

Dynamics in Games: Algorithms and Learning

Sylvain Sorin
sylvain.sorin@imj-prg.fr

Institut de Mathématiques Jussieu-PRG
Sorbonne Université, UPMC - Paris 6
CNRS UMR 7586

PGMO Course
February 2021

Abstract

Game theory studies interactions between agents with specific aims, be they rational actors, genes, or computers. This course is intended to provide the main mathematical concepts and tools used in game theory with a particular focus on their connections to learning and convex optimization. The first part of the course deals with the basic notions: value, (Nash and Wardrop) equilibria, correlated equilibria. We will give several dynamic proofs of the minmax theorem and describe the link with Blackwell's approachability. We will also study the connection with variational inequalities.

The second part will introduce no-regret properties in on-line learning and exhibit a family of unilateral procedures satisfying this property. When applied in a game framework we will study the consequences in terms of convergence (value, correlated equilibria). We will also compare discrete and continuous time approaches and their analog in convex optimization (projected gradient, mirror descent, dual averaging). Finally we will present the main tools of stochastic approximation that allow to deal with random trajectories generated by the players.

Part B

ALGORITHMS AND LEARNING

B.1 No-regret procedures (I) and applications

This section relies in part on the following :

Lectures on Dynamics in Games (2008) Université Paris 6, UPMC,
unpublished lectures notes.

Tutorial on learning (2015) *Stochastic Methods in Game Theory*,
IMS-NUS, Singapore.

1. No-regret procedures (I)

- 1.1 External regret
- 1.2 Internal regret
- 1.3 Calibrating
- 1.4 Extensions
- 1.5 Imperfect monitoring

2. Application to games

- 2.1 Global procedures
- 2.2 External consistency and Hannan's set
- 2.3 Internal consistency and correlated equilibria
- 2.4 From calibrating to correlated equilibrium
- 2.5 No convergence to Nash
- 2.6 Weak calibration and deterministic procedure
- 2.7 Adaptive procedures

No-regret procedures (I)

Unilateral procedures

We consider an agent acting in discrete time and facing an unknown environment.

At each stage n , she chooses k_n in a finite set K , then observes a **reward vector** $U_n \in \mathcal{U} = [-1, 1]^K$ and her payoff is the k_n^{th} component:

$$\omega_n = U_n^{k_n}.$$

We work in an adversarial framework where no assumption is done on the reward process that is a function of the past history

$$h_{n-1} = \{k_1, U_1, \dots, k_{n-1}, U_{n-1}\} \in H_{n-1}.$$

A strategy of the agent is a map σ from $H = \cup_{m=0}^{+\infty} H_m$ to $\Delta(K)$ (set of probabilities on K). $\sigma(h_{n-1})$ is her mixed move at stage n .

1. External regret

Introduce the **external regret** given $k \in K$ and $U \in \mathcal{U} \subset \mathbb{R}^K$ as the vector $R(k, U) \in \mathbb{R}^K$ defined by:

$$R(k, U)^\ell = U^\ell - U^k, \ell \in K.$$

The evaluation of the procedure σ is through the sequence of external regret vectors, where the external regret at stage n is $R_n = R(k_n, U_n)$ thus :

$$R_n^\ell = U_n^\ell - \omega_n, \ell \in K.$$

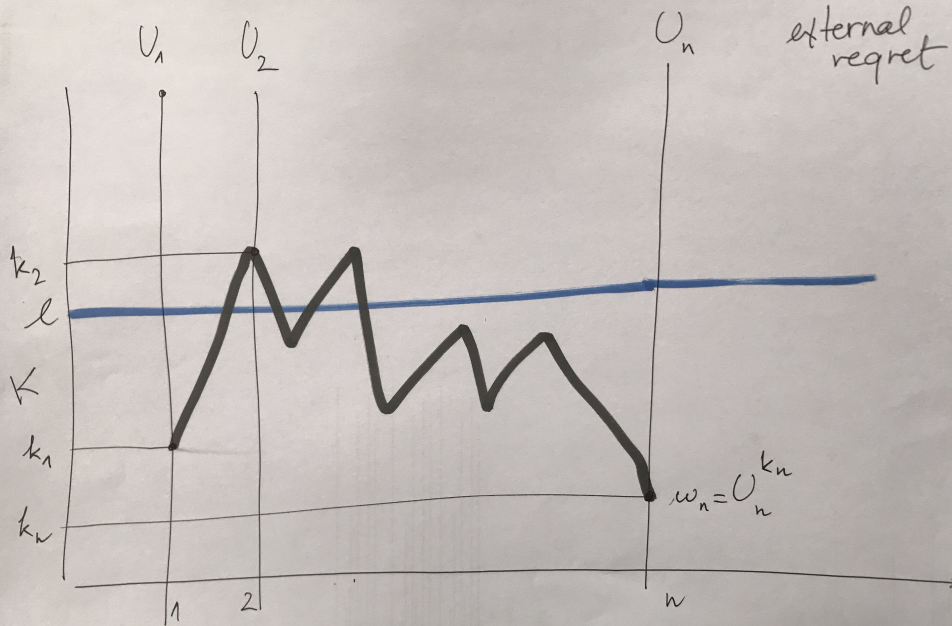
with $\omega_n = U_n^{k_n}$.

The average external regret vector at stage n is $\bar{R}_n = \frac{1}{n} \sum_{m=1}^n R_m$, thus :

$$\bar{R}_n^\ell = \bar{U}_n^\ell - \bar{\omega}_n, \ell \in K.$$

This compares the actual average payoff to the payoff corresponding to the choice of a constant component.

See Hannan, 1957 [31], Fudenberg and Levine, 1995 [28], Foster and Vohra, 1999 [27], ...



Definition 1.1

A strategy σ satisfies **external consistency** (EC) (or exhibits no **external regret**) if, for every process $\{U_m\} \in \mathcal{U}$:

$$\max_{k \in K} [\bar{R}_n^k]^+ \longrightarrow 0 \text{ a.s., as } n \rightarrow +\infty$$

or, equivalently :

$$\sum_{m=1}^n (U_m^k - \omega_m) \leq o(n), \quad \forall k \in K.$$

We prove the existence of a strategy satisfying (EC) by showing that the **negative orthant** $D = \mathbb{R}_-^K$ is approachable by the sequence of regret $\{R_n\}$.

Lemma 1.1 (Average lemma)

$\forall x \in \Delta(K), \forall U \in \mathcal{U} :$

$$\langle \mathbb{E}_x[R(\cdot, U)], x \rangle = 0.$$

Proof :

One has :

$$\mathbb{E}_x[R(\cdot, U)] = \sum_{k \in K} x_k R(k, U) = \sum_{k \in K} x_k (U - U^k \mathbf{1}) = U - \langle x, U \rangle \mathbf{1}$$

(where $\mathbf{1}$ is the K -vector of ones), thus $\langle x, \mathbb{E}_x[R(\cdot, U)] \rangle = 0$. ■

The strategy is as follows:

at stage n , if $\bar{R}_n^+ \neq 0$, let $\sigma(h_n)$ be proportional to this vector.

Proposition 1.1

σ satisfies (EC).

Proof:

Recall that $\Pi_D(Z) = Z^-$, $Z = Z^+ + Z^-$ and $\langle Z^-, Z^+ \rangle = 0$, $\forall Z \in \mathbb{R}^K$.

One has :

$$\langle \mathbb{E}(R_{n+1}|h_n) - \Pi_D(\bar{R}_n), \bar{R}_n - \Pi_D(\bar{R}_n) \rangle = 0 \quad (1)$$

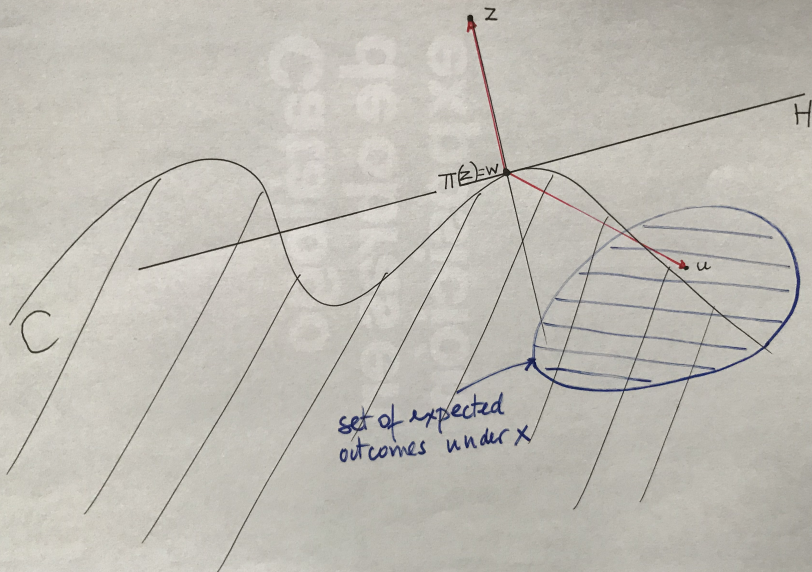
since $\langle \Pi_D(\bar{R}_n), \bar{R}_n - \Pi_D(\bar{R}_n) \rangle = 0$ and using Lemma 1.1:

$$\begin{aligned} \langle \mathbb{E}(R_{n+1}|h_n), \bar{R}_n - \Pi_D(\bar{R}_n) \rangle &= \langle \mathbb{E}(R_{n+1}|h_n), \bar{R}_n^+ \rangle \\ &\div \langle \mathbb{E}(R_{n+1}|h_n), \sigma(h_n) \rangle \\ &= \langle \mathbb{E}_x[R(\cdot, U_{n+1})], x \rangle, \quad \text{for } x = \sigma(h_n) \\ &= 0 \end{aligned}$$

Thus by (1) the **B**-set condition ($\langle z - w, u - w \rangle \leq 0$) is satisfied, so that D is **approachable**, hence $d(\bar{R}_n, \mathbb{R}_-^K)$ goes to 0 and $\max_{k \in K} [\bar{R}_n^k]^+ \rightarrow 0$.

B-set

$$\langle z-w, u-w \rangle \leq 0$$



2. Internal regret

The **internal regret** given (k, U) is the $K \times K$ matrix $S(k, U)$ with components: $S^{j\ell}(k, U) = (U^\ell - U^j) \mathbf{I}_{\{j=k\}}$.

The evaluation at stage n is the matrix $S_n = S(k_n, U_n)$ hence defined by:

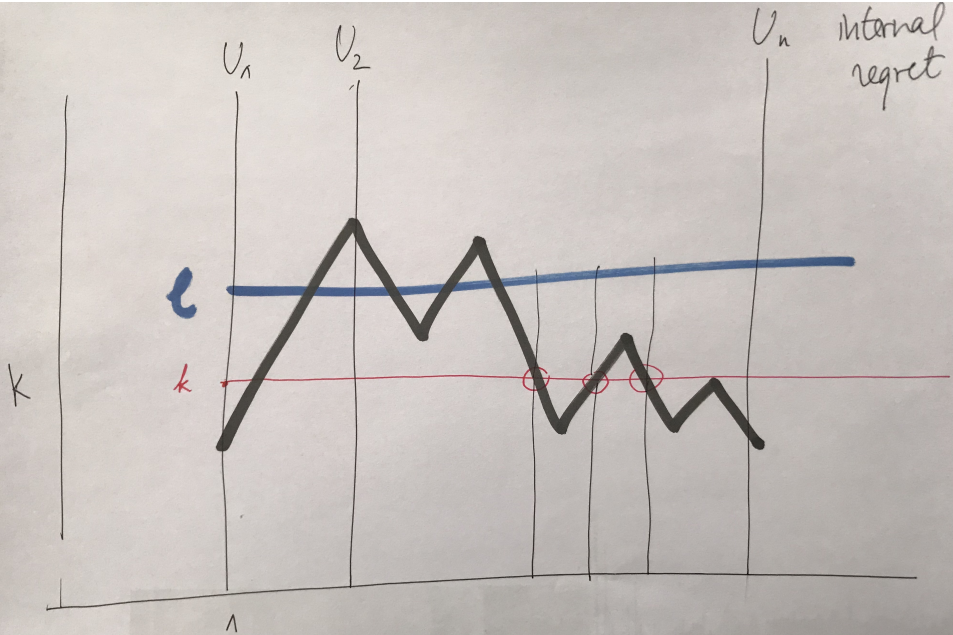
$$S_n^{k\ell} = \begin{cases} U_n^\ell - U_n^k & \text{for } k = k_n \\ 0 & \text{otherwise.} \end{cases}$$

Average **internal regret** matrix:

$$\bar{S}_n^{k\ell} = \frac{1}{n} \sum_{m=1, k_m=k}^n (U_m^\ell - U_m^k)$$

Comparison for each component k , of the average payoff obtained on the dates where k was played, to the payoff for an alternative choice ℓ .

See e.g. Foster and Vohra (1999), Fudenberg and Levine (1999).



Definition 1.2

A strategy σ satisfies *internal consistency* (IC) (or exhibits no *internal regret*) if, for every process $\{U_m\} \in \mathcal{U}$ and every couple k, ℓ :

$$[\bar{S}_n^{k\ell}]^+ \longrightarrow 0 \text{ a.s.,} \quad \text{as } n \rightarrow +\infty$$

Definition 1.3

Given a $K \times K$ real matrix A with nonnegative coefficients, let $\text{Inv}[A]$ be the non-empty set of *invariant measures* for A , namely vectors $\mu \in \Delta(K)$ satisfying:

$$\sum_{k \in K} \mu^k A^{k\ell} = \mu^\ell \sum_{k \in K} A^{\ell k}, \quad \forall \ell \in K.$$

(The existence follows from the existence of an invariant measure for a finite Markov chain - which is itself a consequence of the minmax theorem).

Lemma 1.2 (Average lemma bis)

Given $A \in \mathbb{R}_+^{K^2}$, let $\mu \in \text{Inv}[A]$ then:

$$\langle E_\mu[S(\cdot, U)], A \rangle = 0, \quad \forall U \in \mathcal{U}.$$

Proof :

$$\langle E_\mu[S(\cdot, U)], A \rangle = \sum_{k, \ell} A^{k\ell} \mu^k (U^\ell - U^k)$$

and the coefficient of each U^ℓ is

$$\sum_{k \in K} \mu^k A^{k\ell} - \mu^\ell \sum_{k \in K} A^{\ell k} = 0$$

■

To prove the existence of a strategy satisfying internal consistency, we show that $\Delta = \mathbb{R}_{-}^{K \times K}$ is approachable by the sequence of internal regret matrices $\{S_n\}$.

The **strategy** σ is as follows:

define at stage $n + 1$, if $B = \bar{S}_n^+ \neq 0$, $\sigma(h_n)$ to be an invariant measure of B .

Proposition 1.2

σ satisfies (IC).

Proof :

One has :

$$\langle \mathbb{E}(S_{n+1}|h_n) - \Pi_{\Delta}(\bar{S}_n), \bar{S}_n - \Pi_{\Delta}(\bar{S}_n) \rangle = 0$$

since again $\langle \Pi_{\Delta}(\bar{S}_n), \bar{S}_n - \Pi_{\Delta}(\bar{S}_n) \rangle = 0$ and using Lemma 1.2:

$$\begin{aligned} \langle \mathbb{E}(S_{n+1}|h_n), \bar{S}_n - \Pi_{\Delta}(\bar{S}_n) \rangle &= \langle \mathbb{E}(S_{n+1}|h_n), \bar{S}_n^+ \rangle \\ &= \langle \mathbb{E}(S_{n+1}|h_n), B \rangle \\ &= \langle \mathbb{E}_{\mu}[S(\cdot, U_{n+1})], B \rangle, \quad \text{for } \mu = \sigma(h_n) \\ &= 0 \end{aligned}$$

Then Δ is approachable hence $\max_{k,\ell} [\bar{S}_n^{k,\ell}]^+ \longrightarrow 0$.

3. Calibrating

One considers a sequence of random variables X_m with values in a finite set Ω (that will be written as a basis of \mathbb{R}^Ω).

Obviously any deterministic prediction algorithm ϕ_m of X_m - where the loss is measured by $\|X_m - \phi_m\|$ - will have a worst loss 1 and any random predictor a loss at least 1/2 (take $X_m = 1$ iff $\phi_m(1) \leq 1/2$).

We introduce here a predictor with values in a **finite discretization V of $D = \Delta(\Omega)$** with the following interpretation: “ $\phi_m = v$ ” means that the anticipated probability that $X_m = \omega$ (or $X_m^\omega = 1$) is v^ω .

Definition 1.4

ϕ is ε -calibrated if, for any $v \in V$:

$$\lim_{n \rightarrow +\infty} \frac{1}{n} \left\| \sum_{\{m \leq n, \phi_m = v\}} (X_m - v) \right\| \leq \varepsilon$$

Dawid, 1982 [21].

This means that if the average number of times v is predicted does not vanish, the average value of X_m on these dates is close to v .

More precisely let B_n^v the set of stages before n where v is announced, let N_n^v be its cardinal and $\bar{X}_n(v)$ the empirical average of X_m on these stages.

Then the condition writes :

$$\lim_{n \rightarrow +\infty} \frac{N_n^v}{n} \|\bar{X}_n(v) - v\| \leq \varepsilon, \quad \forall v \in V.$$

a) From internal consistency to calibrating

Foster and Vohra, 1997 [25].

Consider the online algorithm where the choice set of the forecaster is V and the outcome given v and X_m is given by :

$$U_m^v = \|X_m - v\|^2$$

(where we use the L^2 norm).

Given an internal consistent procedure ϕ one obtains (the outcome is here a loss):

$$\frac{1}{n} \sum_{m \in B_n^v} (U_m^v - U_m^w) \leq o(n), \quad \forall w \in V,$$

which is:

$$\frac{1}{n} \sum_{m \in B_n^v} (\|X_m - v\|^2 - \|X_m - w\|^2) \leq o(n), \quad \forall w \in V,$$

hence implies:

$$\frac{N_n^v}{n} (\|\bar{X}_n(v) - v\|^2 - \|\bar{X}_n(v) - w\|^2) \leq o(n), \quad \forall w \in V.$$

In particular by choosing a point $w \in V$ closest to $\bar{X}_n(v)$:

$$\frac{N_n^v}{n} (\|\bar{X}_n(v) - v\|^2) \leq \delta^2 + o(n)$$

where δ is the L^2 mesh of V , from which calibration follows. ■

b) From calibrating to approachability

Foster and Vohra, 1997 [25].

We use calibrating to prove approachability of non excludable convex sets.

Assume that C satisfies:

$$\forall y \in Y, \exists x \in X \text{ such that } xAy \in C.$$

Consider a δ -grid of Y defined by $\{y_v, v \in V\}$.

A stage has a **label** v if player 1 predicts y_v and then plays a mixed move x_v such that $x_v A y_v \in C$.

By using a calibrated procedure, the average move of player 2 on the stages with label v will be δ close to y_v .

By a standard martingale argument the average payoff on these stages will then be ε close to $x_v A y_v$ for δ small enough and n large enough.

Finally the total average payoff is a convex combination of such amounts hence is close to C by convexity.

There is a huge literature on the relations between approachability, no-regret and calibrating.

We recommend in particular:
Mannor and Stoltz, 2010 [53],
Abernethy, Bartlett and Hazan, 2011 [1],
Perchet, 2014 [59].

4. Extensions

1. Conditional expectation

Recall that the total regret at stage n that the agent wants to control is:

$$\sum_{m=1}^n U_m^k - \omega_m, \quad k \in K$$

where $\omega_m = U_m^{k_m}$ is the random payoff at stage m .

Let $x_m \in \Delta(K)$ be the strategy of the player at stage m , then

$$\mathbb{E}(\omega_m | h_{m-1}) = \langle U_m, x_m \rangle$$

so that $\omega_m - \langle U_m, x_m \rangle$ is a **bounded martingale difference**.

Hoeffding-Azuma's concentration inequality, [5], [39], for a process $\{Z_n\}$ of martingale differences with $|Z_n| \leq L$ states that:

$$\mathbb{P}\{|\bar{Z}_n| \geq \varepsilon\} \leq 2 \exp\left(-\frac{n \varepsilon^2}{2L^2}\right).$$

Hence the average difference between the payoff and its conditional expectation is controlled.

Thus we will also study quantities of the form:

$$\sum_{m=1}^n U_m^k - \langle U_m, x_m \rangle, \quad k \in K.$$

or equivalently, because of the linearity:

$$\sum_{m=1}^n \langle U_m, x \rangle - \langle U_m, x_m \rangle, \quad x \in \Delta(K).$$

This will be the notion of regret used in section B2.

Similarly the internal no-regret condition becomes:

$$\sum_{m=1}^n x_m^i [U_m^j - U_m^i] \leq o(n), \quad \forall i, j \in K.$$

2. Procedures in law

Assume that the actual move k_n is not observed and define a pseudo-process \tilde{R} defined through the conditional expected regret:

$$R_n = U_n - \omega_n \mathbf{1}, \quad \tilde{R}_n = U_n - \langle U_n, x_n \rangle \mathbf{1}$$

and introduce the associated strategy $\tilde{\sigma}$.

Then consistency holds both for the pseudo and the realized processes under $\tilde{\sigma}$, Benaim, Hofbauer and Sorin, 2006 [8].

3. Experts and generalized consistency

External consistency can be considered as a robustness property of σ facing a given finite family of “external” experts using procedures $\phi \in \Phi$:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \left[\sum_{m=0}^n \langle \phi_m - x_m, U_m \rangle \right]^+ = 0, \quad \forall \phi \in \Phi.$$

The typical case corresponds to a constant choice : $\phi = k$ and $\Phi = K$.

In general “ k ” will correspond to the (random) move of expert k , that the player follows with probability x_m^k at stage m .

U_m^k has to be understood as the payoff to expert k at stage m .

Internal consistency corresponds to experts adjusting their behavior to the one of the predictor.

4. From external to internal consistency

We describe briefly two ways of getting (IC) procedures by using combinations of (EC) procedures running on adequate datas.

1) We follow Stoltz and Lugosi, 2005 [64].

Consider a family $\psi^{ij}, (i,j) \in M = K \times K$ of experts and θ an algorithm that satisfies external consistency with respect to this family.

Define σ inductively as follows.

Given some element $p \in \Delta(K)$, let $p(ij)$ be the vector obtained by moving p^i from the i^{th} to the j^{th} component of p .

Let $q_{n+1}(p)$ be the distribution on $\Delta(K)$ induced by the composition of θ at stage $n+1$ given the history h_n , which is a probability on M and the behavior $\psi^{ij}(h_n) = p(ij)$ of the experts.

Assume that the map $p \mapsto q_{n+1}(p)$ is continuous and let \hat{p}_{n+1} be a fixed point which defines $\sigma(h_n) = x_{n+1}$.

The fact that σ is an incarnation of θ implies that it performs well facing any ψ^{ij} hence:

$$\left[\sum_{m=0}^n \langle \psi_m^{ij} - x_m, U_m \rangle \right] \leq o(n), \quad \forall i, j$$

which is:

$$\left[\sum_{m=0}^n \langle \hat{p}(ij)_m - \hat{p}_m, U_m \rangle \right] \leq o(n), \quad \forall i, j$$

hence:

$$\left[\sum_{m=0}^n \hat{p}_m^i (U_m^j - U_m^i) \right] \leq o(n), \quad \forall i, j$$

and this is the internal consistency condition.

2) We follow Blum and Mansour, 2007 [10].

Consider K parallel algorithms $\{\phi[k]; k \in K\}$ having no external regret, that generates each a (row) vector $Q[k] \in \Delta(K)$ then define σ by the invariant measure p satisfying:


$$p = pQ.$$

Given the outcome $U \in \mathbb{R}^K$, add $p^k U$ to the entry of algorithm $\phi[k]$. Expressing the fact that $\phi[k]$ satisfies no external regret gives, at stage m , for all $j \in K$:

$$\left[\sum_{m=0}^n p_m^k U_m^j - \langle Q[k]_m, p_m^k U_m \rangle \right] \leq o(n)$$

Note that $\sum_k \langle Q[k]_m, p_m^k U_m \rangle = \sum_k \langle p_m^k Q[k]_m, U_m \rangle = \langle p_m, U_m \rangle$, hence by summing over k , for any function $L : K \mapsto K$, corresponding to a perturbation σ_L of σ with $j = L(k)$ the difference between the performances of σ_L and σ will satisfy as well :

$$\left[\sum_{m=0}^n \sum_k p_m^k U_m^{L(k)} - \langle p_m, U_m \rangle \right] \leq o(n).$$

This is the internal consistency for “swap experts”. 

5. Large range

Consider an even larger set of experts that are allowed (in addition to be adapted to the past history) to choose their actions and to be active as a function of the choice of the predictor.

Explicitly, every expert $s \in S$ (finite) is characterized, at stage m , by :

- a choice function $f_m^s : K \rightarrow K$
 - an activity function $\tau_m^s : K \rightarrow [0, 1]$.
- both conditional to the past.

Given a predictor ϕ which prediction at stage m has a law p_m the regret facing s is :

$$r_m^s = \sum_k p_m^k \tau_m^s(k) [U_m^{f_m^s(k)} - U_m^k]$$

We assume that the functions f^s, τ^s are known by the predictor.
Then there exists a **consistent procedure**.

Lehrer, 2003 [48]; Cesa-Bianchi and Lugosi, 2006 [17]; Blum and Mansour, 2007 [10].

6. Bandit framework

This is the case where, given the move k and the vector U , the only information to the agent is the realization $\omega = U^k$ (the vector U is not announced).

Define the **pseudo regret vector** at each stage n by:

$$\hat{U}_n^k = \frac{\omega_n}{\sigma_n^k} \mathbf{1}_{\{k_n=k\}}$$

and note that it is an unbiased estimator of the true regret.

To keep the outcome bounded one may have to consider a slight perturbation of the strategy but the same asymptotic properties hold, see Auer, Cesa-Bianchi, Freund and Shapire, 1995 [2], 2002 [3].

For recent advances and precise evaluations of the convergence rates, see Bubeck and Cesa-Bianchi, 2012 [14].

5. Imperfect monitoring

1. The model

Consider a finite zero-sum two person repeated game defined by a function G from $I \times J$ to \mathbb{R} .

In addition there is a finite signal set S and a map M from $I \times J$ to $\Delta(S)$. At each stage n , given a profile of moves (i_n, j_n) , a signal s_n with law $M(i_n, j_n) \in \Delta(S)$ is sent to player 1 and this is his only information.

Player 2 is Nature and knows the full history.

Given $y \in Y = \Delta(J)$, let $M(i, y) = \sum_j y^j M(i, j) \in \Delta(S)$ be the linear extension and denote by $m(y) \in F = \Delta(S)^I = \{ M(i, y), i \in I \}$ be the **flag** induced by y .

This is the maximal information that player 1 can obtain if player 2 uses y i.i.d..

This model appears in the theory of repeated games with incomplete information, Kohlberg, 1975 [46], Mertens, Sorin and Zamir, 2015 [54], and the analysis of external regret in this framework is due to Rustichini, 1999 [63].

Given a n -stage play, the average flag is $\bar{\mu}_n \in F$, where $\mu_r = m(j_r)$ (hence $\bar{\mu}_n = m(\bar{y}_n)$) and the evaluation of player 1 is $d(\bar{\mu}_n)$ where:

$$d(\mu) = \max_{x \in \Delta(I)} \min_{y \in \Delta(J); m(y) = \mu} G(x, y).$$

Note that in general best replies are not pure.

The **external regret** is then defined by:

$$r_n = d(\bar{\mu}_n) - \bar{G}_n$$

2. Internal regret

Related results are in Cesa-Bianchi, Lugosi and Stoltz, 2006 [18], Lehrer and Solan, 2015 [49], Lugosi, Mannor and Stoltz, 2008 [51], Perchet, 2009 [56].

To specify a notion of internal consistency, we use the regularity of the model to introduce for each $\varepsilon > 0$ finite discretizations

$(\mu[\ell], x[\ell]; \ell \in L)$ of F and X , such that there exists $\delta > 0, \eta > 0$ with:

- the set of flags is covered by balls $B(\mu(\ell), \delta), \ell \in L$,
- for any $\mu \in B(\mu(\ell), \delta)$ and $x \in B(x(\ell), \eta)$, x is a ε -best reply to μ for the evaluation $d(\mu)$.

One can now introduce the vector of **internal regret**.

Let $A_n[\ell]$ be the set of stages up to n where player 1 uses $x[\ell]$ and $N_n[\ell]$ its cardinality.

$\bar{\mu}_n[\ell]$, resp. $\bar{G}_n[\ell]$, are the corresponding average flag, resp. payoff.
Then let:

$$R_n[\ell] = d(\bar{\mu}_n[\ell]) - \bar{G}_n[\ell], \quad \ell \in L$$

and define ε —**internal consistency** as:

$$\limsup_{n \rightarrow +\infty} \frac{N_n[\ell]}{n} [R_n[\ell] - \varepsilon]^+ \rightarrow 0, \quad \forall \ell \in L.$$

Proposition 1.3 (Perchet, 2009 [56])

There exist ε –internal consistent strategies.

Proof :

Assume first that player 1 is informed of the vector of signals (indexed by I) at each stage.

He can use a calibrated strategy associated to L such that, at a stage with label ℓ , he “predicts ” $\mu[\ell]$ and plays $x[\ell]$.

Then asymptotically on these stages (if their frequency is large enough) his prediction will be correct, his average moves closed to $x[\ell]$ hence the average regret $R_n[\ell]$ small.

To reduce the analysis to the previous case, where the flag is known, one constructs an estimator of the flag via a pertubation of the strategy (like in the bandit framework above). ■

Using a specific discretization trough Laguerre diagrams allows to get a speed of convergence of $O(n^{-1/3})$ which is optimal, Perchet, 2011 [58].

3. Approachability with random signals

The framework is as above except that G is now from $I \times J$ to \mathbb{R}^K . $G(x, y)$ is the multilinear extension to $X = \Delta(I) \times Y = \Delta(J)$.

Let

$$P(x, \mu) = \{G(x, y); m(y) = \mu, y \in Y\} \subset \mathbb{R}^K$$

be the set of outcomes compatible with the strategy $x \in X$ and the flag μ (that could be generated in expectation by some y).

Proposition 1.4 (Perchet 2011 [57])

A closed convex set $C \subset \mathbb{R}^K$ is approachable (by player 1) if and only if:

$$\forall \mu \in m(Y), \exists x \in X \quad \text{such that} \quad P(x, \mu) \subset C.$$

Note that this is exactly Blackwell's condition in the full monitoring case, where the signal to the agent is the move of the opponent.

Proof :

1) Assume that the condition holds.

Then for each $\varepsilon > 0$ one constructs as above a finite family

$\{\mu[\ell], x[\ell], \ell \in L\}$ with $P(x[\ell], \mu[\ell]) \subset C$.

A calibrated strategy associated to this set L , such that $x[\ell]$ is played when $\mu[\ell]$ is predicted, will induce on average, on stages with label ℓ , a payoff near C .

One then uses the convexity of C to deduce approachability.

2) For the converse, if there exists a signal μ_0 such that

$$\forall x \in X, \exists y = y(x) \in Y, m(y) = \mu_0 \quad \text{and} \quad G(x, y) \notin C$$

one can assume $d(G(x, y), C) \geq \delta > 0$ by compactness.

Given σ strategy of Player 1 in the n -stage game let:

$$z_n = \mathbb{E}_{\sigma, \mu_0} [\bar{i}_n]$$

be the expectation of the average move of Player 1 facing signals with distribution μ_0 at each stage.

Then, let τ be $y(z_n)$ i.i.d. and by convexity:

$$\mathbb{E}_{\sigma, \tau} [d(\bar{g}_n, C)] \geq d(G(z_n, y(z_n)), C) \geq \delta > 0$$



In addition there are convex sets that are neither approachable nor excludable.’

Recent results with optimal rates of convergence are in Kwon and Perchet (2017 [47])

For extensions to games with payoff correspondence see Mannor, Perchet and Stoltz, 2014 [52].

The natural extension is to consider games on signal's distributions and to check approachability at this level.

A first approach was in Kohlberg, 1975 [46] and this program has been developed in Perchet and Quincampoix, 2014 [61], 2019 [62].

It is one of the main direction of research.

Application to games

1. Global procedures

Let G a finite game in strategic form.

There are finitely many players $i = 1, 2, \dots, I$.

S^i is the finite set of moves of player i , $S = \prod_i S^i$, and $Z = \Delta(S)$ is the set of probabilities on S (correlated moves).

The payoff is $g : S \rightarrow \mathbb{R}^I$.

We will consider repeated interaction in discrete time where at each stage the players observe the actions of their opponents.

There is an important literature on this topic.

Basic references

Cesa-Bianchi N. and G. Lugosi (2006) *Prediction, Learning and Games*, Cambridge University Press.

Fudenberg D. and D.K. Levine (1998) *Theory of Learning in Games*, M.I.T. Press.

Fudenberg D. and D. K. Levine (2009) Learning and equilibrium, *Annual Review of Economics*, **1**, 385-420.

Hart S. and A. Mas Colell (2013) *Simple Adaptive Strategies: From Regret-Matching to Uncoupled Dynamics*, World Scientific Publishing.

Young P. (1998) *Theory of Learning in Games*, M.I.T. Press.

see also a recent survey: Faure, Gaillard, Gaujal and Perchet, 2015 [22].

We want evaluate the joint impact on the play of the prescribed behavior of the players (no-regret).

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), $L = \prod_{i \neq 1} S^i$, $Y = \Delta(L)$ (correlated moves of Player 1's opponents) and $Z = \Delta(K \times L)$ (correlated distributions).

$F : S \rightarrow \mathbb{R}$ 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 .

$R(z)$ 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 her payoff using a given move k to her actual payoff, assuming the other players' behavior, z^{-1} , given.)

Let us recall:

Definition 2.1 (Hannan, 1957, [31])

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

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

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

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

will converge to H .

Proposition 2.1

If Player 1 follows any (EC) 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:

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

which gives :

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

and this expression is:

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



Alternative proof: Blackwell, 1956 [9].

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

D_1 is the convex set:

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

Theorem 2.1

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 $\phi(s^1, \theta) \in D_1$. ■

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

Recall that one defines similarly H^i for each player and $H = \cap_i H^i$ which is the global Hannan set.

Proposition 2.2

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

Note that no coordination is required and different (EC) procedures can be used.

3. Internal consistency and correlated equilibria

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 $(z | j)$ denote the conditional probability on L induced by z given $j \in K$ and z^1 the marginal on K , then:

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

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

Definition 2.2

The set of distributions satisfying 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 2.3

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 equilibria distributions of the game G is defined by

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

Hence one has :

Lemma 2.1

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

Thus we obtain:

Proposition 2.4

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

Note that this provides an alternative proof of existence of correlated equilibrium through the existence of (IC) procedures.

For an alternative algorithm (but defined jointly) sharing this property, see Hart and Mas Colell, 2001 [35] and Cahn, 2004 [15].

An extension to the compact case is achieved in Stoltz and Lugosi, 2007 [65].

4. From calibrating to correlated equilibrium

Foster and Vohra, 1997 [25].

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 an (approximate) best reply to $z|k$: hence z is (approximately) in C^1 . If all players use calibrated strategies the empirical average frequency of moves 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 [37].

Similarly there are no learning process with finite memory such that the stage behavior will converge to Nash equilibrium: Hart and Mas-Colell, 2005 [38].

Similar results were obtained for MAD dynamics, Hofbauer and Swinkels, 1995 [41], see also Young, 2004 [66].

6. Weak calibration and deterministic procedure

We follow Kakade and Foster, 2004 [44], 2008 [45].

Weak calibration

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

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

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

In the standard framework the prediction ϕ belongs to a finite set (a grid V of $\Delta(\Omega)$) and γ_v is the indicator of $v \in V$.

The next result will apply for Γ , 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 ε -neighborhood of $D = \Delta(\Omega)$ (for the L^1 norm).

For $p \in D'$ consider the barycentric decomposition:

$$p = \sum_{v \in V} W_v(p)v$$

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 evaluation associated to the test function W_v for each $v \in V$.

Define a map ρ_n on V by:

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

and then by linear interpolation on D' 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:

$$\begin{aligned} \rho_n(v) &= v + \frac{1}{n} \sum_{m=1}^n W_v(\phi_m)(X_m - \phi_m) \\ &= \left(1 - \frac{1}{n} \sum_{m=1}^n W_v(\phi_m)\right)v + \frac{1}{n} \sum_{m=1}^n W_v(\phi_m)(X_m + v - \phi_m) \end{aligned}$$

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 , 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 2.2

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 over $v \in V$ 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 :

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

by induction.

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

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

and note that $|\hat{\gamma}(p) - \gamma(p)| \leq \varepsilon L$.

The evaluation associated to γ and the above forecast ϕ is:

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

Then $|\mu_n[\gamma]| \leq |\mu_n[\hat{\gamma}]| + \varepsilon C_1 L$, while $|X - \phi| \leq C_1$ on D' . But :

$$\begin{aligned} |\mu_n[\hat{\gamma}]| &= \left| \frac{1}{n} \sum_{m=1}^n \sum_v W_v(\phi_m) \gamma(v) (X_m - \phi_m) \right| = \left| \sum_v \gamma(v) \mu_n(v) \right| \\ &\leq \sum_v |\mu_n(v)| \leq \sqrt{(\#V) \sum_v \|\mu_n(v)\|^2}. \end{aligned}$$

Finally one obtains:

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

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 associated 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 evaluation is in 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 evaluations one obtains a finite sum ($v \in V$) of evaluations 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 one shows that with probability one:

Believing Nash:

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

Playing Nash:

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

Merging:

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

The previous convergence result implies that on the stages where x_m is predicted (and where using a martingale argument y_m is realized) the average distribution is also x_m , hence the fixed point and the equilibrium condition.

Convergence to Nash equilibria is obtained by requiring all the players to use the same calibrated algorithm ϕ .

Recent extensions are presented in Foster and Hart, 2018 [23].

7. 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 $G^1 : \prod_i S^i \rightarrow \mathbb{R}$ 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; Player 1 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".

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 of this fact is through 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 2).

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 x_n of the current action k_n as a function of the parameter:

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

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

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

Remark

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

Example 1: Fictitious 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 $z_n = \bar{U}_n$ is the average vector payoff thus satisfies:

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

and

$$\sigma(z) \in br(z) \quad \text{or} \quad \sigma(z) = br^\varepsilon(z),$$

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 $\nabla P(z)$, Hart and Mas-Colell, 2003 [36], Cesa-Bianchi and Lugosi, 2003 [16].

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 [6], Börgers and Sarin, 1997 [11], Hopkins, 2002 [42], Hopkins and Posch, 2005 [43], Pemantle, 2007 [55], and the references therein.

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 $x_n = \sigma(z_{n-1})$. Write:

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

and define:

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

where Co stands for the convex hull and:

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

Thus:

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

The related differential inclusion is:

$$\dot{z} \in S(z) - z \quad (2)$$







and the process z_n is a Discrete Stochastic Approximation of (2).

For further results with explicit applications of this procedure see e.g. Hofbauer and Sandholm, 2002 [40], Leslie and Collins, 2005 [50], Benaïm, Hofbauer and Sorin, 2006 [8], Cominetti, Melo and Sorin, 2010 [19], Coucheney, Gaujal and Mertikopoulos, 2015 [20], Bravo, 2015 [12], Bravo and Faure, 2015 [13]...







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.








-  Abernethy J., P. L. Bartlett and E. Hazan (2011) Blackwell approachability and no-regret learning are equivalent, *JMLR Workshop and Conference Proceedings*, **19**, 27-46.
-  Auer, P., N. Cesa-Bianchi, Y. Freund and R.E. Schapire (1995) Gambling in a rigged casino: the adversarial multi-armed bandit problem, *Proceedings of the 36 th Annual Symposium on Foundations of Computer Science*, 322-331.
-  Auer P., Cesa-Bianchi N., Freund Y. and R.E. Shapire (2002) The nonstochastic multiarmed bandit problem, *SIAM J. Comput.*, **32**, 48-77.
-  Aumann R. J. and M. Maschler (1995) *Repeated Games with Incomplete Information*, MIT Press.
-  Azuma K. (1967) Weighted sum of certain dependent random variables, *Tohoku Math. J.*, **68**, 357-367.
-  Beggs A. (2005) On the convergence of reinforcement learning, *Journal of Economic Theory*, **122**, 1-36.







-  Benaim M., J. Hofbauer and S. Sorin (2005) Stochastic approximations and differential inclusions, *Siam J. Control Optim*, **44**, 328-348.
-  Benaim M., J. Hofbauer and S. Sorin (2006) Stochastic approximations and differential inclusions. Part II: applications, *Mathematics of Operations Research*, **31**, 673-695.
-  Blackwell D. (1956) Controlled random walks, *Proceedings of the International Congress of Mathematicians, 1954, Amsterdam*, Erven P. Noordhoff N.V., North-Holland, **III**, 336-338.
-  Blum A. and Y. Mansour (2007) From external to internal regret, *Journal of Machine Learning Research*, **8**, 1307-1324.
-  Börgers T. and R. Sarin (1997) Learning through reinforcement and replicator dynamics, *Journal of Economic Theory*, **77**, 1-14.
-  Bravo M.(2016) An adjusted payoff-based procedure for normal form games, *Mathematics of Operations Research*, **41**, 1469-1483.







-  Bravo M. and M. Faure (2015) Reinforcement learning with restrictions on the action set, *Siam J Control Optim*, **53**, 287-312.
-  Bubeck S. and N. Cesa-Bianchi (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems, *Fondations and Trends in Machine Learning*, **5**, 1-122.
-  Cahn A. (2004) General procedures leading to correlated equilibria, *International Journal of Game Theory*, **33**, 21-40.
-  Cesa-Bianchi N. and G. Lugosi (2003) Potential-based algorithms in on-line prediction and game theory, *Machine Learning*, **51**, 239-261.
-  Cesa-Bianchi N. and G. Lugosi (2006) *Prediction, Learning and Games*, Cambridge University Press.
-  Cesa-Bianchi N., G. Lugosi and G. Stoltz (2006) Regret minimization under partial monitoring, *Mathematics of Operations Research*, **31**, 562-580.






-  Cominetti R., E. Melo and S. Sorin (2010) A payoff-based learning procedure and its application to traffic games, *Games and Economic Behavior*, **70**, 71-83.
-  Coucheney P., B. Gaujal and P. Mertikopoulos (2015) Penalty-regulated dynamics and robust learning procedures in games , *Mathematics of Operations Research*, , article in advance.
-  Dawid A. (1982) “The well-calibrated Bayesian”, *Journal of the American Statistical Association*, **77**, 605-613.
-  Faure M., P. Gaillard, B. Gaujal and V. Perchet (2015) Online learning and game theory. A quick overview with recent results and applications, *ESAIM: Proceedings and Surveys*, **51**, 246-271.
-  Foster D. and S. Hart (2018) Smooth calibration, leaky forecasts, finite recall, and Nash dynamics, *Games and Economic Behavior*, **109**, 271-293.
-  Foster D. and R. Vohra (1993) A randomization rule for selecting forecasts, *Operations Research*, **41**, 704-707.






-  Foster D. and R. Vohra (1997) Calibrated learning and correlated equilibria, *Games and Economic Behavior*, **21**, 40-55.
-  Foster D. and R. Vohra (1998) Asymptotic calibration, *Biometrika*, **85**, 379-390.
-  Foster D. and R. Vohra (1999) Regret in the on-line decision problem, *Games and Economic Behavior*, **29**, 7-35.
-  Fudenberg D. and D. K. Levine (1995) Consistency and cautious fictitious play, *Journal of Economic Dynamics and Control*, **19**, 1065-1089.
-  Fudenberg D. and D. K. Levine (1998) *The Theory of Learning in Games*, MIT Press.
-  Fudenberg D. and D. K. Levine (1999) Conditional universal consistency, *Games and Economic Behavior*, **29**, 104-130.
-  Hannan J. (1957) Approximation to Bayes risk in repeated plays, *Contributions to the Theory of Games, III*, Drescher M., A.W. Tucker and P. Wolfe eds., Princeton University Press, 97-139.






-  Hart S. (2005) Adaptive heuristics, *Econometrica*, **73**, 1401-1430.
-  Hart S. and A. Mas-Colell (2000) A simple adaptive procedure leading to correlated equilibrium, *Econometrica*, **68**, 1127-1150.
-  Hart S. and A. Mas-Colell (2001) A general class of adaptive strategies, *Journal of Economic Theory*, **98**, 26-54.
-  Hart S. and A. Mas-Colell (2001) A reinforcement procedure leading to correlated equilibria, in *Economic Essays: A Festschrift for W. Hildenbrandt* ed. by G. Debreu, W. Neuefeind and W. Trockel, Springer, 181-200.
-  Hart S. and A. Mas-Colell (2003) Regret-based continuous time dynamics, *Games and Economic Behavior*, **45**, 375-394.
-  Hart S. and A. Mas-Colell (2003) Uncoupled dynamics do not lead to Nash equilibrium, *Am. Econ. Rev.*, **93**, 1830-1836.
-  S. Hart and A. Mas-Colell (2006) Stochastic uncoupled dynamics and Nash equilibrium, *Games and Economic Behavior*, **57**, 286-303.

-  Hoeffding W. (1963) Probability inequalities for sum of bounded random variables, *J. Amer. Statist. Assoc.*, **58**, 13-30.
-  Hofbauer J. and W. H. Sandholm (2002) On the global convergence of stochastic fictitious play, *Econometrica*, **70**, 2265-2294.
-  Hofbauer J and J. Swinkels (1996) A universal Shapley example, Preprint.
-  Hopkins Ed. (2002) Two competing models of how people learn in games, *Econometrica*, **70**, 2141-2166.
-  Hopkins Ed. and M. Posch (2005) Attainability of boundary points under reinforcement learning, *Games and Economic Behavior*, **53**, 110-125.
-  Kakade S.M. and D. P. Foster (2004) Deterministic calibration and Nash equilibrium, in John Shawe-Taylor, Yoram Singer (Eds.): *Learning Theory, 17th Annual Conference on Learning Theory, COLT 2004*, Lecture Notes in Computer Science, **3120**, Springer, 33-48.

-  Kakade S. and D. P. Foster (2008) Deterministic calibration and Nash equilibrium, *Journal of Computer and System Sciences*, **74**, 115-130.
-  Kohlberg E. (1975) Optimal strategies in repeated games with incomplete information, *Int. J. Game Theory*, **4**, 7-24.
-  Kwon J. and V. Perchet (2017) Online learning and Blackwell approachability with partial monitoring: optimal convergence rates, *Proceedings of Machine Learning Research (AISTATS 2017)*, **54**, 604-613.
-  Lehrer E. (2003) A wide range no-regret theorem, *Games and Economic Behavior*, **42**, 101-115.
-  Lehrer E. and E. Solan (2016) A general internal regret-free strategy, *Dynamic Games and Applications*, **6**, 112-138.
-  Leslie and Collins (2016) Generalised weakened fictitious play, *Games and Economic Behavior*, **56**, 285-298.

-  Lugosi G., S. Mannor and G. Stoltz (2008) Strategies for prediction under imperfect monitoring, *Mathematics of Operations Research*, **33**, 513-528.
-  Mannor S., V. Perchet and G. Stoltz (2014) Set-valued approachability and online learning with partial monitoring, *Journal of Machine Learning Research*, **15**, 3247-3295.
-  Mannor S. and G. Stoltz (2010) A geometric proof of calibration, *Mathematics of Operations Research*, **35**, 721-727.
-  Mertens, J.-F, S. Sorin and S. Zamir (2015), *Repeated Games*, Cambridge University Press.
-  Pemantle R. (2007) A survey of random processes with reinforcement, *Probability Surveys*, **4**, 1-79.
-  Perchet V. (2009) Calibration and internal no-regret with random signals, *Proceedings of the 20th International Conference on Algorithmic Learning Theory*, LNAI **5809**, 68-82.

-  Perchet V. (2011a) Approachability of convex sets in games with partial monitoring, *Journal of Optimization Theory and Applications*, **149**, 665-677.
-  Perchet V. (2011b) Internal regret with partial monitoring calibration-based optimal algorithms, *Journal of Machine Learning Research*, **12**, 1893-1921.
-  Perchet V. (2014) Approachability, regret and calibration: implications and equivalences, *Journal of Dynamics and Games*, **1**, 181-254.
-  Perchet V. (2015) Exponential weight approachability, applications to calibration and regret minimization, *Dynamic Games and Applications*, **5**, 136-153.
-  Perchet V. and M. Quincampoix (2014) On a unified framework for approachability with full or partial monitoring, *Mathematics of Operations Research*, **40**, 596-610.

-  Perchet V. and M. Quincampoix (2019) A differential game on Wasserstein space, application to weak approachability with partial monitoring, *Journal of Dynamics and Games*, **6**, 65-85.
-  Rustichini A. (1999) Minimizing regret: the general case, *Games and Economic Behavior*, **29**, 224-243.
-  Stoltz G. and G. Lugosi (2005) Internal regret in on-line portfolio selection, *Machine Learning*, **59**, 125-159.
-  Stoltz G. and G. Lugosi (2007) Learning correlated equilibria in games with compact sets of strategies, *Games and Economic Behavior*, **59**, 187-208.
-  Young P. (2004) *Strategic Learning and Its Limits*, Oxford U. P. .