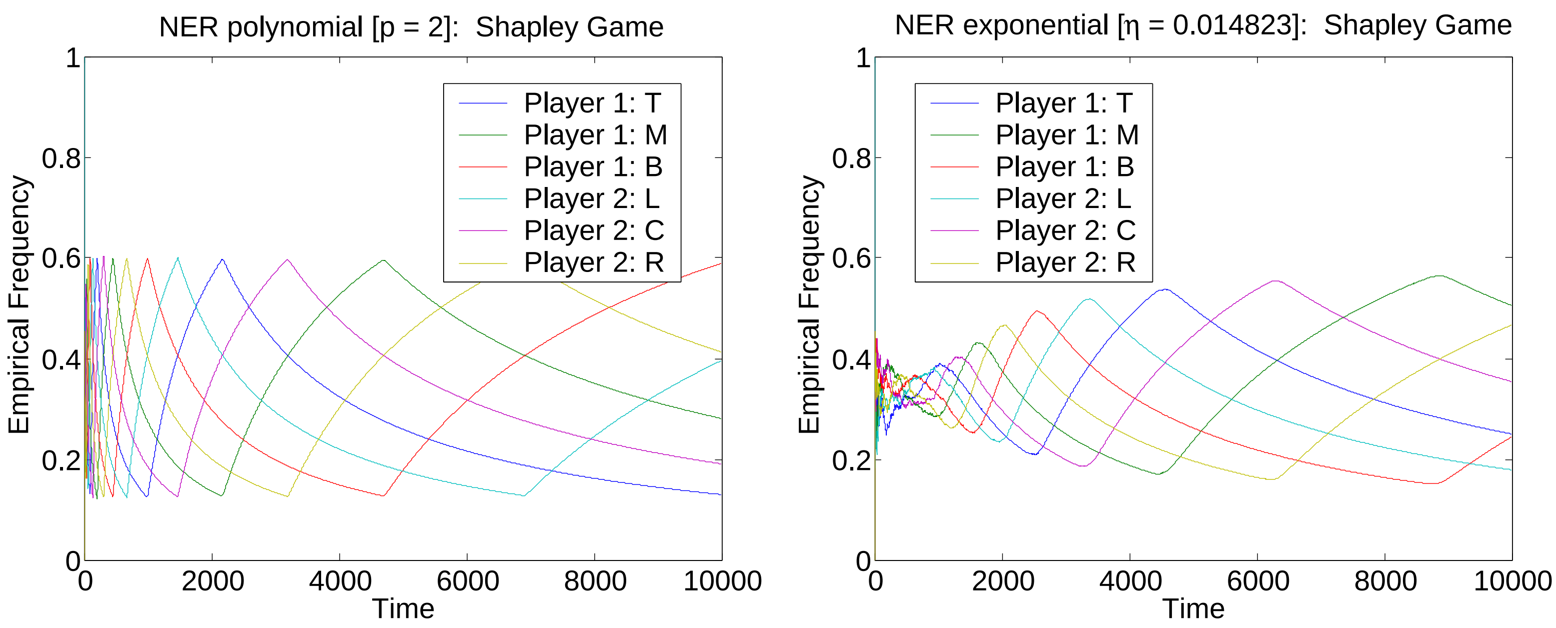
Joint Frequencies





Shapley Game: No External Regret Learning

Frequencies



II. No-Regret Learning in Repeated Games

Summary of Observations

- No-Φ-regret learning algorithms exist for a natural class of Φs.
- The empirical distribution of play of no-Φ-regret learning converges to the set of Φ-equilibria in repeated general-sum games.

Open Questions

- Equilibrium selection problem: QWERTY Game

	d	q
D	5,5	0,0
Q	0,0	4,4

III. Stochastic Stability

Definition

Given a Markov matrix M (i.e., $M \geq 0$ and $JM = J$), a perturbed Markov process M_ϵ is a family of Markov matrices with entries $M_{ij} = \epsilon^{r_{ij}} c_{ij}(\epsilon)$.

Theorem

Given $\epsilon > 0$, the Markov matrix M_ϵ has a unique stable distribution, call it v_ϵ.

Definition

The limit of the sequence $\{v_\epsilon\}$, as $\epsilon \to 0$, exists, is unique, and is called the stochastically stable distribution of the perturbed Markov process.

Algorithm [WG 2005]

An exact algorithm to compute the stochastically stable distribution of a perturbed Markov process.

Adaptive Learning in Repeated Games

Model [Young 1993]

- A variant of Fictitious Play [Brown 1951]
- Finite memory m, Sample size s
- Play a best-response

QWERTY: $m = s = 1$

M_0	Dd	Qd	Dq	Qq
Dd	1	0	0	0
Qd	0	0	1	0
Dq	0	1	0	0
Qq	0	0	0	1

Adaptive Learning in Repeated Games

Model [Young 1993]

- A variant of Fictitious Play [Brown 1951]
- Finite memory m, Sample size s
- Mistake probability ϵ
  - Play arbitrarily with probability ϵ
  - Play a best-response with probability $1 - \epsilon$

QWERTY: $m = s = 1$

M_ϵ	Dd	Qd	Dq	Qq
Dd	$(1-\epsilon)(1-\epsilon)$	$(1-\epsilon)\epsilon$	$\epsilon(1-\epsilon)$	ϵ^2
Qd	$\epsilon(1-\epsilon)$	ϵ^2	$(1-\epsilon)(1-\epsilon)$	$(1-\epsilon)\epsilon$
Dq	$(1-\epsilon)\epsilon$	$(1-\epsilon)(1-\epsilon)$	ϵ^2	$\epsilon(1-\epsilon)$
Qq	ϵ^2	$\epsilon(1-\epsilon)$	$(1-\epsilon)\epsilon$	$(1-\epsilon)(1-\epsilon)$

Equilibrium Selection

QWERTY′

	d	q
D	5,5	0,3
Q	3,0	4,4

m	s	Equilibrium
2	2	Qq
3	2	Qq
3	3	Qq
4	2	Qq
4	3	Qq
4	4	Qq

In QWERTY′, Qq is the risk-dominant equilibrium.

Equilibrium Selection

QWERTY′

	d	q
D	5,5	0,3
Q	3,0	4,4

m	s	Equilibrium
2	2	Qq
3	2	Qq
3	3	Qq
4	2	Qq
4	3	Qq
4	4	Qq

Coordination Game

	l	c	r
T	3,3	0,0	0,0
M	0,0	2,2	0,0
B	0,0	0,0	1,1

In QWERTY′, Qq is the risk-dominant equilibrium.

III. Adaptive Learning in Repeated Games

Summary of Observations

- The theory of stochastic stability can be applied to predict the dynamics of adaptive learning in repeated games.

Open Questions

- Can this theory be applied to predict the dynamics of no-regret learning in repeated games or multiagent Q-learning in Markov games?

Summary

What is the outcome of multiagent learning in games?

- Multiagent value iteration in Markov games $\longrightarrow$ cyclic equilibria.
- No-Φ-regret learning in repeated games $\longrightarrow$ the set of Φ-equilibria.
- Adaptive learning in repeated games selects risk-dominant equilibria.

References

ZGL Martin Zinkevich, Amy Greenwald, and Michael Littman. "Cyclic Equilibria in Markov Games." *2005 Proceedings of the Neural Information Processing Systems Conference*.

GJ Amy Greenwald and Amir Jafari. "A General Class of No-Regret Learning Algorithms and Game-Theoretic Equilibria." *2003 Proceedings of the Computational Learning Theory Conference.*

WG John Wicks and Amy Greenwald. "An Algorithm for Computing Stochastically Stable Distributions with Applications to Multiagent Learning in Repeated Games." *2005 Proceedings of the Uncertainty in Artificial Intelligence Conference.*

GHZ Amy Greenwald, Keith Hall, and Martin Zinkevich. "Correlated Q-Learning." *Brown University Technical Report CS–05–08.* Earlier version: *2003 Proceedings of the International Conference on Machine Learning*.