Topics on strategic learning

Sylvain Sorin

IMJ-PRG
Université P. et M. Curie - Paris 6
IMJ-PRG
sylvain.sorin@imj-prg.fr

Stochastic Methods in Game Theory
National University of Singapore
Institute for Mathematical Sciences
November 16, 2015

Topics on strategic learning III:

Global procedures

1. Application to games

Let $\mathscr{G}$ a finite game in strategic form.
There are finitely many players labeled $i = 1, 2, \ldots, I$.
S^i is the finite moves set of player i, $S = \prod_i S^i$, and $Z = \Delta(S)$ is the set of probabilities on S (correlated moves).
We will consider repeated interaction in discrete time where at each stage the players observe the actions of their opponents.
We want evaluate the joint impact on the play of the prescribed behavior of the players.
Since we will study the procedure from the view point of player 1 it is convenient to set $S^1 = K, X = \Delta(K)$ (mixed moves of player 1), $L = \prod_{i \neq 1} S^i$, and $Y = \Delta(L)$ (correlated moves of player 1's opponents) hence $Z = \Delta(K \times L)$.
$F : S \to \mathbb{R}$ denotes the payoff function of player 1 and we still denote by F its linear extension to Z, and its bilinear extension to $X \times Y$.

2. External consistency and Hannan's set

Let m be the cardinality of K and $R(z)$ denote the m-dimensional vector of regrets for player 1 at z in Z, defined by

$$R(z) = \{F(k, z^{-1}) - F(z)\}_{k \in K}$$

where z^{-1} stands for the marginal of z on L.
(Player 1 compares his payoff using a given move k to his actual payoff, assuming the other players' behavior, z^{-1}, given.)

Definition

H^1 (for Hannan's set) is the set of correlated moves in Z satisfying the no-regret condition for player 1.
Formally:

$$H^1 = \{z \in Z : F(k, z^{-1}) \leq F(z), \forall k \in K\} = \{z \in Z : R(z) \in D = \mathbb{R}^K_-\}.$$

The main property is that if player 1 uses a procedure with no external regret in the on line problem corresponding to the repeated game where the outcome vector at stage m is $\{F(k,\ell_m\}_{k\in K}\}$, where ℓ_m is the profile of moves of his opponents, the empirical average distribution

$$z_n = \frac{1}{n}\sum_{m=1}^{n}(k_m,\ell_m) \in Z$$

will converge to H.

Proposition

If Player 1 follows some external consistent procedure, the empirical distribution of moves converges a.s. to the Hannan set H^1.

Proof

The proof is straightforward due to the linearity of the payoff: the consistency property is

$$\frac{1}{n}\sum_{m=1}^{n} F(k,\ell_m) - \frac{1}{n}\sum_{m=1}^{n} F(k_m,\ell_m) \leq o(n) \qquad \forall k \in K$$

which gives

$$F(k,\frac{1}{n}\sum_{m=1}^{n} \ell_m)) - F(\frac{1}{n}\sum_{m=1}^{n} (k_m,\ell_m)) \leq o(n) \qquad \forall k \in K$$

and this expression is

$$F(k,z_n^{-1}) - F(z_n) \leq o(n) \qquad \forall k \in K.$$

■

Alternative proof

We consider an auxiliary game with vector payoffs in $\mathbb{R}^M$, where the dimension is $M = L+1$, and the payoff $g(s) = (F(s), s^{-1})$ is the couple of the current payoff in the original game and of the opponent(s) profile.
D_1 is the convex set:

$$D_1 = \{(u,\theta) \in \mathbb{R} \times \Delta(S^{-1}); u \geq \max_{s^1 \in S^1} F(s,\theta)\}.$$

Theorem

D_1 is approachable.

Proof

The proof that D_1 is approachable is that it is not excludable: namely, for any $\theta \in \Delta(S^{-1})$, there is some $s^1 \in S^1$ such that $g(s,\theta) \in D_1$. ■

This obviously implies the non emptiness of H^1 since by approachability $d(\bar{g}_n, D_1)$ goes to 0 hence also $[\max_{k \in S^1} F(k, \bar{z}_n^{-1}) - F(\bar{z}_n)]^+$.

One defines similarly H^i for each player and $H = \cap_i H^i$ which is the global Hannan's set.

Proposition

If all players follow some external consistent procedure, the empirical distribution of moves converges a.s. to the Hannan set H.

Note that no coordination is required.

3. Correlated equilibria

Examples

Battle of the sexes

$3,1$	$0,0$
$0,0$	$1,3$

There are 2 pures and efficient equilibria and one mixed equilibrium $(3/4,1/4);1/4;3/4)$ Pareto dominated.
One can use a public device (fair coin) to get a symmetric Pareto efficient equilibrium.
Play $(3,1)$ if Head and $(1,3)$ if Tail.
The induced distribution is

$1/2$	0
0	$1/2$

Consider now

	L	R
T	$2,7$	$6,6$
B	$0,0$	$7,2$

Consider a signal space: $(\mathbf{w},\mathbf{g},\mathbf{b})$ with the uniform probability $(1/3,1/3,1/3)$.
Assume that the players get private messages:
1 knows $a=\{\mathbf{w},\mathbf{g}\}$ or $b=\{\mathbf{b}\}$ and
2 knows $\alpha=\{\mathbf{w}\}$ or $\beta=\{\mathbf{g},\mathbf{b}\}$.
Consider the strategies:
T if a, B if b for player 1;
L if α, R if β for player 2.
They induce on S the correlation matrix

$1/3$	$1/3$
0	$1/3$

and no deviation is profitable.

Information structure and extended game

Definition

An information structure $\mathscr{I}$ is given by:
- a probability space $(\Omega, \mathscr{A}, P)$
- a family of measurable maps θ^i from $(\Omega, \mathscr{A})$ to A^i (set of signals for player i) (or a sub σ-algebra $\mathscr{A}^i$).

Let $\mathscr{G}$ be a strategic game defined by $G : S = \Pi_i S^i \to \mathbb{R}^I$.

Definition

The game $\mathscr{G}$ extended par $\mathscr{I}$, denoted $[\mathscr{G}, \mathscr{I}]$, is the game in extensive form played in 2 stages:
stage 0 : the random variable ω is selected according to P and the signal $\theta^i(\omega)$ is sent to player i.
stage 1: the players play in the game $\mathscr{G}$.

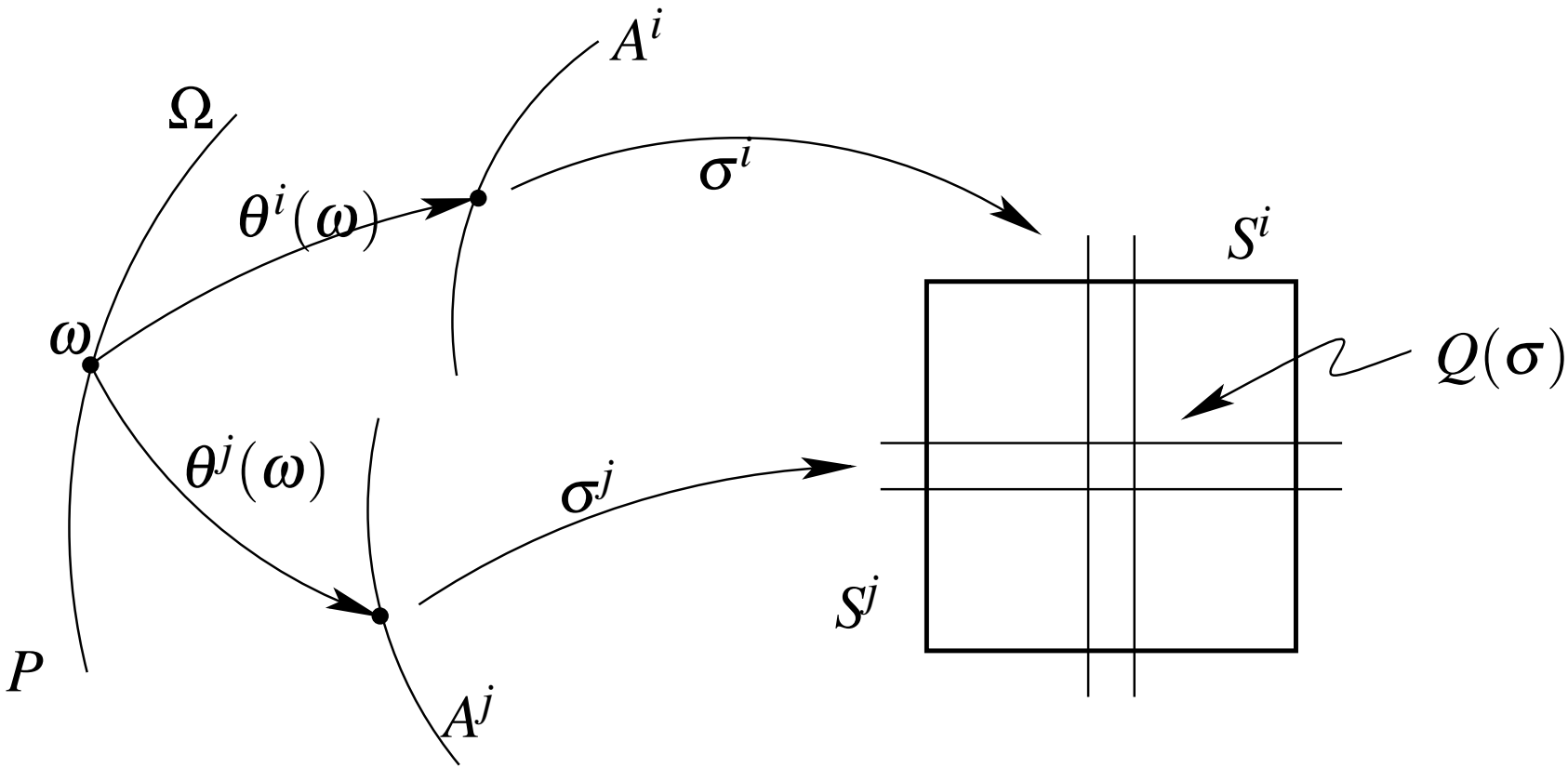
A strategy σ^i of player i in $[\mathscr{G}, \mathscr{I}]$ is a measurable map from A^i to S^i (or a $\mathscr{A}^i$-measurable map from Ω to S^i).

Correlated equilibria

Definition

A correlated equilibrium of $\mathscr{G}$ is a Nash equilibrium of $[\mathscr{G}, \mathscr{I}]$.

A profil σ of strategies in $[\mathscr{G}, \mathscr{I}]$ maps the initial probability P on Ω into a probability $Q(\sigma)$ on S: random variable $\rightarrow$ signal $\rightarrow$ move.



Explicitly, for each ω, $Q(\omega, \sigma)$ is the product probability on S given by $\prod_i \sigma^i(\theta^i(\omega))$ and $Q(\sigma)$ is the expectation w.r.t. P.

Definition

$CED(\mathscr{G})$ is the set of equilibrium correlated distributions in $\mathscr{G}$:

$$CED(\mathscr{G}) = \cup_{\mathscr{I},\sigma}\{Q(\sigma); \sigma \text{ equilibrium in } [\mathscr{G}, \mathscr{I}]\}$$

Note that $CED(\mathscr{G})$ is a convex set: consider the convex combination of information structures.

Canonical correlated equilibria

Definition

A canonical information structure for $\mathscr{G}$ corresponds to the case where: $\Omega = S$; $\theta^i : S \to S^i, \theta^i(s) = s^i$.
P is a probability on the space of profiles S and each player is informed on his component.
A canonical correlated equilibrium is an equilibrium of $\mathscr{G}$ extended by a canonical information structure where moreover the equilibrium strategies satisfy :

$$\sigma^i(\omega) = \sigma^i(s) = \sigma^i(s^i) = s^i \quad \text{“each player follows his signal”}$$

The associated canonical correlated equilibrium distribution ($CCED$) is obviously P.

Characterization

Theorem (Aumann, 1974)

$$CCED(\mathscr{G}) = CED(\mathscr{G})$$

Proof

Let σ an equilibrium profile in an extension $[\mathscr{G}, \mathscr{I}]$ and $Q = Q(\sigma)$ the induced distribution.
Then Q belongs to $CCED(\mathscr{G})$. In fact each player i gets less information in the canonical information structure: his move s^i rather than the signal a^i such that $\sigma^i(a^i) = s^i$. Then s^i is a best reply to the correlated strategy of $-i$ conditionally to a^i. We use then the convexity of BR^i on $\Delta(S^{-i})$. ■

Theorem

$Q \in CED(\mathscr{G})$ *iff*

$$\forall s^i, t^i \in S^i, \forall i = 1, \ldots, n \qquad \sum_{s^{-i} \in S^{-i}} [G^i(s^i, s^{-i}) - G^i(t^i, s^{-i})] Q(s^i, s^{-i}) \geq 0.$$

Proof

One can assume $Q \in CCED(\mathscr{G})$. If s^i is announced (i.e. its marginal $Q^i(s^i) = \sum_{s^{-i}} Q(s^i, s^{-i}) > 0$) introduce the conditional distribution on S^{-i}, $Q(.|s^i)$, and the equilibrium condition writes:

$$s^i \in BR^i(Q(.|s^i).$$

s^i is a best reply of player i to the conditional distribution of the signals, hence moves of the other payers given s^i . ■

Corollary

$CED(\mathscr{G})$ *is the convex hull of finitely many points.*

Hart and Schmeidler (1989) give an elementary existence proof of correlated equilibria by using the minmax theorem.
There are $CCED$ outside the convex hull of Nash distributions:

$0,0$	$5,4$	$4,5$
$4,5$	$0,0$	$5,4$
$5,4$	$4,5$	$0,0$

The only Nash equilibrium is symetrical $(1/3,1/3,1/3)$ with payoff 3.
But a correlated equilibrium distribution is given by:

0	$1/6$	$1/6$
$1/6$	0	$1/6$
$1/6$	$1/6$	0

inducing a payoff 9/2.

Application lo learning procedures

Back to the repeated game framework we still consider only player 1 and denote by F his payoff.

Given $z = (z_s)_{s \in S} \in Z$, introduce the family of m comparison vectors of dimension m (testing k against j with $(j,k) \in K^2$) defined by

$$C(j,k)(z) = \sum_{\ell \in L} [F(k,\ell) - F(j,\ell)] z_{(j,\ell)}.$$

(This corresponds to the change in the expected gain of Player 1 at z when replacing move j by k.)

Remark that if one let $(z \mid j)$ denote the conditional probability on L induced by z given $j \in K$ and z^1 the marginal on K, then

$$\{C(j,k)(z)\}_{k \in K} = z^1_j R((z \mid j))$$

where we recall that $R((z \mid j))$ is the vector of regrets for player 1 at $(z \mid j)$.

Definition

The set of no internal regret (for player 1) is

$$C^1 = \{z \in Z; C(j,k)(z) \leq 0, \forall j,k \in K\}.$$

It is obviously a subset of H^1 since

$$\sum_j \{C(j,k)(z)\}_{k \in I} = R(z).$$

As above, when considering the payoff vectors generated by the moves of the opponents in the repeated game one obtains:

Proposition

If Player 1 follows some internal consistency procedure, the empirical distribution of moves converges a.s. to the set C^1.

Recall that the set of correlated equilibrium distribution of the game $\{F\}$ is defined by

$$C = \{z \in Z; \sum_{\ell \in L} [F^i(k,\ell) - F^i(j,\ell)] z_{(j,\ell)} \leq 0, \qquad \forall j,k \in S^i, \forall i \in I\}.$$

Hence one has :

Proposition

The intersection over all $i \in I$ of the sets C^i is the set of correlated equilibrium distribution of the game.

Thus we obtain:

Corollary

If each player follows some internal consistency procedure, the empirical distribution of moves converges a.s. to the set of correlated equilibria.

Note that this provides an alternative proof of existence of correlated equilibrium through the existence of internally consistent procedures.

Alternative joint algorithm

(Hart and Mas Colell (2000)

The procedure is defined (for player 1) by x^1_{n+1} being a function of his average regret, his last move $s^1_n = j$ and some large parameter L, as follows:

$$x^1_{n+1}(k) = R^+_n(j,k)/L, \quad k \neq j \qquad ; \qquad x^1_{n+1} = 1 - \sum_{k \neq j} x^1_{n+1}(k).$$

Theorem

If all players use procedure B, the empirical distribution of moves converge to the set of correlated equilibria.

4. From calibrating to correlated equilibrium

Foster and Vohra (1997)

Consider the case where Player 1 is forecasting the behavior (a profile in L) of his opponents.

Given a precision level δ, Player 1 is thus predicting points in a δ-grid $\{p[v], v \in V\}$ of $\Delta(L)$ and then plays a (pure) best reply to his forecast.

It is thus clear that if the forecast is calibrated the empirical distribution of the moves of the opponents, will converge to the forecast, on each event of the form $\{m; p_m = p[v] \in \Delta(L)\}$, hence eventually the action chosen by Player 1, k, will be a best reply to the frequency near $p[v]$.

When looking at the average empirical distribution z, the conditional distribution $z|k$ of z given k, will correspond to a convex combination of distributions $p[v]$ to which k is best reply, hence k will again be (approximate) best reply to $z|k$: hence z is (approximately) in C^1.

If all players use calibrated strategies the empirical average frequency converges to C.

5. No convergence to Nash

There is no uncoupled deterministic smooth dynamic that converges to Nash equilibrium in all finite 2-person games: Hart and Mas-Colell (2003).
Similarly there are no learning process with finite memory such that the stage behavior will converge to Nash equilibrium: Hart and Mas-Colell (2005).
Similar results were obtained for MAD dynamics, Hofbauer and Swinkels (1995)
see also Foster and Young (2001) On the impossibility of predicting.
Young (2002) On the limits to rational learning .

6. Weak calibration and deterministic procedure

We follow Kakade and Foster (2004, 2008)

Weak calibration

A general definition of calibrating for X with values in Ω (or $\mathbb{R}^\Omega$) is , given a family of test functions from $\Delta(\Omega)$ to $\mathbb{R}$, say $g \in G$, a procedure ϕ such that for each sequence X_m and each g

$$\frac{1}{n}\sum_{m=1}^{n} g(\phi_m)(X_m - \phi_m) \to 0$$

where the convergence is in $\mathbb{R}^\Omega$ and ϕ can be random (then the cv is a.s.).

In the basic framework the prediction belongs to a finite set (a grid V of $\Delta(\Omega)$) and g_v is the indicator of v.

The next result will apply for G, the set of Lipschitz functions and moreover ϕ will be deterministic.

Let V be a simplicial subdivision of $D' \subset \mathbb{R}^{\Omega}$ which is an $\varepsilon(L^1)$-neighborhood of $D = \Delta(\Omega)$.
For $p \in D'$ consider the barycentric decomposition

$$p = \sum_v W_v(p) v$$

where $W_v(p) \geq 0$, $\sum_v W_v(p) = 1$, the support of the sum is V_p, and $|p - v| \leq \varepsilon$ for $v \in V_p$.
Given a forecast ϕ with values in D let:

$$\mu_n(v) = \frac{1}{n} \sum_{m=1}^{n} W_v(\phi_m)(X_m - \phi_m)$$

be the error associated to the test function W_v for each $v \in V$.
Define a map ρ_n on V and then by linear interpolation on D' by

$$\rho_n(v) = v + \mu_n(v)$$

thus

$$\rho_n(p) = p + \sum_v W_v(p) \mu_n(v)$$

Claim: ρ_n is a continuous map from D' to itself.
The continuity is clear and for $v \in V$ one writes

$$\rho_n(v) = v + \frac{1}{n}\sum_{m=1}^{n} W_v(\phi_m)(X_m - \phi_m)$$

$$= (1 - \frac{1}{n}\sum_{m=1}^{n} W_v(\phi_m))v + \frac{1}{n}\sum_{m=1}^{n} W_v(\phi_m)(X_m + v - \phi_m)$$

and in the last term the coefficient is 0 if $|v - \phi_m| > \varepsilon$ which implies that the sum is a convex combination of v and points within ε of X_m that are thus in D', as well as the combination.

■

Define inductively ϕ_{n+1} to be a fixed point of ρ_n, in particular it satisfies

$$\sum_v W_v(\phi_{n+1})\mu_n(v) = 0.$$

Lemma

There exists C_2 such that

$$\sum_v \|\mu_n(v)\|^2 \leq \frac{C_2}{n}$$

Proof

Let $r_n(v) = n\mu_n(v) = \sum_{m=1}^n W_v(\phi_m)(X_m - \phi_m)$ so that

$$\|r_n(v)\|^2 = \|r_{n-1}(v)\|^2 + W_v(\phi_n)^2\|X_n - \phi_n\|^2 + 2W_v(\phi_n)\langle X_n - \phi_n, r_{n-1}(v)\rangle$$

Now the sum in v of the last term is 0 since it writes

$$\langle X_n - \phi_n, \sum_v W_v(\phi_n) r_{n-1}(v)\rangle.$$

For the second term one has $\|X - \phi\|^2$ uniformly bounded by some C_2 on D' and $\sum_v W_v(\phi_n)^2 \leq \sum_v W_v(\phi_n) = 1$ hence

$$\sum_v \|r_n(v)\|^2 \leq \sum_v \|r_{n-1}(v)\|^2 + C_2 \leq C_2 n$$

Consider now a L Lipschitz function g from D' to $[0,1]$. Define an approximation $\hat{g}$ through

$$\hat{g}(p) = \sum_v W_v(p) g(v)$$

and note that $|\hat{g}(p) - g(p)| \leq \varepsilon L$.
The error associated to g and the above forecast ϕ is

$$\mu_n[g] = \frac{1}{n} \sum_{m=1}^{n} g(\phi_m)(X_m - \phi_m).$$

Then $|\mu_n[g]| \leq |\mu_n[\hat{g}]| + \varepsilon C_1 L$ whith $|X - \phi| \leq C_1$ on D'. But

$$|\mu_n[\hat{g}]| = |\frac{1}{n} \sum_{m=1}^{n} \sum_v W_v(\phi_m) g(v)(X_m - \phi_m)| = |\sum_v g(v)\mu_n(v)|$$

$$\leq \sum_v |\mu_n(v)| \leq \sqrt{(\#V) \sum_v \|\mu_n(v)\|^2}.$$

Finally we obtain

$$|\mu_n[g]| \leq \sqrt{\frac{C_2 \# V}{n}} + \varepsilon C_1 L$$

hence given any positive η, choose ε small enough and then let n greater than some $N(\varepsilon)$ to get a bound of η.

To avoid forecasting in $D' \setminus D$ one projects ϕ on D by Π_D which is Lipschitz and satisfies $\|\Pi_D(p) - p\| \leq (\#\Omega)\varepsilon$.

Application to random calibration

Let V a simplicial subdivision of $\Delta(\Omega)$ and recall the asociated barycentric representation $p = \sum_v W_v(p)v$.
Given a deterministic forecast adapted to L Lipschitz functions as above consider the random forecast having values in V with law defined by the splitting above.
Then the error is on expectation

$$E_m = \sum_v W_v(\phi_m)(X_m - v)$$

which is within ε of

$$E_m = \sum_v W_v(\phi_m)(X_m - \phi_m)$$

since $W_v(\phi_m) = 0$ if $\|\phi_m - v\|$ exceeds ε.
When summing the errors on has a finite sum ($v \in V$) of errors adapted each to ϕ and a Lipschitz test function W_v.

Convergence to Nash equilibria

The random variable is the joint profile $s_m \in S$ of the players.
Each prediction using a deterministic procedure is a mixed profile say $x_m \in \Delta(S)$.
Given a smooth (ε-)best reply function for each player, this defines a profile of mixed strategies $y_m \in \prod_i \Delta(S^i)$.
Then, with probability one:
Believing Nash

$$\frac{1}{n}\sum_{m=1}^{n} d(x_m, NE^{\varepsilon}) \to 0$$

Playing Nash

$$\frac{1}{n}\sum_{m=1}^{n} d(y_m, NE^{\varepsilon}) \to 0$$

Merging

$$\frac{1}{n}\sum_{m=1}^{n} d(x_m, y_m) \to 0$$

Convergence is obtain by requiring all the players to use the same calibrated algorithm ϕ.

7. Hypothesis testing

We present here procedures that corresponds to a random search of an equilibrium profile.
A first approach is based on prediction of the behavior of the opponents and hypothesis testing, Foster and Young (2003).
Each player state variable has 3 components:
- the empirical frequency of the moves of the opponent during the last s periods
- an hypothesis on this variable
- a counting variable relevant to the mode of the player.

If the hypothesis is rejected, the player chooses a new one at random. Then for specific choices of the parameters convergence in probability to Nash equilibria will occur.

A more direct process can be described as follows:
Consider a δ-discretization of the set of mixed strategies $X = \prod_i X^i = \prod_i \Delta(S^i)$ denoted by $\{x_v; v \in V\}$. Given the payoff function G and $\varepsilon > 0$ at least one of the x_v is for δ small enough an ε-equilibrium of G.

Assume that each player i plays by large blocks L an i.i.d. strategy (in the grid), say x_v^i while occasionally testing all his moves in S^i. Given a tolerance bound $\eta > 0$, if one move s^i gives more than the average payoff the block $+\eta$, he chooses at random a new point in his grid. Otherwise he keeps playing x_v^i for another block.
One wants that the proportion of blocks played with $\{v \in V^*; x_v$ ε-equilibrium of $G\}$ approaches 1.

This model has been proposed by Foster and Young (2006) and improved by Germano and Lugosi (2007).
Note that this strategy is radically uncoupled, in the sense that not only it does not depend on the payoff function of the opponents but it does not depend on the knowledge of their moves. It is simply a function of the realized payoffs of the player.
Characteristics of this procedure are:
inertia (keep playing if there are small variations)
search (with positive probability experiment)

Note that the procedure generates a Markov chain on V.
Then one makes a precise analysis of the exit probability from state v in terms of ε as a function of x_v being an ε-equilibrium.

Results:
The players plays with probability $1-\varepsilon$ an ε-equilibrium at each stage n, for n large enough (2 player case; Foster and Young, 2006).
Almost sure convergence holds (I players, generic; Germano and Lugosi, 2007) adding:
experimentation (even when no positive gain) with small probability λ
localization via annealing
to obtain geometric mixing and the proof that the length of the search phase is negligible compare to the time spent at equilibrium.

8. Adaptive procedures

We consider here random processes corresponding to adaptive behavior in repeated interactions.
There are at least three different levels of information.

1) Knowing the fact that one plays a game; the payoff function $G^1 : \prod_i S^i \to \mathbb{R}$ is known (hence player 1 knows both $K = S^1$ and $L = S^{-1}$).
After each stage n the opponent 's move s_n^{-1} is announced; one deduces the stage vector outcome $U_n = G^1(., s_n^{-1})$.
One can then speak about "learning" in terms of predicting, after each observation, the opponent's behavior.
Note nevertheless that if the payoff of the opponent is unknown it is difficult to predict anything on a rational basis, except in special situations like facing the same random event : strategic experimentation (Bolton and Harris, Rosenberg, Solan and Vieille) .
ADAPTIVE/LEARNING PROCEDURE

2) Here the information is simply the vector U_n (one may face a sequence of different opponents in terms of strategies or payoffs) the only “stationarity " in the model is the fact that the outcome are bounded and the set of moves K is given.
One uses also this approach if the payoff is not linear with respect to the opponents’ move - so that empirical distribution of moves has no interpretation).
The knowledge of the move played (s_n^1) may be needed (in no-regret procedures) or not (fictitious play); the explanation is though the "procedure in law" properties.)
NO REGRET/COMPARISON PROCEDURE

3) Only the payoff $g_n = G^1(s_n)$ (the component k_n of U_n) is announced.
A first kind of procedure is “payoff-based" using the knowledge of the move s_n^1.
REINFORCEMENT PROCEDURE
A second kind constructs from the observation g_n (and the move played s_n^1 and its law) a pseudo vector $\tilde{U}_n$ and applies the previous procedure ii).
PSEUDO COMPARISON PROCEDURE

In most of the procedures the behavior of the player depends upon a parameter $z \in Z$.
At stage n, the state is z_{n-1} and the process is defined by two functions:
a **decision map** σ from Z to $\Delta(K)$ (the simplex on K) defining the law π_n of the current action k_n as a function of the parameter:

$$\pi_n = \sigma(z_{n-1})$$

and given the observation α_n of the player, after the play at stage n, an **updating rule** for the state variable, that depends upon the stage:

$$z_n = \Phi_n(z_{n-1}, \alpha_n).$$

Remark
Note that the decision map is stationary but that the updating rule may depend upon the stage.

Example 1: Fictious Play

The state space is usually the empirical distribution of actions of the opponents $z_n = \{z_n^j\}$ with $z_n^j = \bar{s}_n^j$ if $\alpha_n = s_n^{-1}$, but one can as well take $\alpha_n = U_n$, the vector payoff, then $z_n = \overline{U}_n$ is the average vector payoff thus satisfies:

$$z_n = \frac{(n-1)z_{n-1} + U_n}{n}$$

and

$$\sigma(z) \in br(z) \qquad \text{or} \qquad \sigma(z) = br^{\varepsilon}(z).$$

with $br(z) = \{x \in \Delta(K); \langle z, x - y \rangle \geq 0, \forall y \in \Delta(K)\}$ being the payoff-based (rather than strategy-based) best reply.

Example 2: Potential regret dynamics

Here $\alpha_n = U_n$ and

$$R_n = U_n - g_n \mathbf{1}$$

is the “regret vector" at stage n. The updating rule $z_n = \Phi_n(z_{n-1}, \alpha_n)$ is simply

$$z_n = \overline{R}_n.$$

Choose P to be a “potential function" for the negative orthant $D = \mathbb{R}^K_-$ and for $z \notin D$ let $\sigma(z)$ be proportional to $\nabla P(z)$.

Example 3: Cumulative proportional reinforcement

The observation α_n is only the stage payoff g_n (we assume all payoffs ≥ 1).

The updating rule is

$$z_n^k = z_{n-1}^k + g_n \, \mathbf{I}_{\{k_n = k\}}$$

and the decision map is $\sigma(z)$ proportional to the vector z.

There is an important literature on such reinforcement dynamics, see e.g. Beggs (2005), Börgers, Morales and Sarin (2004), Börgers and Sarin (1997), Hopkins (2002), Hopkins and Posch (2005), Laslier, Topol and Walliser (2001), Leslie and Collins (2005), Pemantle (2007), Posch (1997).

Note that these three procedures can be written as

$$z_n = \frac{(n-1)z_{n-1} + v_n}{n} \quad \text{or} \quad z_n - z_{n-1} = \frac{1}{n}[v_n - z_{n-1}].$$

where v_n is a random variable depending on the actions ℓ of the opponents and on the action k_n having distribution $\sigma(z_{n-1})$.
Write

$$v_n = E_{\pi_n}(v_n|z_1, \ldots, z_{n-1}) + [v_n - E_{\pi_n}(v_n|z_1, \ldots, z_{n-1})]$$

and define

$$S(z_{n-1}) = Co\{E_{\pi_n}(v_n|z_1, \ldots, z_{n-1}); \ell \in L\}$$

where Co stands for the convex hull and

$$W_n = v_n - E_{\pi_n}(v_n|z_1, \ldots, z_{n-1}).$$

Thus

$$z_n - z_{n-1} \in \frac{1}{n}[S(z_{n-1}) - z_{n-1} + W_n].$$

The related differential inclusion is

$$\dot{z} \in S(z) - z \tag{1}$$

and the process z_n is a Discrete Stochastic Approximation of (1).
For further results with explicit applications of this procedure see e.g. Hofbauer and Sandholm (2002), Leslie and Collins (2005), Benaïm, Hofbauer and Sorin (2006), Cominetti, Melo and Sorin (2010), Coucheney, Gaujal and Mertikopoulos (2014), Bravo (2015), Bravo and Faure (2015)...
In conclusion, a large class of adaptive dynamics can be expressed in discrete time as a random difference equation with vanishing step size. Information on the asymptotic behavior can then be obtained by studying the continuous time deterministic analog obtained as above.