Reinforcement learning with restrictions on the action set

Mario Bravo
Universidad de Chile

Joint work with Mathieu Faure (AMSE-GREQAM)

Outline

Motivation

- Most debated and studied learning procedure in game theory : Fictitious play [Brown51]

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

- Consider an N-player normal form game which is repeated in discrete time. At each time, players compute a *best response* to the opponent's empirical average play.
  The idea is to study the asymptotic behavior of the empirical frequency of play of player i, v_n^i.

Motivation

Large body of literature devoted to the question of identifying classes of games where the empirical frequencies of play converge to the set of Nash equilibria of the underlying game.

- Zero-sum games [Robinson 51]
- General (non-degenerate) 2×2 [Miyasawa 61]
- Potential games [Monderer and Shapley 96]

Motivation

Large body of literature devoted to the question of identifying classes of games where the empirical frequencies of play converge to the set of Nash equilibria of the underlying game.

- Zero-sum games [Robinson 51]
- General (non-degenerate) 2×2 [Miyasawa 61]
- Potential games [Monderer and Shapley 96]

Recall that

A game $\mathcal{G} = (N, (S^i)_{i \in N}, (G^i)_{i \in N})$ is a potential game if it exists a function $\Phi : \Pi_{k=1}^{N} S^k \to \mathbb{R}$ such that

$$G^i(s^i, s^{-i}) - G^i(r^i, s^{-i}) = \Phi(s^i, s^{-i}) - \Phi(r^i, s^{-i}),$$

for all $s^i, r^i \in S^i$ and $s^{-i} \in S^{-i}$.
Primary example : Congestion games [Rosenthal 73]

Motivation

Large body of literature devoted to the question of identifying classes of games where the empirical frequencies of play converge to the set of Nash equilibria of the underlying game.

- Zero-sum games [Robinson 51]
- General (non-degenerate) 2×2 [Miyasawa 61]
- Potential games [Monderer and Shapley 96]

- 2-player games where one of the players has only two actions [Berger 05]
- New proofs and generalizations using stochastic approximation techniques [Benaim et al 05, Hofbauer and Sorin 06]
- Several variations and applications in multiple domains (transportation, telecomunications, etc)

Problem

Players need a lot of information !

Problem

Players need a lot of information !

Three main assumptions are made here :

(i) Each player knows the structure of the game, i.e. she knows her own payoff function, so she can compute a best response.

Problem

Players need a lot of information !

Three main assumptions are made here :

(i) Each player knows the structure of the game, i.e. she knows her own payoff function, so she can compute a best response.

(ii) Each player is informed of the action selected by her opponents at each stage ; thus she can compute the empirical frequencies

Problem

Players need a lot of information !

Three main assumptions are made here :

(i) Each player knows the structure of the game, i.e. she knows her own payoff function, so she can compute a best response.

(ii) Each player is informed of the action selected by her opponents at each stage ; thus she can compute the empirical frequencies

(iii) Each player is allowed to choose any action at each time, so that she can actually play a best response.

Dropping (i) and (ii)

- One approach (among many others) is to assume that the agents observe only their realized payoff at each stage.
- Payoff function are unkown
- This is the minimal information framework of the so-called *reinforcement learning* procedures [Borgers and Sarin 97, Erev and Roth 98]

Most work in this direction proceeds as follows :

a) construct a sequence of mixed strategies which are updated taking into account the payoff they receive (which is the only information agents have access to)

b) Study the convergence (or non-convergence) of this sequence.

Dropping (i) and (ii)

- One approach (among many others) is to assume that the agents observe only their realized payoff at each stage.
- Payoff function are unkown
- This is the minimal information framework of the so-called *reinforcement learning* procedures [Borgers and Sarin 97, Erev and Roth 98]

	A	B	C	D	E
R	?	?	?	?	?
S	?	?	?	?	?
P	?	?	?	?	?

Actions played :
Payoff received :

Actions played
Payoff received :

Dropping (i) and (ii)

- One approach (among many others) is to assume that the agents observe only their realized payoff at each stage.
- Payoff function are unkown
- This is the minimal information framework of the so-called *reinforcement learning* procedures [Borgers and Sarin 97, Erev and Roth 98]

	A	B	C	D	E
R	?	?	?	1	?
S	?	?	?	?	?
P	?	?	?	?	?

Actions played : R
Payoff received :**1**

Actions played : D
Payoff received :-**1**

Dropping (i) and (ii)

- One approach (among many others) is to assume that the agents observe only their realized payoff at each stage.
- Payoff function are unkown
- This is the minimal information framework of the so-called *reinforcement learning* procedures [Borgers and Sarin 97, Erev and Roth 98]

	A	B	C	D	E
R	?	?	?	1	?
S	?	?	-1	?	?
P	?	?	?	?	?

Actions played : R, S
Payoff received :**1, -1**

Actions played : D, C
Payoff received :**-1, 1**

Dropping (i) and (ii)

- One approach (among many others) is to assume that the agents observe only their realized payoff at each stage.
- Payoff function are unkown
- This is the minimal information framework of the so-called *reinforcement learning* procedures [Borgers and Sarin 97, Erev and Roth 98]

	A	B	C	D	E
R	?	?	?	1	?
S	?	2	-1	?	?
P	?	?	?	?	?

Actions played : R, S, S
Payoff received :**1, -1, 2**

Actions played : D, C, B
Payoff received :**-1, 1, -2**

Dropping (i) and (ii)

- One approach (among many others) is to assume that the agents observe only their realized payoff at each stage.
- Payoff function are unkown
- This is the minimal information framework of the so-called *reinforcement learning* procedures [Borgers and Sarin 97, Erev and Roth 98]

	A	B	C	D	E
R	?	?	?	1	?
S	?	2	-1	?	?
P	?	?	-10	?	?

Actions played : R, S, S, P
Payoff received :**1, -1, 2, -10**

Actions played : D, C, B, C
Payoff received :**-1, 1, -2, 10**

Dropping (i) and (ii)

- One approach (among many others) is to assume that the agents observe only their realized payoff at each stage.
- Payoff function are unkown
- This is the minimal information framework of the so-called *reinforcement learning* procedures [Borgers and Sarin 97, Erev and Roth 98]

How do players use the available information ?

Tipically, it is supposed that players are given a rule of behavior (a *choice rule*) which depends on a *state variable* constructed by means of the aggregate information they gather.

Dropping (iii)

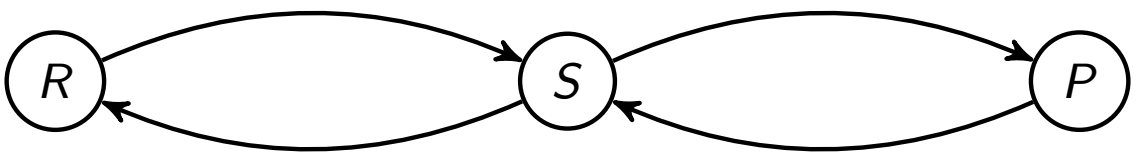
- Players have restrictions on their action set, due to limited computational capacity or even to physical restrictions.
- Some hypotheses are needed regarding payers' ability to explore their action set.

Dropping (iii)

- Players have restrictions on their action set, due to limited computational capacity or even to physical restrictions.
- Some hypotheses are needed regarding payers' ability to explore their action set.

For example :

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



This kind of restrictions were introduced recently by [Benaim and Raimond 10] in the fictitious play information framework.

Our contribution

In this work

We drop all the three assumptions.

Outline

Setting

- Let $\mathcal{G} = (N, (S^i)_{i \in N}, (G^i)_{i \in N})$ be a given finite normal form game
- $S = \prod_i S^i$ is the set of action profiles.
- $\Delta(S^i)$ is the mixed action set for player i, i.e

$$\Delta(S^i) = \left\{ \sigma^i \in \mathbb{R}^{|S^i|} \ : \ \sum_{s^i \in S^i} \sigma^i(s^i) = 1, \ \sigma^i(s^i) \geq 0, \ \forall s^i \in S^i \right\},$$

  and $\Delta = \prod_i \Delta(S^i)$.
- As usual, we use the notation $-i$ to exclude player i, namely S^{-i} denotes the set $\prod_{j \neq i} S^j$ and Δ^{-i} the set $\prod_{j \neq i} \Delta(S^i)$.

Reinforcement learning

A *reinforcement learning* procedure can be defined in the following manner. Let us assume that, at the end of stage $n \in \mathbb{N}$, player i has constructed a *state variable* X_n^i. Then

Reinforcement learning

A *reinforcement learning* procedure can be defined in the following manner. Let us assume that, at the end of stage $n \in \mathbb{N}$, player i has constructed a *state variable* X_n^i. Then

(*a*) at stage $n+1$, player i selects a mixed strategy σ_n^i according to a *decision rule*, which can depend on state variable X_n^i the time n.

Reinforcement learning

A *reinforcement learning* procedure can be defined in the following manner. Let us assume that, at the end of stage $n \in \mathbb{N}$, player i has constructed a *state variable* X_n^i. Then

(*a*) at stage $n+1$, player i selects a mixed strategy σ_n^i according to a *decision rule*, which can depend on state variable X_n^i the time n.

(*b*) Player i's action s_{n+1}^i at time $n+1$ is randomly drawn according to σ_n^i.

Reinforcement learning

A *reinforcement learning* procedure can be defined in the following manner. Let us assume that, at the end of stage $n \in \mathbb{N}$, player i has constructed a *state variable* X_n^i. Then

(*a*) at stage $n+1$, player i selects a mixed strategy σ_n^i according to a *decision rule*, which can depend on state variable X_n^i the time n.

(*b*) Player i's action s_{n+1}^i at time $n+1$ is randomly drawn according to σ_n^i.

(*c*) She only observes $g_{n+1}^i = G^i(s_{n+1}^1, \ldots, s_{n+1}^N)$, as a consequence of the realized action profile $(s_{n+1}^1, \ldots, s_{n+1}^N)$.

Reinforcement learning

A *reinforcement learning* procedure can be defined in the following manner. Let us assume that, at the end of stage $n \in \mathbb{N}$, player i has constructed a *state variable* X_n^i. Then

(*a*) at stage $n+1$, player i selects a mixed strategy σ_n^i according to a *decision rule*, which can depend on state variable X_n^i the time n.

(*b*) Player i's action s_{n+1}^i at time $n+1$ is randomly drawn according to σ_n^i.

(*c*) She only observes $g_{n+1}^i = G^i(s_{n+1}^1, \ldots, s_{n+1}^N)$, as a consequence of the realized action profile $(s_{n+1}^1, \ldots, s_{n+1}^N)$.

(*d*) Finally, this observation allows her to update her state variable to X_{n+1}^i through an *updating rule*, which can depend on observation g_{n+1}^i, state variable X_n^i, and time n.

Reinforcement learning

A *reinforcement learning* procedure can be defined in the following manner. Let us assume that, at the end of stage $n \in \mathbb{N}$, player i has constructed a *state variable* X_n^i. Then

- (a) at stage $n+1$, player i selects a mixed strategy σ_n^i according to a *decision rule*, which can depend on state variable X_n^i the time n.
- (b) Player i's action s_{n+1}^i at time $n+1$ is randomly drawn according to σ_n^i.
- (c) She only observes $g_{n+1}^i = G^i(s_{n+1}^1, \ldots, s_{n+1}^N)$, as a consequence of the realized action profile $(s_{n+1}^1, \ldots, s_{n+1}^N)$.
- (d) Finally, this observation allows her to update her state variable to X_{n+1}^i through an *updating rule*, which can depend on observation g_{n+1}^i, state variable X_n^i, and time n.

An interesting example when such a framework naturally arises : **Congestion games**

Restrictions on the action set

- When an agent i plays a pure strategy $s \in S^i$ at stage $n \in \mathbb{N}$, her available actions at stage $n+1$ are reduced to a subset of S^i.
- Each player has a *exploration matrix* $M_0^i \in \mathbb{R}^{|S^i|}$: if at stage n player i plays $s \in S^i$, she can switch to action $r \neq s$ at stage $n+1$ if and only if $M_0^i(s,r) > 0$.
- The matrix M_0^i is assumed to be irreducible and reversible with respect to its unique invariant measure π_0^i, i.e.

$$\pi_0^i(s) M_0^i(s,r) = \pi_0^i(r) M_0^i(r,s),$$

for every $s, r \in S^i$.

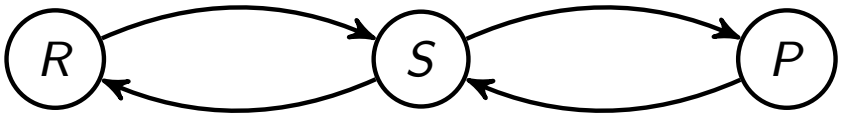
Restrictions on the action set : Examples

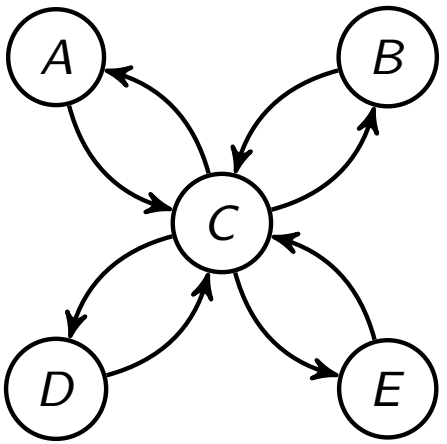
$$M_0^1 = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \end{pmatrix}$$

$$\pi_0^1 = \begin{pmatrix} 2/7 & 3/7 & 2/7 \end{pmatrix}$$

$$M_0^2 = \begin{pmatrix} 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \end{pmatrix}$$

$$\pi_0^2 = \begin{pmatrix} 2/13 & 2/13 & 5/13 & 2/13 & 2/13 \end{pmatrix}$$





Comments on the literature

- Most of the decision rules considered in the literature are *stationary* in the sense that they are defined through a time-independent function of the state variable.
  - 2×2 games [Posch 97]
  - 2-players games with positive payoff [Borgers and Sarin 97, Beggs 05, Hopkins 02, Hopkins and Posch 05]
  - Convergence to perturbed equilibria in 2-player games [Leslie and Collins 03] or multiplayer games [Cominetti, Melo and Sorin 10], [Bravo 12].

Comments on the literature

- Most of the decision rules considered in the literature are *stationary* in the sense that they are defined through a time-independent function of the state variable.
  - 2×2 games [Posch 97]
  - 2-players games with positive payoff [Borgers and Sarin 97, Beggs 05, Hopkins 02, Hopkins and Posch 05]
  - Convergence to perturbed equilibria in 2-player games [Leslie and Collins 03] or multiplayer games [Cominetti, Melo and Sorin 10], [Bravo 12].
- Examples of non-homogeneous (time-dependent) choice rule
  - Convergence of mixed actions is shown for zero-sum games and multiplayer potential games [Leslie and Collins 06]
  - Based on consistent procedures, [Hart and Mas-Colell 01], construct a procedure where, for any game, the joint empirical frequency of play converges to the set of correlated equilibria. (The choice rule is Markovian).

Comments on the literature

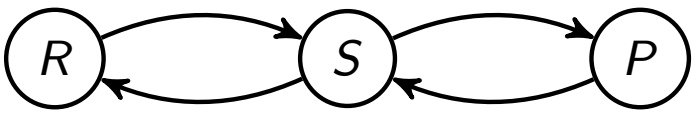
- Most of the decision rules considered in the literature are *stationary* in the sense that they are defined through a time-independent function of the state variable.
  - 2×2 games [Posch 97]
  - 2-players games with positive payoff [Borgers and Sarin 97, Beggs 05, Hopkins 02, Hopkins and Posch 05]
  - Convergence to perturbed equilibria in 2-player games [Leslie and Collins 03] or multiplayer games [Cominetti, Melo and Sorin 10], [Bravo 12].
- Examples of non-homogeneous (time-dependent) choice rule
  - Convergence of mixed actions is shown for zero-sum games and multiplayer potential games [Leslie and Collins 06]
  - Based on consistent procedures, [Hart and Mas-Colell 01], construct a procedure where, for any game, the joint empirical frequency of play converges to the set of correlated equilibria. (The choice rule is Markovian).
- However, in all the examples described above, players can use any action at any time.

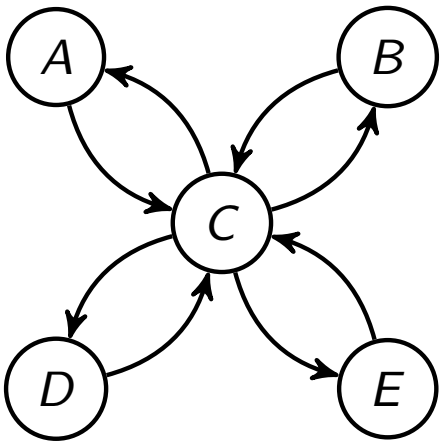
Intuition on the discrete dynamics (zero-sum game)

	A	B	C	D	E
R	?	?	?	?	?
S	?	?	?	?	?
P	?	?	?	?	?

Actions played :
Payoff received :

Actions played
Payoff received :





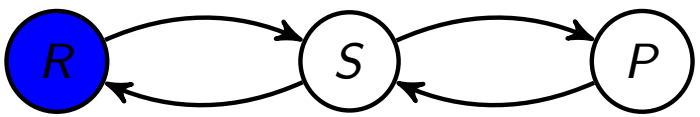
We are interested in the asymptotic behavior of the empirical frequencies of play, i.e. the limit set of the occupation measures on the graphs.

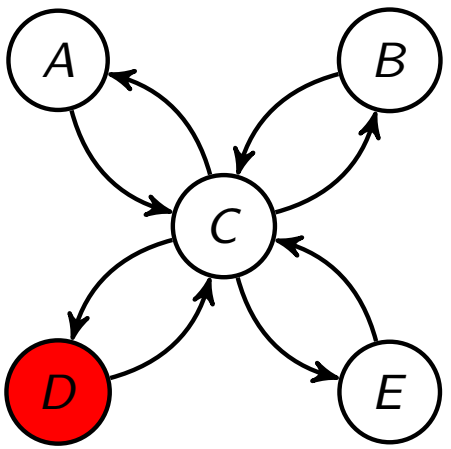
Intuition on the discrete dynamics (zero-sum game)

	A	B	C	D	E
R	?	?	?	1	?
S	?	?	?	?	?
P	?	?	?	?	?

Actions played : R
Payoff received :**1**

Actions played : D
Payoff received :-**1**





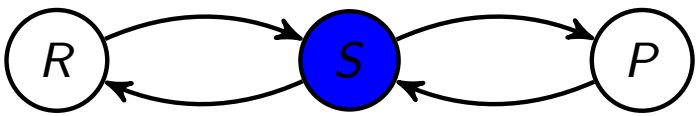
We are interested in the asymptotic behavior of the empirical frequencies of play, i.e. the limit set of the asymptotic occupation measures on the graphs.

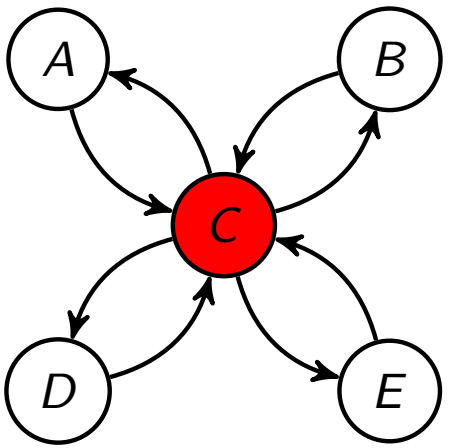
Intuition on the discrete dynamics (zero-sum game)

	A	B	C	D	E
R	?	?	?	1	?
S	?	?	-1	?	?
P	?	?	?	?	?

Actions played : R, S
Payoff received :**1, -1**

Actions played : D, C
Payoff received :**-1, 1**





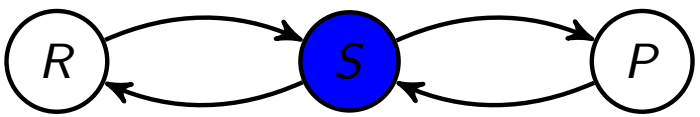
We are interested in the asymptotic behavior of the empirical frequencies of play, i.e. the limit set of the asymptotic occupation measures on the graphs.

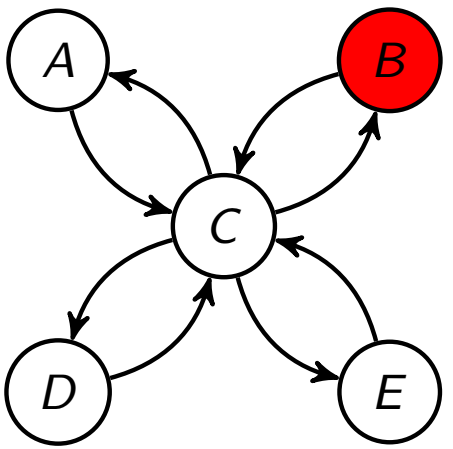
Intuition on the discrete dynamics (zero-sum game)

	A	B	C	D	E
R	?	?	?	1	?
S	?	2	-1	?	?
P	?	?	?	?	?

Actions played : R, S, S
Payoff received :**1, -1, 2**

Actions played : D, C, B
Payoff received :**-1, 1, -2**





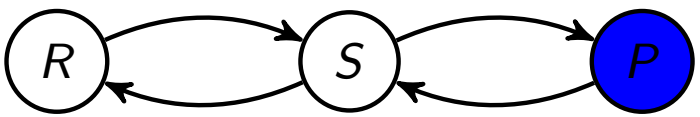
We are interested in the asymptotic behavior of the empirical frequencies of play, i.e. the limit set of the asymptotic occupation measures on the graphs.

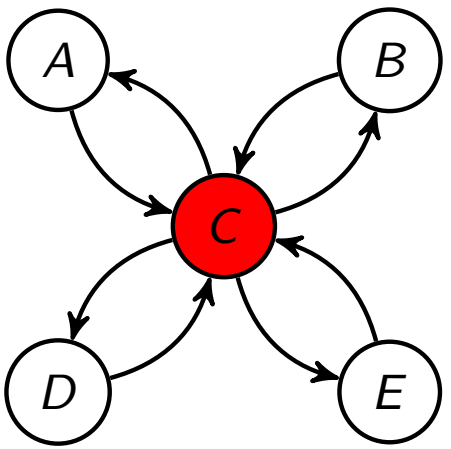
Intuition on the discrete dynamics (zero-sum game)

	A	B	C	D	E
R	?	?	?	1	?
S	?	2	-1	?	?
P	?	?	-10	?	?

Actions played : R, S, S, P
Payoff received :**1, -1, 2, -10**

Actions played : D, C, B, C
Payoff received :**-1, 1, -2, 10**





We are interested in the asymptotic behavior of the empirical frequencies of play, i.e. the limit set of the asymptotic occupation measures on the graphs.

Payoff-based Markovian procedure

Q : How to define precise rules for the players in order to achieve convergence to the set of Nash equilibria of the underlying game ?

Payoff-based Markovian procedure

Q : How to define precise rules for the players in order to achieve convergence to the set of Nash equilibria of the underlying game ?

We need some notation :

- For $\beta > 0$ and a vector $R \in \mathbb{R}^{|S^i|}$, we define the stochastic matrix $M^i[\beta, R]$ as

$$M^i[\beta, R](s, r) = \begin{cases} M_0^i(s, r) \exp(-\beta |R(s) - R(r)|_+) & s \neq r \\ 1 - \sum_{s' \neq s} M^i[\beta, R](s, s') & s = r, \end{cases}$$

- The matrix $M^i[\beta, R]$ is irreducible and its invariant measure of the matrix is given explicitly by

$$\pi^i[\beta, R](s) = \frac{\pi_0^i(s) \exp(\beta R(s))}{\sum_{r \in S^i} \pi_0^i(r) \exp(\beta R(r))},$$

for any $\beta > 0$, $R \in \mathbb{R}^{|S^i|}$, and $s \in S^i$.

Choice rule of player i

- At the end of the stage n, player i has a state variable $R_n^i \in \mathbb{R}^{|S^i|}$
- Let $M_n^i = M^i[\beta_n, R_n^i]$ and $\pi_n^i = \pi^i[\beta_n^i, R_n^i]$, where $(\beta_n^i)_n$ is a strictly positive sequence

Choice rule

The choice rule of player i is

$$\begin{aligned}\sigma_n^i(s) &= \mathbb{P}(s_{n+1}^i = s \,|\, \mathcal{F}_n) \\ &= M_n^i(s_n^i, s), \\ &= \begin{cases} M_0^i(s_n^i, s)\exp(-\beta_n^i |R_n^i(s_n^i) - R_n^i(s)|_+) & s \neq s_n^i \\ 1 - \sum_{s' \neq s} M_n^i(s_n^i, s') & s = s_n^i. \end{cases}\end{aligned} \tag{CR}$$

Updating rule of player i

- After observing the realized payoff $g^i_{n+1} = G^1(s_{n+1^i}, s^{-i}_{n+1})$, player updates the state variable R^i_n as

Updating Rule

$$R^i_{n+1}(s) = R^i_n(s) + \gamma^i_{n+1}(s)\left(g^i_{n+1} - R^i_n(s)\right)\mathbb{1}_{\{s^1_{n+1}=s\}}, \qquad \text{(UR)}$$

where,

$$\gamma^i_{n+1}(s) = \min\left\{1\,,\,\frac{1}{(n+1)\pi^i_n(s)}\right\},$$

and $\mathbb{1}_E$ is the indicator of the event E.

Updating rule of player i

- After observing the realized payoff $g^i_{n+1} = G^1(s_{n+1^i}, s^{-i}_{n+1})$, player updates the state variable R^i_n as

Updating Rule

$$R^i_{n+1}(s) = R^i_n(s) + \frac{1}{(n+1)\pi^i_n(s)} \left(g^i_{n+1} - R^i_n(s) \right) \mathbb{1}_{\{s^1_{n+1} = s\}}, \qquad \text{(UR)}$$

where $\mathbb{1}_E$ is the indicator of the event E.

Payoff-based Markovian procedure

PBMP

We call *Payoff-based Markovian procedure* the adaptive process where, for any $i \in N$, agent i plays according to the choice rule (CR), and updates R_n^i through the updating rule (UR).

Outline

Assumptions

In the case of a 2-player game, we introduce our major assumption on the positive sequence $(\beta_n^i)_n$.

Assumption

Let us assume that, for $i \in \{1, 2\}$,

$$\begin{aligned} &\text{(i) } \beta_n^i \longrightarrow +\infty, \\ &\text{(ii) } \beta_n^i \leq A_n^i \ln(n), \text{ where } A_n^i \longrightarrow 0. \end{aligned} \tag{H}$$

Assumptions

In the case of a 2-player game, we introduce our major assumption on the positive sequence $(\beta_n^i)_n$.

Assumption

Let us assume that, for $i \in \{1,2\}$,

$$\begin{aligned} &\text{(i) } \beta_n^i \longrightarrow +\infty, \\ &\text{(ii) } \beta_n^i \leq A_n^i \ln(n), \text{ where } A_n^i \longrightarrow 0. \end{aligned} \tag{H}$$

- Let us denote by v_n^i and $\overline{g}_n^i$ the empirical frequency of play and the average payoff obtained by player i up to time n, i.e., respectively

$$v_n^i = \frac{1}{n}\sum_{m=1}^{n} \delta_{s_m^i} \text{ and } \overline{g}_n^i = \frac{1}{n}\sum_{m=1}^{n} G^i(s_m^1, s_m^2).$$

- For a sequence $(z_n)_n$, we call $\mathcal{L}((z_n)_n)$ its limit set , i.e.

$$\mathcal{L}((z_n)_n) = \left\{z : \text{ there exists a subsequence } (z_{n_k})_k \text{ such that } \lim_{k\to+\infty} z_{n_k} = z\right\}.$$

We say that the sequence $(z_n)_n$ converges to a set A if $\mathcal{L}((z_n)_n) \subseteq A$.

Main result

Theorem

Under assumption (H), the Payoff-based Markovian procedure enjoys the following properties :

(a) In a zero-sum game, $(v_n^1, v_n^2)_n$ converges almost surely to the set of Nash equilibria and the average payoff $(\overline{g}_n^1)_n$ converges almost surely to the value of the game.

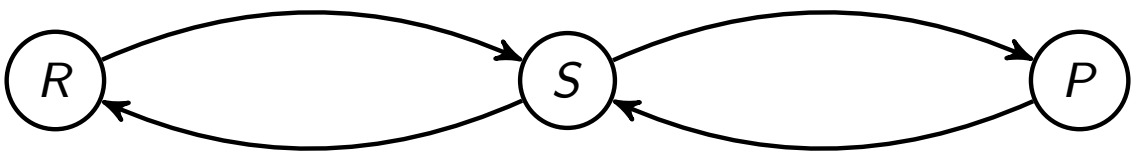
(b) In a potential game with potential Φ, $(v_n^1, v_n^2)_n$ converges almost surely to a connected subset of the set of Nash equilibria on which Φ is constant, and $\frac{1}{n}\sum_{m=1}^{n} \Phi(s_m^1, s_m^2)$ converges to this constant.
In the particular case $G^1 = G^2$, then $(v_n^1, v_n^2)_n$ converges almost surely to a connected subset of the set of Nash equilibria on which G^1 is constant ; moreover $(\overline{g}_n^1)_n$ converges almost surely to this constant.

(c) If either $|S^1| = 2$ or $|S^2| = 2$, then $(v_n^1, v_n^2)_n$ converges almost surely to the set of Nash equilibria.

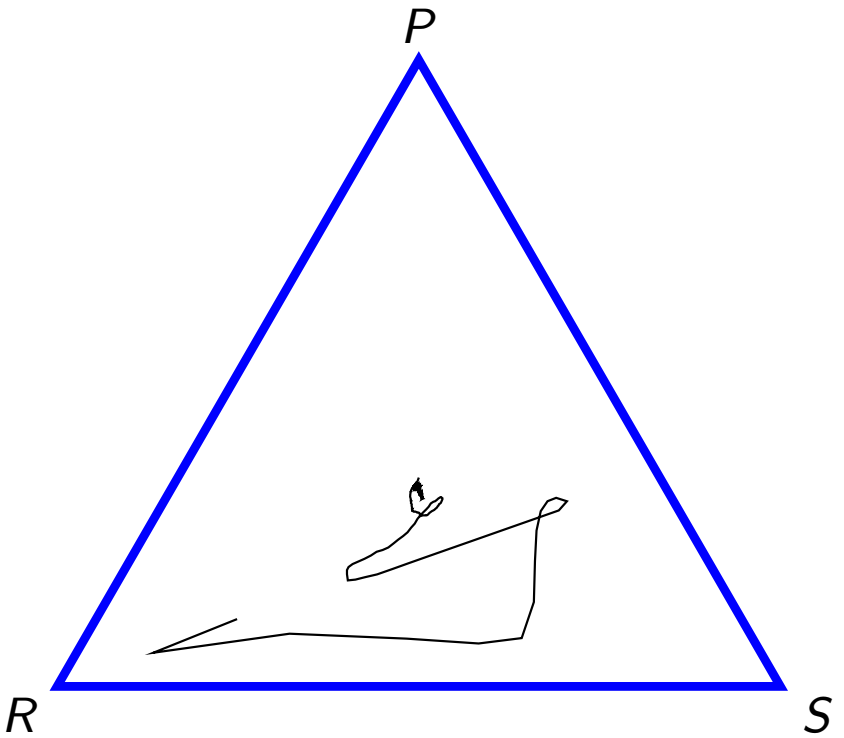
Outline

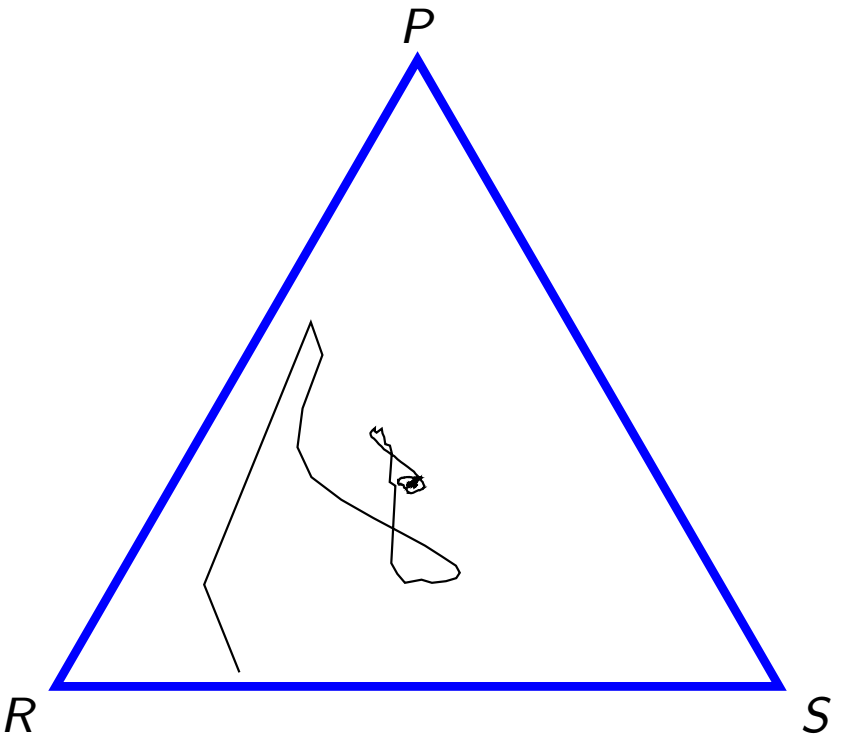
Blind-restricted RSP

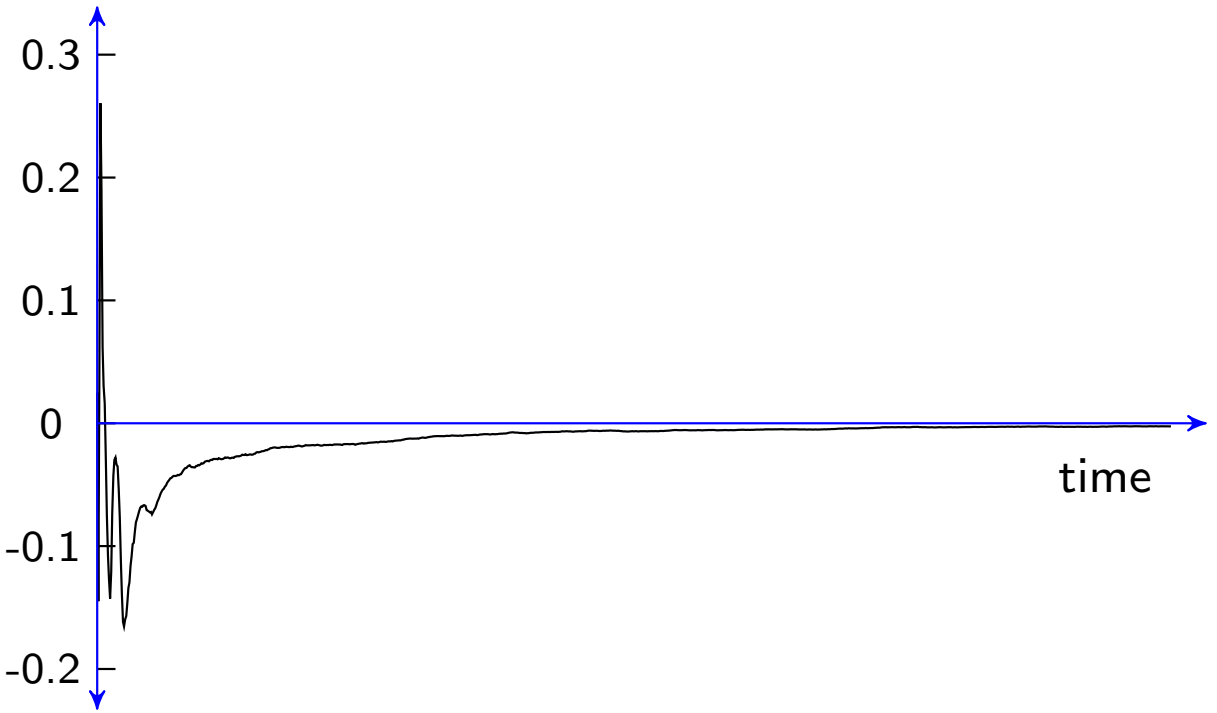
	R	S	P
R	0	1	-1
S	-1	0	1
P	1	-1	0



Optimal strategies are given by $((1/3, 1/3, 1/3), (1/3, 1/3, 1/3)) \in \Delta$ and the value of the game is 0.


P
R
S


P
R
S

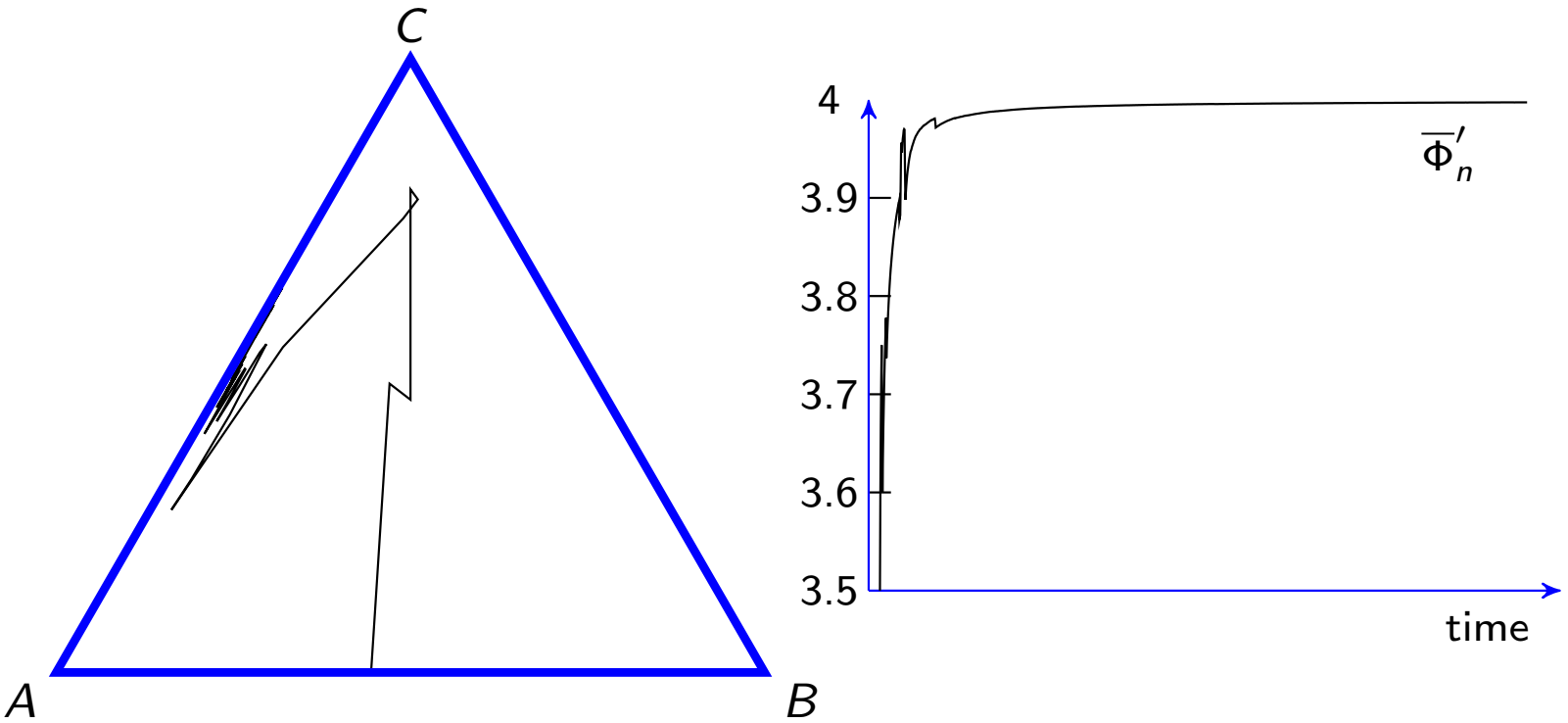

0.3
0.2
0.1
0
-0.1
-0.2
time

A 3×3 Potential game

$$G = \begin{array}{c|ccc} & a & b & c \\ A & 1,1 & 9,0 & 1,0 \\ B & 0,9 & 6,6 & 0,8 \\ C & 1,2 & 8,0 & 2,2 \end{array} \quad \text{and} \quad \Phi = \begin{array}{c|ccc} & a & b & c \\ A & 4 & 3 & 3 \\ B & 3 & 0 & 2 \\ C & 4 & 2 & 4 \end{array}$$

Here we see that the set of Nash equilibria is connected and equal to

$$NE = \{((x, 0, 1-x), a),\ x \in [0,1]\} \cup \{(C, (y, 0, 1-y)),\ y \in [0,1]\}.$$


C
A
B
4
3.9
3.8
3.7
3.6
3.5
$\overline{\Phi}'_n$
time

A 3×3 Potential game

$$G' = \begin{array}{c|ccc} & a & b & c \\ A & 1,1 & 9,0 & 1,0 \\ B & 0,9 & 6,6 & 0,8 \\ C & 0,1 & 9,0 & 2,2 \end{array} \quad \text{and} \quad \Phi' = \begin{array}{c|ccc} & a & b & c \\ A & 4 & 3 & 3 \\ B & 3 & 0 & 2 \\ C & 3 & 2 & 4 \end{array} \qquad (\mathcal{G})$$

There is a mixed Nash equilibrium, and two strict Nash equilibria (A, a) and (C, c), with same potential value (equal to 4). However,

$$\mathbb{P}\left[\mathcal{L}((v_n)_n) = \{(A, a), (C, c)\}\right] = 0.$$

Outline

Definition

The Best-Response correspondence for player $i \in \{1,2\}$, $\mathrm{BR}^i : \Delta^{-i} \rightrightarrows \Delta(S^i)$, is defined as

$$\mathrm{BR}^i(\sigma^{-i}) = \mathrm{argmax}_{\sigma^i \in \Delta(S^i)} G^i(\sigma^i, \sigma^{-i}).$$

for any $\sigma^{-i} \in \Delta^{-i}$. The Best-Response correspondence $\mathrm{BR} : \Delta \rightrightarrows \Delta$ is given by

$$\mathrm{BR}(\sigma) = \prod_{i \in \{1,2\}} \mathrm{BR}^i(\sigma^{-i}),$$

for all $\sigma \in \Delta$.

Definition

The Best-Response correspondence for player $i \in \{1,2\}$, $\mathrm{BR}^i : \Delta^{-i} \rightrightarrows \Delta(S^i)$, is defined as

$$\mathrm{BR}^i(\sigma^{-i}) = \mathrm{argmax}_{\sigma^i \in \Delta(S^i)} G^i(\sigma^i, \sigma^{-i}).$$

for any $\sigma^{-i} \in \Delta^{-i}$. The Best-Response correspondence $\mathrm{BR} : \Delta \rightrightarrows \Delta$ is given by

$$\mathrm{BR}(\sigma) = \prod_{i \in \{1,2\}} \mathrm{BR}^i(\sigma^{-i}),$$

for all $\sigma \in \Delta$.

In fact we show a more general result

Theorem

Under hypothesis (H), assume that players follow the Payoff-based adaptive Markovian procedure. Assume that the Best-Response dynamics

$$\dot{v} \in \mathrm{BR}(v) - v$$

has an attractor $\mathcal{A}$. Then $\mathcal{L}((v_n)_n) \subseteq \mathcal{A}$.

Then we will use known results on the Best-Response dynamics

Evolution on v_n^i

$$\begin{aligned} v_{n+1}^i - v_n^i &= \frac{1}{n+1}\left(\delta_{s_{n+1}^i} - v_n^i\right), \\ &= \frac{1}{n+1}\left(\pi_n^i - v_n^i + W_{n+1}^1\right) \end{aligned}$$

where

$$W_{n+1}^i = \delta_{s_{n+1}^i} - \pi_n^i.$$

Evolution on v_n^i. It would be very nice that...

$$v_{n+1}^i - v_n^i = \frac{1}{n+1}\left(\delta_{s_{n+1}^i} - v_n^i\right),$$
$$\in \frac{1}{n+1}\left(\mathrm{BR}^i(v_n^{-i}) - v_n^i + W_{n+1}^i\right)$$

where

$$W_{n+1}^i = \delta_{s_{n+1}^1} - \pi_n^i.$$

Evolution on v_n^i. It would be very nice that...

$$v_{n+1}^i - v_n^i = \frac{1}{n+1}\left(\delta_{s_{n+1}^i} - v_n^i\right),$$
$$\in \frac{1}{n+1}\left(\mathrm{BR}^i(v_n^{-i}) - v_n^i + W_{n+1}^i\right)$$

where

$$W_{n+1}^i = \delta_{s_{n+1}^i} - \pi_n^i.$$

- **Major problem :** π_n^i depends on R_n^i and also is a function of the time n !.
- **We would like to replace** π_n^1 **by** $\pi^i[\beta_n^1, G^1(\cdot, v_n^2)]$ **when** n **is large.**

The difficult part

Proposition 1

For any $i \in \{1, 2\}$, we have that $R_n^i - G^i(\cdot, v_n^{-i}) \to 0$ goes to zero almost surely as n goes to infinity.

Then we can show

Property

For any almost sure limit point $(v^1, v^2, \pi^1, \pi^2) \in (\Delta(S^1) \times \Delta(S^2))^2$ of the random process $(v_n^1, v_n^2, \pi_n^1, \pi_n^2)_n$ we have that, for $i \in \{1, 2\}$

$$\pi^i \in \mathrm{BR}^i(v^{-i})$$

The difficult part

Proposition 1

For any $i \in \{1, 2\}$, we have that $R_n^i - G^i(\cdot, v_n^{-i}) \to 0$ goes to zero almost surely as n goes to infinity.

Then we can show

Evolution on v_n^i.

$$\begin{aligned} v_{n+1}^i - v_n^i &= \frac{1}{n+1}\left(\delta_{s_{n+1}^i} - v_n^i\right), \\ &\in \frac{1}{n+1}\left([\mathrm{BR}^i]^\epsilon(v_n^{-i}) - v_n^i + W_{n+1}^i\right) \end{aligned}$$

for any $\varepsilon > 0$.

To conclude, we use some recent results on stochastic approximaxion theory for differential inclusions [Benaim and Raimond 10], [Benaim, Hofbauer and Sorin 05]).
If, in adition,

- For $i \in \{1, 2\}$, $\epsilon\left(\frac{1}{n+1} W_{n+1}^i, T\right)$ goes to zero almost surely for all $T > 0$, where

$$\epsilon(u_n, T) = \sup \left\{ \left\| \sum_{j=n}^{l-1} u_{j+1} \right\| ; \ \ l \in \{n+1, \ldots, m(\tau_n + T)\} \right\},$$

  for a sequence $(u_n)_n$

To conclude, we use some recent results on stochastic approximaxion theory for differential inclusions [Benaim and Raimond 10], [Benaim, Hofbauer and Sorin 05]).
If, in adition,

- For $i \in \{1,2\}$, $\epsilon\left(\frac{1}{n+1} W^i_{n+1}, T\right)$ goes to zero almost surely for all $T > 0$, where

$$\epsilon(u_n, T) = \sup\left\{\left\|\sum_{j=n}^{l-1} u_{j+1}\right\|;\ \ l \in \{n+1, \ldots, m(\tau_n + T)\}\right\},$$

  for a sequence $(u_n)_n$

Therefore, if the Best-Response dynamics

$$\dot{v} \in \mathrm{BR}(v) - v$$

has an attractor $\mathcal{A}$.
Then

$$\mathcal{L}((v_n)_n) \subseteq \mathcal{A}$$

.

Proof of Proposition 1

$$
\begin{aligned}
R^i_{n+1}(s) - R^i_n(s) &= \frac{1}{(n+1)\pi^i_n(s)} \left[\mathbb{1}_{\{s^i_{n+1}=s\}} G^i(s, s^{-i}_{n+1}) - \mathbb{1}_{\{s^i_{n+1}=s\}} R^i_n(s) \right], \\
&= \frac{1}{(n+1)\pi^i_n(s)} \Big[\pi^i_n(s) \left(G^i(s, \pi^{-i}_n) - R^i_n(s) \right) + \\
&+ \left(\mathbb{1}_{\{s^i_{n+1}=s\}} G^i(s, s^{-i}_{n+1}) - \pi^i_n(s) G^i(s, \pi^{-i}_n) \right) + \\
&+ R^i_n(s) \left(\pi^i_n(s) - \mathbb{1}_{\{s^i_{n+1}=s\}} \right) \Big], \\
&= \frac{1}{n+1} \left[G^i(s, \pi^{-i}_n) - R^i_n(s) + \overline{W}^i_{n+1}(s) \right]
\end{aligned}
$$

If $U^i_n = G^i(\cdot, v^{-i}_n)$. Then

$$
U^i_{n+1} - U^i_n = \frac{1}{n+1} \left(G^i(\cdot, \pi^{-i}_n) - U^i_n + \tilde{W}^i_{n+1} \right).
$$

We define $\zeta^i_n = R^i_n - G^i(\cdot, v^{-i}_n) = R^i_n - U^i_n$.

Therefore

$$\zeta^i_{n+1} - \zeta^i_n = \frac{1}{n+1}\big[-\zeta^i_n + \mathcal{W}^i_{n+1}\big],$$

Log-Sobolev estimation via the spectral gap for Markov chains are needed to show that $\epsilon(\frac{1}{n+1}\mathcal{W}^i_{n+1}, T)$ goes to zero almost surely for any $T > 0$. This is the really hard part !

Finally, using the standard stochastic approximation theory and the fact that the ODE $\dot{\zeta} = -\zeta$ has the set $\{0\}$ as a attractor we can conclude.

Thanks for your attention

If you want to get into more details, the paper is available at
http://arxiv.org/abs/1306.2918