Stochastic Stability of Perturbed Learning Automata in Positive Utility Games

Georgios C. Chasparis

Department of Data Analysis Systems
Software Competence Center Hagenberg GmbH, Austria
(Johannes Kepler University, Linz, Austria)

LEG'2019

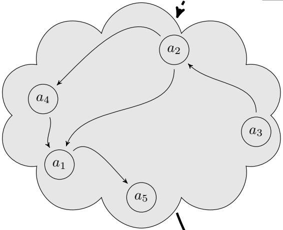
Tel Aviv, Israel
June 25th, 2019

Centralized vs Decentralized Optimization

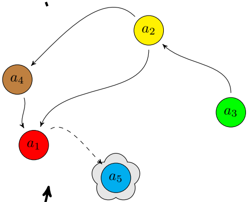
Optimization Uncertainty

Challenges of decentralized opt.:
- (C1) Dynamic optimization
- (C2) Measurement-based opt.
- (C3) Efficiency and robustness
- (C4) Fast response

Centralized optimization



Multi-agent optimization

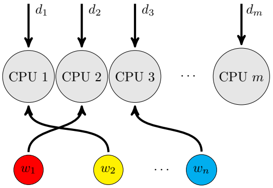


Issues of centralized optimization:
- (I1) Unknown components' details
- (I2) Unknown disturbances
- (I3) Unknown interdependencies
- (I4) Limited computational resources

Real-time adaptivity & robustness

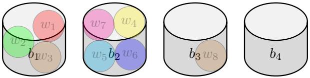
Example: *Resource-Aware Applications*

- *Challenges*
  - Unknown objective function
  - Unknown disturbances
- *Instead:*
  - *Distributed sensing/actuation*
  - *Measurement-based opt.*
- *New challenges:*
  - Optimization uncertainty
  - Adaptivity
  - Noisy measurements
  - Convergence speed

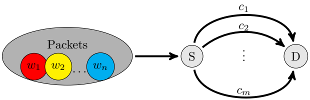


Other relevant examples

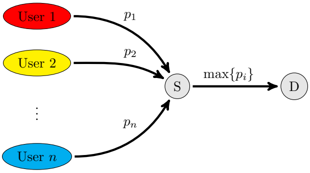
- *Bin-packing*



- *Routing*

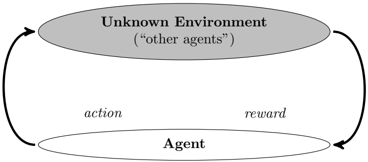


- *Channel access*



Approach

- *Main elements*
  - Payoff-based learning
  - Large (coordination) games
  - Convergence guarantees
- *Specifically, this work is about*
  - Reinforcement learning
  - Convergence guarantees in large games
  - Specialization to coordination games

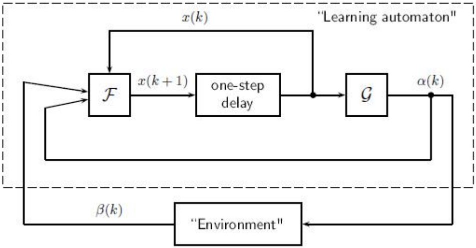


Outline

Outline

Learning Automata

- **Learning Automata:**
  - Agents revise their decisions *repeatedly*
  - Information is only *local*
    - Agents observe only their own utility
  - Agents reinforce an action through
    - repeated selection
    - reward size
  - Introduced/analyzed first by Tsetlin (1973)



Strategic-form Games: Basic Notation/Terminology

- Each agent i has a finite set of *actions* $\mathcal{A}_i$
- Each agent i select actions based on *strategy*

$$\sigma_i \triangleq \begin{pmatrix} \sigma_{i1} \\ \vdots \\ \sigma_{i|\mathcal{A}_i|} \end{pmatrix} \in \Delta\left(|\mathcal{A}_i|\right)$$

- Each agent i receives a *utility* (or *payoff*),

$$u_i : \mathcal{A} \to \mathbb{R}_+$$

Strategic-form Games: Basic Notation/Terminology

- Each agent i has a finite set of *actions* $\mathcal{A}_i$
- Each agent i select actions based on *strategy*

$$\sigma_i \triangleq \begin{pmatrix} \sigma_{i1} \\ \vdots \\ \sigma_{i|\mathcal{A}_i|} \end{pmatrix} \in \Delta\left(|\mathcal{A}_i|\right)$$

- Each agent i receives a *utility* (or *payoff*),

$$u_i : \mathcal{A} \to \mathbb{R}_+$$

- Example:
  - 2 players, 2 actions
  - strategy: e.g., $\sigma_i = (0.2, 0.8)$
  - utility: e.g., $u_i(A, A) = 2$.

	A	B
A	$2, 2$	$0, 0$
B	$0, 0$	$1, 1$

(Variable structure) Learning Automata

At each time period $k = 0, 1, 2, ...$, each agent i

1. **Action update:** Randomize using strategy $\sigma_i(k) = x_i(k)$,

$$\alpha_i(k) = \mathrm{rand}_{\sigma_i}[\mathcal{A}_i]$$

2. **Performance Observation:**

$$u_i = u_i(\alpha(k))$$

3. **Strategy update:**

$$x_i(k+1) = x_i(k) + \epsilon(k) \cdot u_i(\alpha(k)) \cdot (e_{\alpha_i(k)} - x_i(k))$$

(Variable structure) Learning Automata

At some time k, agent i

1. **Action update:** Selects $\alpha_i(k) = A$ based on strategy

$$x_i(k) = \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix}$$

2. **Performance Observation:**

$$u_i = u_i(A,A)=2$$

3. **Strategy update:**

$$\begin{pmatrix} 0.2 + 1.6\epsilon \\ 0.8 - 1.6\epsilon \end{pmatrix} \leftarrow \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix} + \epsilon \cdot 2 \cdot \left[\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix} \right]$$

Example:

	A	B
A	2, 2	0, 0
B	0, 0	1, 1

(Variable structure) Learning Automata

At some time k, agent i

1. **Action update:** Selects $\alpha_i(k) = A$ based on strategy

$$x_i(k) = \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix}$$

2. **Performance Observation:**

$$u_i = u_i(A,A) = 2$$

3. **Strategy update:**

$$\begin{pmatrix} 0.2 + 1.6\epsilon \\ 0.8 - 1.6\epsilon \end{pmatrix} \leftarrow \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix} + \epsilon \cdot 2 \cdot \left[\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix} \right]$$

Note:

- $x_i(k)$ increases *in the direction of* the selected action
- $x_i(k)$ increases *proportionally to* the observed performance

Prior Schemes: Reinforcement-Learning

Action update:

$$\alpha_i(t) = \text{rand}_{\sigma_i(k)}[\mathcal{A}_i], \quad \sigma_i(k) = x_i(k)$$

Strategy update:

$$x_i(k+1) = x_i(k) + \epsilon_i(k) \cdot u_i(\alpha(k)) \cdot \left[e_{\alpha_i(k)} - x_i(k)\right]$$

- *Arthur (1993), Posch (1997) models:*

$$\epsilon_i(k) \triangleq \frac{1}{ck^{\nu} + u_i(\alpha(k))}$$

  - Excluding convergence to non-Nash equilibria.

Prior Schemes: Reinforcement-Learning

Action update:

$$\alpha_i(t) = \text{rand}_{\sigma_i(k)}[\mathcal{A}_i], \quad \sigma_i(k) = x_i(k)$$

Strategy update:

$$x_i(k+1) = x_i(k) + \epsilon_i(k) \cdot u_i(\alpha(k)) \cdot \left[e_{\alpha_i(k)} - x_i(k)\right]$$

- *Arthur (1993), Posch (1997) models:*

$$\epsilon_i(k) \triangleq \frac{1}{ck^{\nu} + u_i(\alpha(k))}$$

  – Excluding convergence to non-Nash equilibria.

- *Urn Process:* [Hopkins & Posch (2005), Erev & Roth (1998)]

$$\epsilon_i(k) \triangleq \frac{1}{V_i(k) + u_i(\alpha(k))}$$

  + Excluding convergence to non-Nash equilibria.
  – Convergence to Nash equilibria only in 2-player partnership games

Prior Schemes: Learning automata

Action update:

$$\alpha_i(t) = \text{rand}_{\sigma_i(k)}[\mathcal{A}_i], \quad \sigma_i(k) = x_i(k)$$

Strategy update:

$$x_i(k+1) = x_i(k) + \epsilon_i(k) \cdot u_i(\alpha(k)) \cdot \left[e_{\alpha_i(k)} - x_i(k)\right]$$

- *Narendra & Thathachar (1989):*

$$u_i(\alpha(k)) \in [0, 1]$$

  - Convergence to Nash equilibria only in *identical interest games*
  - Extension to large games requires an *absolute monotonocity* condition.

Prior Schemes: Learning automata

Action update:

$$\alpha_i(t) = \text{rand}_{\sigma_i(k)}[\mathcal{A}_i], \quad \sigma_i(k) = x_i(k)$$

Strategy update:

$$x_i(k+1) = x_i(k) + \epsilon_i(k) \cdot u_i(\alpha(k)) \cdot \left[e_{\alpha_i(k)} - x_i(k)\right]$$

- *Narendra & Thathachar (1989):*

$$u_i(\alpha(k)) \in [0, 1]$$

  – Convergence to Nash equilibria only in *identical interest games*
  – Extension to large games requires an *absolute monotonocity* condition.

- *Verbeeck et al (2007):*
  – Introduced a *coordinated exploration phase*
  + Convergence to efficient Nash equilibria

Prior Schemes: Perturbed Learning automata

Action update:

$$\alpha_i(t) = \text{rand}_{\sigma_i(k)}[\mathcal{A}_i], \quad \sigma_i(k) = (1-\lambda)x_i(k) + \lambda\mathbf{1}/n$$

Strategy update:

$$x_i(k+1) = x_i(k) + \epsilon_i(k) \cdot u_i(\alpha(k)) \cdot \left[e_{\alpha_i(k)} - x_i(k)\right]$$

- *Chasparis, Shamma & Rantzer (2014)*

$$\sigma_i(k) = (1-\lambda)x_i(k) + \lambda\mathbf{1}/n$$

  - \+ excludes convergence to non-Nash equilibria
  - \+ guarantees global convergence to pure Nash equilibria in potential games
  - − global convergence in generic coordination games is not shown

Why learning automata?

	A	B
A	$2, 2$	$0, 0$
B	$0, 0$	$1, 1$

- *equilibrium-selection* mechanism
  - We can get convergence to desirable outcomes
  - Modified selection rules may be required
- *measurement-based* dynamics
  - Agents only observe performance measurements
- "handles" *noisy observations*
  - noise is filtered out through the strategy-vector formulation
  - demonstrated in the analysis of Hopkins and Posch (2005)

Issues?

	A	B
A	$2, 2$	$0, 0$
B	$0, 0$	$1, 1$

- *Issues*
  - global convergence to efficient outcomes is difficult to show.
  - excluding convergence to mixed strategies.
  - Lyapunov-based techniques are not appropriate for large games

Issues?

	A	B
A	$2, 2$	$0, 0$
B	$0, 0$	$1, 1$

- *Issues*
  - global convergence to efficient outcomes is difficult to show.
  - excluding convergence to mixed strategies.
  - Lyapunov-based techniques are not appropriate for large games

- *Contributions*
  - a *stochastic stability* analysis for perturbed learning automata
  - global convergence guarantees (circumvents issues of Lyapunov-based analysis)
  - specialization to coordination games

Outline

Stochastic Stability for constant step-size

Strategy Update:

$$x_i(k+1) = x_i(k) + \epsilon \cdot u_i(\alpha(k)) \cdot \left[e_{\alpha_i(k)} - x_i(k)\right]$$

Action selection:

$$\sigma_i(k) = (1-\lambda)x_i(k) + \lambda\mathbf{1}/n$$

Note:

- Defines an induced Markov chain in:

$$\mathcal{Z} \doteq \mathcal{A} \times \Delta(n)$$

- Infinite dimensional with t.p.f. P_λ

Assumption: $u_i(\alpha) > 0$ for all i and $\alpha \in \mathcal{A}$.

Stochastic Stability for constant step-size

Strategy Update:

$$x_i(k+1) = x_i(k) + \epsilon \cdot u_i(\alpha(k)) \cdot \left[e_{\alpha_i(k)} - x_i(k)\right]$$

Action selection:

$$\sigma_i(k) = (1-\lambda)x_i(k) + \lambda\mathbf{1}/n$$

Proposition

For $\lambda = 0$, the probability that eventually agents play the same action profile is 1

Stochastic Stability for constant step-size

Strategy Update:

$$x_i(k+1) = x_i(k) + \epsilon \cdot u_i(\alpha(k)) \cdot \left[e_{\alpha_i(k)} - x_i(k)\right]$$

Action selection:

$$\sigma_i(k) = (1-\lambda)x_i(k) + \lambda \mathbf{1}/n$$

Proposition

For $\lambda = 0$, the probability that eventually agents play the same action profile is 1

Remark

Reduce infinite dimensional P_λ to finite dimensional π (isomorphic with $\mathcal{A}$).

Stochastic Stability for constant step-size

Strategy Update:

$$x_i(k+1) = x_i(k) + \epsilon \cdot u_i(\alpha(k)) \cdot \left[e_{\alpha_i(k)} - x_i(k)\right]$$

Action selection:

$$\sigma_i(k) = (1-\lambda)x_i(k) + \lambda \mathbf{1}/n$$

Theorem

There exists a unique probability vector π such that:

1. $\mu_\lambda \Rightarrow \sum_{\alpha \in \mathcal{A}} \pi_\alpha \delta_\alpha(\cdot)$ *as* $\lambda \downarrow 0$,
2. π *is an invariant distribution of the (finite-state) Markov chain* $\hat{P}$

$$\hat{P}_{\alpha\alpha'} \doteq \lim_{t\to\infty} QP^t(\alpha, \mathcal{N}_\varepsilon(\alpha')),$$

for any $\varepsilon > 0$, *where* Q *is the t.p.f. of one player trembling.*

Stochastic Stability for constant step-size

Strategy Update:

$$x_i(k+1) = x_i(k) + \epsilon \cdot u_i(\alpha(k)) \cdot \left[e_{\alpha_i(k)} - x_i(k)\right]$$

Action selection:

$$\sigma_i(k) = (1-\lambda)x_i(k) + \lambda \mathbf{1}/n$$

Theorem

There exists a unique probability vector π such that:

1. $\mu_\lambda \Rightarrow \sum_{\alpha\in\mathcal{A}} \pi_\alpha \delta_\alpha(\cdot)$ *as* $\lambda \downarrow 0$,
2. π *is an invariant distribution of the (finite-state) Markov chain* $\hat{P}$

$$\hat{P}_{\alpha\alpha'} \doteq \lim_{t\to\infty} QP^t(\alpha, \mathcal{N}_\varepsilon(\alpha')),$$

for any $\varepsilon > 0$, *where* Q *is the t.p.f. of one player trembling.*

Infinite dimensional $\Rightarrow$ *Finite dimensional Markov chain*

δ-resistance

Lemma

For sufficiently small step-size $\epsilon > 0$, the one-step transition probabilities (of the finite approximation) satisfy:

$$\hat{P}_{\alpha\alpha'} \approx \gamma \lim_{\delta \downarrow 0} \exp\left(\frac{\eta(\delta)}{\epsilon u_j(\alpha')}\right)$$

for some negative constant $\eta(\delta)$.

δ-resistance

Lemma

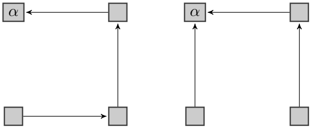
For sufficiently small step-size $\epsilon > 0$, the one-step transition probabilities (of the finite approximation) satisfy:

$$\hat{P}_{\alpha\alpha'} \approx \gamma \lim_{\delta \downarrow 0} \exp\left(\frac{\eta(\delta)}{\epsilon u_j(\alpha')}\right)$$

for some negative constant $\eta(\delta)$.

δ-resistance:

$$\varphi_\delta(\alpha|g) \doteq \sum_{(\alpha^{(k)} \to \alpha^{(\ell)})} \frac{1}{\epsilon u_j(\alpha^{(\ell)})}$$



($\mathcal{W}$-graphs)

δ-resistance

Lemma

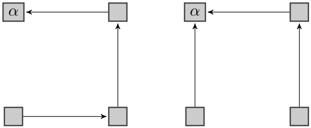
For sufficiently small step-size $\epsilon > 0$, the one-step transition probabilities (of the finite approximation) satisfy:

$$\hat{P}_{\alpha\alpha'} \approx \gamma \lim_{\delta \downarrow 0} \exp\left(\frac{\eta(\delta)}{\epsilon u_j(\alpha')}\right)$$

for some negative constant $\eta(\delta)$.

δ-resistance:

$$\varphi_\delta(\alpha|g) \doteq \sum_{(\alpha^{(k)} \to \alpha^{(\ell)})} \frac{1}{\epsilon u_j(\alpha^{(\ell)})}$$



($\mathcal{W}$-graphs)

Theorem

As $\epsilon \downarrow 0$, the set of stochastically stable action profiles $\mathcal{A}^*$ is such that, for any $\delta > 0$,

$$\max_{\alpha^* \in \mathcal{A}^*} \varphi_\delta^*(\alpha^*) < \min_{\alpha \in \mathcal{A} \backslash \mathcal{A}^*} \varphi_\delta^*(\alpha)$$

where ϕ_δ^* denotes minimum resistance over all g.

Outline

Specialization to Large Coordination Games

Definition (Coordination games)

A strategic-form game satisfying the positive-utility property is a coordination game if, for every action profile α and player i, $u_j(\alpha_i', \alpha_{-i}) \geq u_j(\alpha_i, \alpha_{-i})$ for any $\alpha_i' \in \mathrm{BR}_i(\alpha)$.

Specialization to Large Coordination Games

Definition (Coordination games)

A strategic-form game satisfying the positive-utility property is a coordination game if, for every action profile α and player i, $u_j(\alpha_i', \alpha_{-i}) \geq u_j(\alpha_i, \alpha_{-i})$ for any $\alpha_i' \in \mathrm{BR}_i(\alpha)$.

Theorem

In any coordination game, as $\epsilon \downarrow 0$ and $\lambda \downarrow 0$,

$$\mathcal{S}^* \subseteq \mathcal{S}_{\mathrm{NE}}$$

Specialization to Large Coordination Games

Definition (Coordination games)

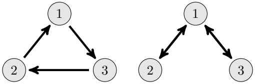
A strategic-form game satisfying the positive-utility property is a coordination game if, for every action profile α and player i, $u_j(\alpha_i', \alpha_{-i}) \geq u_j(\alpha_i, \alpha_{-i})$ for any $\alpha_i' \in \mathrm{BR}_i(\alpha)$.

Theorem

In any coordination game, as $\epsilon \downarrow 0$ and $\lambda \downarrow 0$,

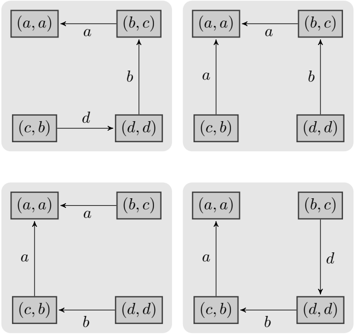
$$\mathcal{S}^* \subseteq \mathcal{S}_{\mathrm{NE}}$$

- **Example:** *Network Formation Games.*



Specialization to 2×2 Coordination Games

	A	B
A	a, a	b, c
B	c, b	d, d



One-step $s_{(A,A)}$-graphs and payoff change.

Procedure

1. Compute resistances of s-graphs
2. Compare minimum resistances

Specialization to 2×2 Coordination Games (cont.)

	A	B
A	a, a	b, c
B	c, b	d, d

Proposition

Consider the 2-player, 2-action game of with $a > c > 0$, $d > b > 0$, and $a > d$. Denote $s_{(A,A)}$ and $s_{(B,B)}$ as the p.s.s.'s corresponding to action profiles (A, A) and (B, B), respectively. The following hold:

(a) *if $a - c < d - b$, then*

$$\lim_{\epsilon \downarrow 0} \lim_{\lambda \downarrow 0} \pi_{s_{(B,B)}} = 1,$$

i.e., (B, B) corresponds to the unique stochastically stable state;

(b) *if $a - c \geq d - b$ and $c \leq b$, then*

$$\lim_{\epsilon \downarrow 0} \lim_{\lambda \downarrow 0} \pi_{s_{(A,A)}} = 1,$$

i.e., (A, A) corresponds to the unique stochastically stable state.

Outline

Contribution Snapshot

Features/Conditions	***Strong* Convergence in Strategic-Form Games**		
	Reinforcement-based learning	Q-learning	*Aspiration-based learning*
(Structural) Assumptions:			
2 players	✓	✓	✓
>2 players	✓	○	✓
Potential games	✓	✓	✓
Coordination games	✓	○	✓
Weakly-acyclic games	○	○	✓
Convergence to:			
Nash equilibria	✓	✓	✓
(Pareto) Efficient Nash equil.	○	○	✓
(Pareto) Efficient outcomes	○	○	✓
Additional features:			
Noisy observations	✓	✓	○
Constant step-size	✓	○	✓

- **Aspiration-based learning:**
  - Benchmark-based learning (Marden, Young, Arslan, Shamma, 2009)
  - Trial-and-error learning (Young, 2011)
  - Mood-based learning (Marden, Young, Pao, 2014)
  - Average Testing (Arieli, Babichenko, 2011)