SIAM J. CONTROL OPTIM.
Vol. 53, No. 1, pp. 287–312

REINFORCEMENT LEARNING WITH RESTRICTIONS ON THE ACTION SET*

MARIO BRAVO† AND MATHIEU FAURE‡

Abstract. Consider a two-player normal-form game repeated over time. We introduce an adaptive learning procedure, where the players only observe their own realized payoff at each stage. We assume that agents do not know their own payoff function and have no information on the other player. Furthermore, we assume that they have restrictions on their own actions such that, at each stage, their choice is limited to a subset of their action set. We prove that the empirical distributions of play converge to the set of Nash equilibria for zero-sum and potential games, and games where one player has two actions.

Key words. learning in games, reinforcement learning, best-response dynamics, stochastic approximations

AMS subject classifications. 91A26, 62L20, 37B25

DOI. 10.1137/130936488

1. Introduction.

First introduced by Brown [12] to compute the value of zero-sum games, *fictitious play* is one of the most intensely studied and debated procedures in game theory. 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.

A major issue in fictitious play is identifying classes of games where the empirical frequencies of play converge to the set of Nash equilibria of the underlying game. A large body of literature has been devoted to this question. Convergence for two-player zero-sum games was obtained by Robinson [34] and for general (nondegenerate) 2×2 games by Miyasawa [31]. Monderer and Shapley [32] proved the same result for potential games, and Berger [9] for two-player games where one of the players has only two actions. Recently, a large proportion of these results has been reexplored using the stochastic approximation theory (see, for example, Benaïm [3], Benveniste, Métivier, and Priouret [8], and Kushner and Yin [28]), where the asymptotic behavior of the fictitious play procedure can be analyzed through related dynamics. For instance, Hofbauer and Sorin [24] obtain more general convergence results for zero-sum games, while Benaïm, Hofbauer, and Sorin [6] extend Monderer and Shapley's result to a general class of potential games, with nonlinear payoff functions on compact convex action sets.

Most of these convergence properties also hold for *smooth fictitious play*, introduced by Fudenberg and Kreps [16] (see also Fudenberg and Levine [17]), where agents use a fictitious play strategy in a game where payoff functions are perturbed by random variables, in the spirit of Harsanyi [20]. For this adaptive procedure, convergence

*Received by the editors September 10, 2013; accepted for publication (in revised form) September 29, 2014; published electronically January 22, 2015. This project was partially funded by Fondecyt grant 3130732, the Nucleo Milenio Información y Coordinación en Redes ICM/FIC P10-024F, and the Complex Engineering Systems Institute (ICM: P-05-004-F, CONICYT: FBO16).
http://www.siam.org/journals/sicon/53-1/93648.html

†Departamento de Ingeniería Industrial and Instituto de Sistemas Complejos de Ingeniería, Universidad de Chile, República 701, Santiago, Chile (mbravo@dii.uchile.cl).

‡Aix-Marseille University (Aix-Marseille School of Economics) CNRS and EHESS, 2 rue de la vieille charité, 13236 Marseille Cedex 02, France (mathieu.faure@univ-amu.fr).

holds in 2×2 games (see Benaïm and Hirsch [5]), zero-sum, potential games (see Hofbauer and Sandholm [23]), and supermodular games (see Benaïm and Faure [4]).

As defined above, in fictitious play or smooth fictitious play, players compute best responses to their opponents' empirical frequencies of play. Three main assumptions are made here: (i) each player knows the structure of the game, i.e., she knows her own payoff function; (ii) each player is informed of the action selected by her opponents at each stage, and thus she can compute the empirical frequencies; and (iii) each player is allowed to choose any action at each time, so that she can actually play a best response.

The next question is usually what happens if assumptions (i) and (ii) are relaxed. One approach is to assume that the agents observe only their realized payoff at each stage. This is the minimal information framework of the so-called reinforcement learning procedures. (See Börgers and Sarin [10] or Erev and Roth [14] for pioneering work on this topic.) Most work in this direction proceeds as follows: (a) construct a sequence of mixed strategies which is updated taking into account the payoff they receive (which is the only information agents have access to) and (b) study the convergence (or nonconvergence) of this sequence. It is supposed that players are given a rule of behavior (a *decision rule*) which depends on a *state variable* constructed by means of the aggregate information they gather and their own history of play.

It is noteworthy that 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. This kind of rule has proved useful in the analysis of simple cases, e.g., 2×2 games (see Posch [33]), two-player games with positive payoff (see Börgers and Sarin [10], Beggs [2], Hopkins [26], and Hopkins and Posch [27]) or in establishing convergence to perturbed equilibria in two-player games (see Leslie and Collins [29]) or multiplayer games (see Cominetti, Melo, and Sorin [13] and Bravo [11]). An example of a nonhomogeneous (time-dependent) decision rule is proposed by Leslie and Collins [30] where, via stochastic approximation techniques, convergence of mixed actions is shown for zero-sum games and multiplayer potential games. Another interesting example that implements a nonhomogeneous decision rule is proposed by Hart and Mas-Colell [22]. Using techniques based on consistent procedures (see Hart and Mas-Colell [21]), the authors show that, for any game, the joint empirical frequency of play converges to the set of correlated equilibria. To our knowledge, this is the only reinforcement learning procedure that uses a decision rule depending explicitly on the last action played (i.e., it is *Markovian*). However, in all the examples described above, assumption (iii) holds; in other words, players can use any action at any time.

A different idea, that of releasing assumption (iii), comes from Benaïm and Raimond [7], who introduced the Markovian fictitious play MFP procedure, where players have restrictions on their action set, due to limited computational capacity or even to physical restrictions. Players know the structure of the game and, at each time, they are informed of opponents' actions, as in the fictitious play framework. Under the appropriate conditions regarding payers' ability to explore their action set, it is shown that this adaptive procedure converges to Nash equilibria for zero-sum and potential games.

Here, we drop all three assumptions (i), (ii), and (iii). The main novelty of this work is that we construct a sophisticated, nonstationary learning procedure in two-player games with minimal information and restrictions on players' action sets. We assume that players do not anticipate opponents' behavior and that they have no information on the structure of the game (in particular, they do not know their own payoff function) nor on opponents' actions at each stage. This means that the only

information allowing agents to react to the environment is their past realized payoffs; the adaptive procedure presented in this work thus belongs to the class of reinforcement learning algorithms. In addition (and in the spirit of the MFP procedure), we suppose that at each stage the agents are restricted to a subset of their action set, which depends on the action they chose at the previous stage. The decision rule we implement is fully explicit, and it is easy for each agent to compute the mixed action which dictates her next action. She actually chooses an action through a nonhomogeneous Markovian rule which depends on a meaningful state variable.

One of the main differences between this procedure and standard reinforcement learning is that the sequence of mixed strategies is no longer a natural choice of state variable. Indeed, the set of mixed strategies available to a given agent at time $n+1$ depends on the action he chose at time n. As a consequence, it is unrealistic to expect good asymptotic behavior from the sequence of mixed strategies, and we turn our attention to the sequence of empirical moves. Our main finding is that the empirical frequencies of play converge to Nash equilibria in zero-sum and potential games, including convergence of the average scored payoffs. We also show convergence in the case where at least one player has only two actions.

This paper is organized as follows. In section 2 we describe the setting and present our model, along with our main result. Section 3 introduces the general framework in which we analyze our procedure. The related MFP procedure is also presented to help the reader better grasp our adaptive procedure. Section 4 gives the proof of our main result, presented as an extended sketch, while the remaining results and technical comments are left to the appendix.

2. The model.

2.1. Setting. Let $\mathcal{G} = (N, (S^i)_{i\in N}, (G^i)_{i\in N})$ be a given finite normal form game and $S = \prod_i S^i$ be the set of action profiles. We call $\Delta(S^i)$ the mixed action set, i.e., $\Delta(S^i) = \{\sigma^i \in \mathbb{R}^{|S^i|} : \sum_{s^i\in S^i} \sigma^i(s^i) = 1, \sigma^i(s^i) \geq 0, \text{ for all } s^i \in S^i\}$, and $\Delta = \prod_i \Delta(S^i)$. More generally, given a finite set S, $\Delta(S)$ denotes the set of probability distributions over S.

In the whole paper, for any agent i, we denote δ_{s^i} the pure action s^i seen as an element of $\Delta(S^i)$. 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 $\prod_{j\neq i} \Delta(S^i)$.

Definition 2.1. *The* best-response correspondence for player $i \in N$, $\mathrm{BR}^i : \Delta^{-i} \rightrightarrows \Delta(S^i)$, *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*

$$\mathrm{BR}(\sigma) = \prod_{i\in N} \mathrm{BR}^i(\sigma^{-i})$$

for $\sigma \in \Delta$.

Recall that a Nash equilibrium of the game $\mathcal{G}$ is a fixed point of the set-valued map BR, namely, a mixed action profile $\sigma^* \in \Delta$ such that $\sigma^* \in \mathrm{BR}(\sigma^*)$.

2.2. Payoff-based Markovian procedure. We consider a situation where the game $\mathcal{G}$ described above is repeated in discrete time. Let $s^i_n \in S^i$ be the action played by player i at time n. We assume that players do not know the game that they are playing, i.e., they know neither their own payoff functions nor opponents'. Also, we assume that the information that a player can gather at any stage of the game is given by her payoff, i.e., at each time n each player $i \in N$ is informed of $g^i_n = G^i(s^1_n, s^2_n, \ldots, s^N_n)$. Players are not able to observe opponents' actions.

In this framework, 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 action σ_n^i according to a *decision rule*, which can depend on state variable X_n^i and time n;
(b) player i's action s_{n+1}^i is randomly drawn according to σ_n^i;
(c) she only observes g_{n+1}^i, as a consequence of the realized action profile $(s_{n+1}^1, \dots, s_{n+1}^N)$;
(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.

In this work we assume that, in addition, players have restrictions on their action set. This idea was introduced by Benaïm and Raimond [7] through the definition of the MFP procedure (see section 3.2 for details). Suppose that when an agent i plays an action $s \in S^i$ at stage $n \in \mathbb{N}$, her available actions at stage $n+1$ are reduced to a subset of S^i. This can be due to physical restrictions, computational limitations, or a large number of available actions. The subset of actions available to player i depends on her last action and is defined through an *exploration graph* $\mathcal{G}^i$ whose vertices are the actions of player i, which is nondirected and strongly connected (that is, for any pair of actions $s, r \in S^i$, there exists a path from s to r). Then, 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 there is an edge between s and r. The connectedness assumption guarantees that agents have access to any of their actions.

In order to incorporate the restriction structure into our reinforcement procedure, we associate an exploration matrix M_0^i to the exploration graph introduced in the model. Namely, we choose a stochastic matrix M_0^i which is compatible with the exploration graph $\mathcal{G}^i$, in the sense that $M_0^i(s,r) > 0$ if and only if there is an edge between r and s. Of course, this matrix is irreducible due to the fact that the exploration graph is strongly connected. We also choose it reversible with respect to its unique invariant measure π_0^i, i.e., the *detailed balance equation* $\pi_0^i(s)M_0^i(s,r) = \pi_0^i(r)M_0^i(r,s)$ holds for every $s, r \in S^i$.

Remark 2.2. Recall that a stochastic matrix M over a finite set S is said to be irreducible if it has a unique recurrent class which is given by S. The reversibility condition is a natural assumption for an exploration matrix. An interpretation is that the distribution of the associated process is invariant by time-reversal. More importantly, it will prove to be very convenient to have a nice explicit expression for the invariant distributions of the stochastic matrices $M^i[\beta, R]$ (see below).

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

$$M^i[\beta,R](s,r) = \begin{cases} M_0^i(s,r)\exp(-\beta|R(s) - R(r)|_+), & s \neq r, \\ 1 - \sum_{s' \neq s} M^i[\beta,R](s,s'), & s = r, \end{cases} \tag{2.1}$$

where, for a number $a \in \mathbb{R}$, $|a|_+ = \max\{a, 0\}$.

From the irreducibility of the exploration matrix M_0^i, we have that $M^i[\beta, R]$ is also irreducible and its unique invariant measure is given by

$$\pi^i[\beta,R](s) = \frac{\pi_0^i(s)\exp(\beta R(s))}{\sum_{r \in S^i} \pi_0^i(r)\exp(\beta R(r))} \tag{2.2}$$

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

Let $(\beta_n^i)_n$ be a deterministic sequence and let $\mathcal{F}_n$ be the σ-algebra generated by the history of play up to time n. Let $R_0^i = 0$. (We choose to initialize the procedure at zero for the sake of simplicity; however, any choice of R_0^i would work.) We suppose that, at the end of stage $n \geq 1$, player i has a state variable $R_n^i \in \mathbb{R}^{|S^i|}$. Let $M_n^i = M^i[\beta_n^i, R_n^i]$ and $\pi_n^i = \pi_n^i[\beta_n^i, R_n^i]$. For $n \geq 0$, Player i selects her action at time $n+1$ through the following *choice rule*:

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

for every $s \in S^i$. As we will see, variable R_n^i will be defined so as to be an estimator of the time-average payoff vector.

At time $n+1$, player i observes her realized action s_{n+1}^i, as well as her realized payoff g_{n+1}^i. The *updating rule* chosen by player i is defined as follows. Agent i updates the vector $R_n^i \in \mathbb{R}^{|S^i|}$, only on the component associated to the action selected at stage n. For every action $s \in S^i$,

$$R_{n+1}^i(s) = R_n^i(s) + \gamma_{n+1}^i(s)\left(g_{n+1}^i - R_n^i(s)\right)\mathbb{1}_{\{s_{n+1}^i = s\}}, \tag{UR}$$

where

$$\gamma_{n+1}^i(s) = \min\left\{1\,, \frac{1}{(n+1)\pi_n^i(s)}\right\},$$

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

Remark 2.3. For the sake of simplicity, we refer to R_n^i as the state variable of player i even if, strictly speaking, the actual state variable is of the form $X_n^i = (R_n^i, s_n^i)$, given that the choice rule (CR) is Markovian.

Note that the step size $\gamma_{n+1}^i(s)$ depends only on π_0^i, β_n^i and R_n^i. Also, as we will see later on, $(\gamma_n^i(s))^{-1} = n\pi_{n-1}^i(s)$ for sufficiently large n (cf. section A.2).

While choosing this step size might appear surprising, we believe that it is actually very natural, as it takes advantage of the fact that the invariant distribution π_n^i is known by player i. To put it another way: a natural candidate for step size $\gamma_n^i(s)$ in (UR) is $\gamma_n^i(s) = 1/\theta_n^i(s)$, where $\theta_n^i(s)$ is equal to the number of times agent i actually played action s during the n first steps. If the Markov process was homogeneous and ergodic, with invariant measure π^i, then the expected value of θ_n^i would be exactly $n\pi^i(s)$.

Consequently, our stochastic approximation scheme (UR) can be interpreted as follows. Assume that, at time $n+1$, action s is played by agent i. Then $R_{n+1}^i(s)$ is updated by taking a convex combination of $R_n^i(s)$ and of the realized payoff playing s at time $n+1$; additionally, the weight that is put on the realized payoff is inversely proportional to the number of times this action *should* have been played (and not the number of times it has *actually* been played).

Let us denote by $(v_n^i)_n$ the sequence of empirical distribution of moves of agent i, i.e., $v_n^i = n^{-1}\sum_{m=1}^n \delta_{s_m^i}$, and $v_n = (v_n^i)_{i \in N} \in \Delta$. Note that, given the physical restrictions on the action set, one cannot expect convergence results on the mixed

actions of players σ_n^i. Therefore, the empirical frequencies of play become the natural focus of our analysis.

DEFINITION 2.4. *We call the* 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).

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

Assumption 2.5. For $i \in \{1,2\}$, the sequence $(\beta_n^i)_n$ is positive and verifies

(i) $\beta_n^i \longrightarrow +\infty$,

(ii) $\beta_n^i = A_n^i \ln(n)$, where A_n^i is nonincreasing and $A_n^i \longrightarrow 0$ as $n \to +\infty$.

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 $\mathcal{L}((z_n)_n) \subseteq A$. In the case where A is a closed set, this amounts to having $\lim_{n\to+\infty} d(z_n, A) = 0$.

First, we establish a relationship between the limit set of the sequence $(v_n^1, v_n^2)_n$ and the attractors (more precisely the *internally chain transitive* (ICT) sets, defined below) of the well-known best-response dynamics (BRD), introduced by Gilboa and Matsui [18],

$$\dot{v} \in -v + \mathrm{BR}(v). \tag{BRD}$$

For this purpose, we need to introduce some notions that will be useful in what follows.

Let $\Sigma \subseteq \mathbb{R}^d$ be a convex compact set and let us consider a set-valued map with nonempty convex values $C : \Sigma \rightrightarrows \Sigma$. Let us suppose that its graph

$$\mathrm{Gr}(C) = \{(z,\mu) : \ z \in \Sigma,\ \mu \in C(z)\}$$

is a closed set in $\Sigma \times \Sigma$.

Under these assumptions, it is well known (see, e.g., Aubin and Cellina [1]) that the differential inclusion

$$\dot{z} \in -z + C(z) \tag{DI}$$

admits at least one solution (i.e., an absolutely continuous mapping $\mathbf{z} : \mathbb{R} \to \mathbb{R}^d$ such that $\dot{\mathbf{z}}(t) \in -\mathbf{z}(t) + C(\mathbf{z}(t))$ for almost every t) through any initial point.

DEFINITION 2.6. *A nonempty compact set $\mathcal{A} \subseteq \Sigma$ is called an* attractor *for* (DI)*, provided*

(i) *it is* invariant*, i.e., for all $v \in \mathcal{A}$, there exists a solution $\mathbf{z}$ to* (DI) *with $\mathbf{z}(0) = v$ and such that $\mathbf{z}(\mathbb{R}) \subseteq \mathcal{A}$,*

(ii) *there exists an open neighborhood $\mathcal{U}$ of $\mathcal{A}$ such that, for every $\epsilon > 0$, there exists $t_\epsilon > 0$ such that $\mathbf{z}(t) \subseteq N^\epsilon(\mathcal{A})$ for any solution $\mathbf{z}$ starting in $\mathcal{U}$ and all $t > t_\epsilon$, where $N^\varepsilon(\mathcal{A})$ is the ε-neighborhood of $\mathcal{A}$. An open set $\mathcal{U}$ with this property is called a* fundamental neighborhood *of $\mathcal{A}$.*

A compact set $D \subseteq \Sigma$ is ICT if it is invariant and connected and has no proper attractors.

Now we can state our first result.

THEOREM 2.7. *Under Assumption* 2.5*, assume that players follow the payoff-based adaptive Markovian procedure. Then the limit set of the sequence $(v_n)_n$ is an*

ICT *set of the best-response dynamics* (BRD). *In particular, if* (BRD) *admits a global attractor* $\mathcal{A}$, *then* $\mathcal{L}((v_n)_n) \subseteq \mathcal{A}$.

As a consequence of Theorem 2.7 as well as some known results (see section 4 for details), we now characterize the asymptotic behavior of the sequence $(v_n)_n$ in several classes of two-player games The most interesting aspect here is that we can also prove convergence results for the actual average payoff scored by the players along the trajectories.

Let us denote by $\overline{g}_n^i$ the average payoff obtained by player i, i.e.,

$$\overline{g}_n^i = n^{-1} \sum_{m=1}^{n} G^i(s_m^1, s_m^2),$$

and $\overline{g}_n = (\overline{g}_n^1, \overline{g}_n^2)$.

Recall that $\mathcal{G}$ is a potential game with potential Φ if, for all $i = 1, 2$, and $s^{-i} \in S^{-i}$, we have $G^i(s^i, s^{-i}) - G^i(t^i, s^{-i}) = \Phi(s^i, s^{-i}) - \Phi(t^i, s^{-i})$ for all $s^i, t^i \in S^i$.

THEOREM 2.8. *Under Assumption* 2.5, *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* $n^{-1} \sum_{m=1}^{n} \Phi(s_m^1, s_m^2)$ *converges to this constant.*
In the particular case $G^1 = G^2$, $(v_n^1, v_n^2)_n$ *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.*

Comments on Theorem 2.8. For potential games, in the general case, the potential is constant on the limit set of $(v_n)_n$, almost surely. Unfortunately, our result does not allow us to know if this is also true for the payoff of a given player. However, we conjecture that it is not necessarily the case.

Consider the game $\mathcal{G}$ with payoff function G and potential Φ:

$(\mathcal{G})$ $\quad G =$

	a	b	c
A	1,1	9,0	1,0
B	0,9	6,6	0,8
C	0,1	8,0	2,2

and $\quad \Phi =$

	a	b	c
A	4	3	3
B	3	0	2
C	3	2	4

.

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

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

because this set is not connected.

Now consider the following modified version $\mathcal{G}'$:

$(\mathcal{G}')$ $\quad G' =$

	a	b	c
A	1,1	9,0	1,0
B	0,9	6,6	0,8
C	1,2	8,0	2,2

and $\quad \Phi' =$

	a	b	c
A	4	3	3
B	3	0	2
C	4	2	4

.

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

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

Consequently, there is no reason to rule out the possibility that the limit set of $(v_n)_n$ is equal to the whole set of Nash equilibria. Therefore, the payoff is not necessarily constant on $\mathcal{L}((v_n)_n)$.

Remark 2.9. As pointed out by Jérôme Renault and an anonymous referee, we say nothing about the convergence or nonconvergence of the empirical joint distributions, that is, the random sequence $n^{-1}\sum_{m=1}^{n}\delta_{s_m}$. This is undoubtedly an interesting and challenging question. The answer is trivial (and positive) when the set of Nash equilibria is restricted to a single pure Nash equilibrium. However, in the general case, this is an open problem.

Remark 2.10. Obviously, when the players are allowed to play any action at any time, our results are still valid. However, in such a scenario, a natural state variable is given by the sequence of mixed actions instead of the sequence of average moves. In [30], this is done by relating the asymptotic behavior of the day-to-day strategy with the attractors of the best-response dynamics (BRD), similarly as in our work. This shows that a Markovian procedure is no longer necessary to reach the same results (now on the sequence of mixed actions). Moreover, convergence for N-player potential games ($N \geq 3$) can be obtained.

Comments on the assumptions. Assumption 2.5 supposes that the sequence β_n^i increases to infinity as $o(\ln(n))$. This assumption is necessary due to the informational constraints on players. For instance, it is not possible to know a priori how far the variables R_0^i are from the set of feasible payoffs.

As we will see later on, in the MFP procedure, sequence β_n^i is supposed to grow more slowly than $A^i \ln(n)$, where A^i is smaller than a quantity which is related to the *energy barrier* of the payoff matrix G^i (see Benaïm and Raimond [7] for details). This quantity is in turn related to the *freezing schedule* of the simulated annealing algorithm (see, for example, Holley and Stroock [25] and Hajek [19], and references therein).

We believe it is worth reformulating our result in this spirit. However, this requires players to have more information about the game. For each $i \in \{1,2\}$, suppose that the initial state variable R_0^i belongs to the set of feasible payoffs. Also, let us define the quantity

$$\omega^i = \max_{s \in S^i} \max_{s^{-i}, r^{-i} \in S^{-i}} |G^i(s, s^{-i}) - G^i(s, r^{-i})|,$$

and let us consider the following assumption.

Assumption 2.11. Each player $i \in \{1,2\}$ can choose a positive constant A^i such that

(i) $\beta_n^i \longrightarrow +\infty$,

(ii) $\beta_n^i \leq A^i \ln(n)$, where $2A^i\omega^i < 1$.

Then, we have the following version of our main result.

THEOREM 2.12. *Under Assumption* 2.11, *the conclusions of Theorem* 2.8 *hold.*

The proof of this result runs along the same lines as the proof of Theorem 2.8 and is therefore omitted.

2.4. Examples. The following simple examples show the scope of Theorem 2.8. In every case presented in this section, we performed a maximum of 5×10^5 iterations.

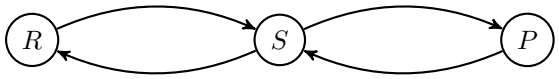


FIG. 1. *Graph representing players' restrictions for the* (RSP) *game (every state has a loop).*

Blind-restricted RSP. Consider the rock-scissor-paper (RSP) game defined by the payoff matrix G^1:

(RSP)

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

.

Then the 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. Players' exploration matrices and their invariant measures are given by

$$M_0^1 = M_0^2 = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \end{pmatrix} \text{ and } \pi_0^1 = \pi_0^2 = \begin{pmatrix} 2/7 \\ 3/7 \\ 2/7 \end{pmatrix}. \tag{2.3}$$

Figure 1 means that if a player's action is *rock* at some time, she cannot select *paper* immediately afterward, and inversely. In Figure 2, we present a realization of $(v_n)_n$, as well as $(g_n^1)_n$.

3×3 potential game. Consider the potential game with payoff matrix G' and potential Φ' (see $(\mathcal{G}')$). We assume that players' exploration matrices are also given by (2.3). Therefore, the graph representing the restriction of players is given by Figure 1 if R, S, and P are replaced by A, B, and C, respectively.

Figure 3 shows a realization of our procedure for the game $(\mathcal{G}')$. On the left, we plot the evolution of v_n^1. On the right, we present the corresponding trajectory of $\overline{\Phi}'_n = n^{-1}\sum_{m=1}^n \Phi'(s_m^1, s_m^2)$, the average value of the potential Φ' along the realization of $(s_n^1, s_n^2)_n$. Note that our results do not stipulate that $(v_n)_n$ converges (which is an open question; see our comments on Theorem 2.8), and that our simulation tends towards nonconvergence of v_n^1. We choose not to display v_n^2 here (which seems to converge to the action a).

5×5 identical interests game. Consider the game with identical interests where both players have five actions and the common payoff matrix is given by

(C)

	A	B	C	D	E
A	2	0	0	0	0
B	0	1	0	0	0
C	0	0	0	0	0
D	0	0	0	1	0
E	0	0	0	0	2

.

Assume that players' exploration matrices are

$$M_0^1 = M_0^2 = \begin{pmatrix} 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \end{pmatrix} \text{ with } \pi_0^1 = \pi_0^2 = \begin{pmatrix} 2/13 \\ 2/13 \\ 5/13 \\ 2/13 \\ 2/13 \end{pmatrix},$$

which corresponds to the graph displayed in Figure 4.

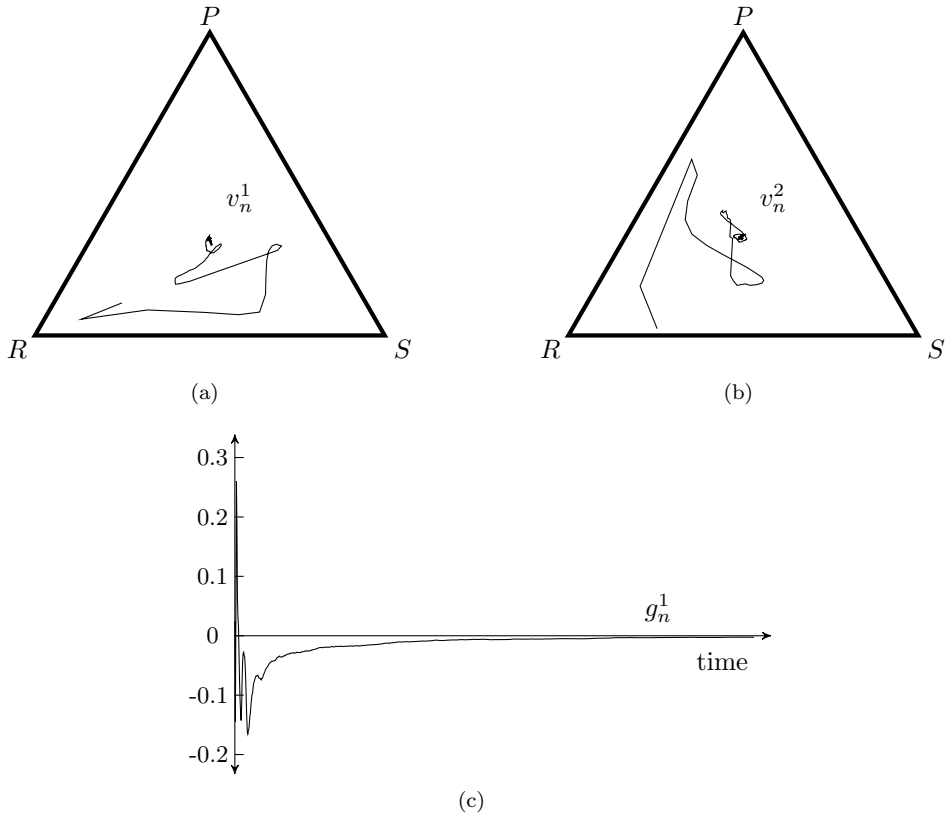


FIG. 2. *At the top, a realization of* v_n. *At the bottom,* g_n^1.

Note that, even if the center action C is bad for both players, the restrictions force them to play C every time they switch to another action.

In Figure 5, on the left, we present a realization where (v_n^1, v_n^2) converges to the NE (B, B). On the right, a trajectory where (v_n^1, v_n^2) converges to the NE (E, E) is displayed. Note that, in both cases, the average realized payoff $\overline{g}_n$ converges to the payoff of the corresponding equilibrium. For simplicity, we only plot the component that converges to one for the first player. This is consistent with the recent finding that the four strict Nash equilibria have a positive probability of being the limit of the random process $(v_n)_n$. (See Faure and Roth [15] for details.)

3. Preliminaries to the proof, related work. The aim of this section is twofold: we introduce the general framework in which we analyze our procedure, and we present the related MFP procedure, where the idea of restrictions on the action set was first introduced.

3.1. A general framework. Let S be a finite set and let $\mathcal{M}(S)$ be the set of Markov matrices over S. We consider a discrete time stochastic process $(s_n, M_n)_n$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $S \times \mathcal{M}(S)$. The space $(\Omega, \mathcal{F}, \mathbb{P})$ is equipped with a nondecreasing sequence of σ-algebras $(\mathcal{F}_n)_n$.

Let us assume the following on the sequence $(s_n, M_n)_n$.

Assumption 3.1. The process $(s_n, M_n)_n$ satisfies the following:

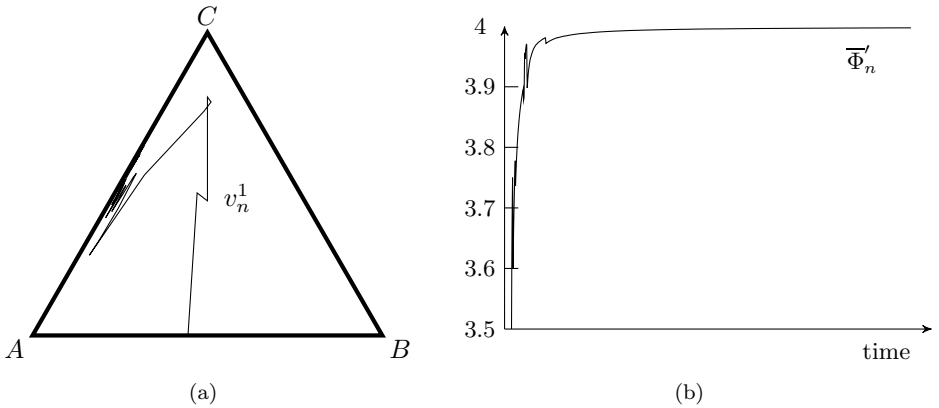


FIG. 3. *Simulation results for the potential game* $(\mathcal{G}')$.

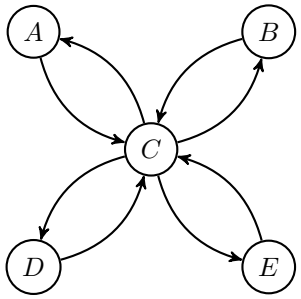


FIG. 4. *Graph representing players' restrictions for the game* (C) *(every state has a loop).*

(i) For all $n \in \mathbb{N}$, (s_n, M_n) is $\mathcal{F}_n$-measurable.
(ii) For all $s \in S$ and $n \in \mathbb{N}$, $\mathbb{P}(s_{n+1} = s \mid \mathcal{F}_n) = M_n(s_n, s)$.
(iii) For all $n \in \mathbb{N}$, the matrix M_n is irreducible with invariant measure $\pi_n \in \Delta(S)$.

Let Σ be again a compact convex subset of an euclidean space $\mathbb{R}^d$ and $H : S \to \Sigma$. For all $n \in \mathbb{N}$, let $V_n = H(s_n) \in \Sigma$. We are interested in the asymptotic behavior of the random sequence $z_n = n^{-1} \sum_{m=1}^{n} V_m$. Let us call

$$\mu_n = \sum_{s \in S} \pi_n(s) H(s) \in \Sigma. \tag{3.1}$$

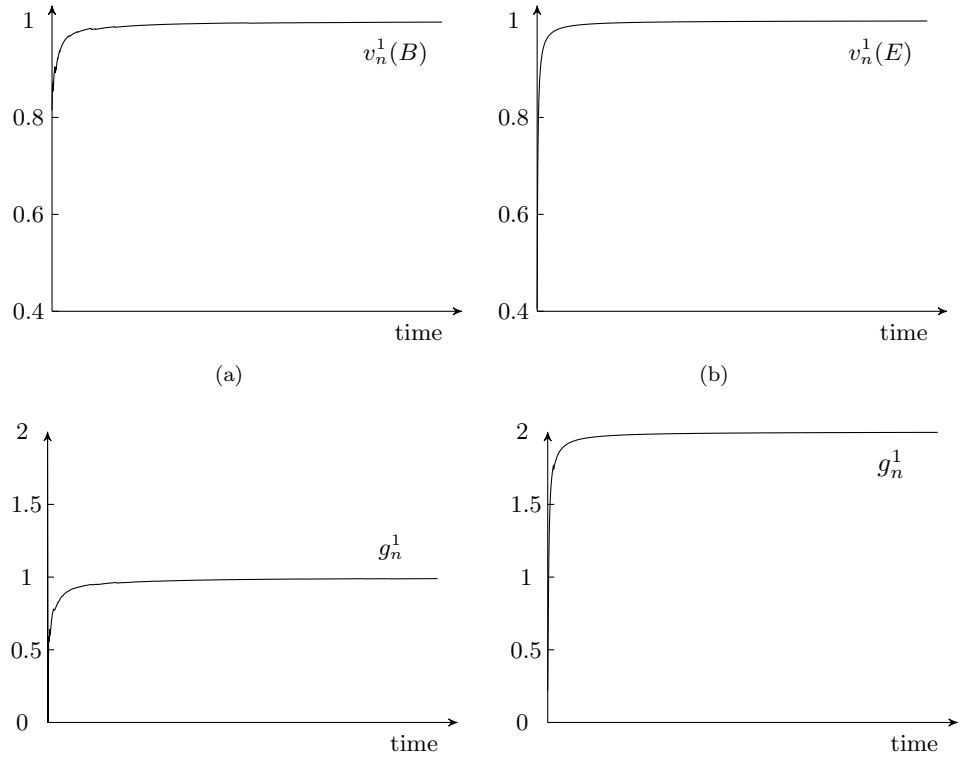
Remark 3.2. This setting is a simplification of that considered by Benaïm and Raimond [7], where a more general observation term V_n is treated. For instance, V_n may depend on other nonobservable variables or explicitly on time.

In order to maintain the original terminology, we introduce the following definition, which is stated in a slightly different form (see [7, Definition 2.4]).

DEFINITION 3.3. *A set-valued map with nonempty convex values* $C : \Sigma \rightrightarrows \Sigma$ *is adapted to the random sequence* $(z_n, \mu_n)_n$ *if*

(i) *its graph* $\mathrm{Gr}(C)$ *is closed in* $\Sigma \times \Sigma$,
(ii) *almost surely, for any limit point* (z, μ) *of* $(z_n, \mu_n)_n$, *we have* $(z, \mu) \in \mathrm{Gr}(C)$.

Given a set-valued map $C : \Sigma \rightrightarrows \Sigma$ adapted to a random sequence $(z_n, \mu_n)_n$, recall that (DI) is the differential inclusion $\dot{z} \in -z + C(z)$.



(a) At the top, $v_n^1(B) \to 1$. At the bottom, $\overline{g}_n^1 \to 1$. (b) At the top, $v_n^1(E) \to 1$. At the bottom, $\overline{g}_n^1 \to 2$.

FIG. 5. *Two realizations of the procedure for the game* (C).

Let $m(t) = \sup\{m \geq 0 : t \geq \tau_m\}$, where $\tau_m = \sum_{j=1}^{m} 1/j$. For a sequence $(u_n)_n$ and a number $T > 0$, we define $\epsilon(u_n, T)$ by

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

Let us denote by $(W_n)_n$ the random sequence defined by $W_{n+1} = H(s_{n+1}) - \mu_n$. The evolution of z_n can be recast as

$$z_{n+1} - z_n = \frac{1}{n+1}(\mu_n - z_n + W_{n+1}). \tag{3.2}$$

A consequence of [7, Theorem 2.6] in this particular framework is the following result.

THEOREM 3.4. *Under Assumption* 3.1, *assume that the set-valued map* C *is adapted to* $(z_n, \mu_n)_n$ *and that for all* $T > 0$

$$\lim_{n \to +\infty} \epsilon\left(\frac{1}{n+1} W_{n+1}, T\right) = 0 \tag{3.3}$$

almost surely. Then, the limit set of $(z_n)_n$ *is, almost surely, an* ICT *set of the differential inclusion* (DI). *In particular, if* $\mathcal{A}$ *is a global attractor for* (DI), *then the limit set of* $(z_n)_n$ *is almost surely contained in* $\mathcal{A}$.

Remark 3.5. Roughly speaking, the fact that the set-valued map C is adapted to (z_n, μ_n) means that (3.2) can be recast as

$$z_{n+1} - z_n \in \frac{1}{n+1}(-z_n + C(z_n) + W_{n+1}).$$

In turn, this recursive form can be seen as a Cauchy–Euler scheme to approximate the solutions of the differential inclusion (DI) with decreasing step sizes and added noise term $(W_n)_n$. Assumption (3.3) guarantees that, on any given time horizon, the noise term asymptotically vanishes. As a consequence, the limit set of $(z_n)_n$ can be described through the deterministic dynamics (DI), in the sense that it needs to be internally chain transitive. If the differential inclusion admits a global attractor, then any ICT set is contained in it. This implies the second point of the theorem. (See Benaïm, Hofbauer, and Sorin [6] for a full discussion on stochastic approximations for differential inclusions.)

3.2. Markovian fictitious play. As in section 2, we consider that players have constraints on their action set, i.e., each player has an exploration matrix M_0^i which is supposed to be irreducible and reversible with respect to its unique invariant measure π_0^i.

The crucial difference between MFP and the procedure introduced in section 2.2 is that players know their own payoff function. Also, at the end of each stage, each player is informed of the opponent's action. The MFP procedure is defined as follows. A player's i action at time $n+1$ is chosen according to the nonhomogeneous Markov matrix

$$\begin{aligned}\mathbb{P}(s_{n+1}^i = s \,|\, \mathcal{F}_n) &= M^i[\beta_n^i, U_n^i](s_n^i, s) \\ &= \begin{cases} M_0^i(s_n^1, s)\exp(-\beta_n^i |U_n^i(s_n^i) - U_n^i(s)|_+), & s \neq s_n^1, \\ 1 - \sum\limits_{s' \neq s} M^i[\beta_n^i, U_n^i](s_n^i, s'), & s = s_n^1, \end{cases}\end{aligned} \tag{3.4}$$

where U_n^i is taken as the vector payoffs of player i, against the average moves of the opponent

$$U_n^i = G^i(\cdot, v_n^{-i}) = \frac{1}{n}\sum_{m=1}^{n} G^i(\cdot, s_m^{-i})$$

for all $s \in S^i$, and the function $M^i[\cdot,\cdot]$ is defined by (2.1). Let $\tilde{M}_n^i = M^i[\beta_n^i, U_n^i]$. Observe that again, from the irreducibility of M_0^i, the matrix $\tilde{M}_n^i$ is also irreducible. Also, $\tilde{\pi}_n^i = \pi^i[\beta_n^i, G^i(\cdot, v_n^{-i})]$ (where $\pi^i[\cdot,\cdot]$ is defined in (2.2)) is the unique invariant measure of $\tilde{M}_n^i$, i.e.,

$$\tilde{\pi}_n(s) = \frac{\pi_0^i(s)\exp(\beta_n^i U_n^i(s))}{\sum\limits_{s' \in S^i} \pi_0^i(s')\exp(\beta_n^i U_n^i(s'))}$$

for every $s \in S^i$.

Benaïm and Raimond [7] obtain the following result.

THEOREM 3.6. *If both players follow the* MFP *procedure, defined by* (3.4)*, then the limit set of the sequence* $v_n = (v_n^1, v_n^2)$ *is an* ICT *set of the best-response dynamics* (BRD)*, provided that for* $i \in \{1, 2\}$ *the positive sequence* $(\beta_n^i)_n$ *satisfies*

(i) $\beta_n^i \to +\infty$ *as* $n \to +\infty$,

(ii) $\beta_n^i \leq A^i \log(n)$ *for a sufficiently small positive constant* A^i.

As a consequence, we have the following:

(a) *In a zero-sum game,* $(v_n^1, v_n^2)_n$ *converges almost surely to the set of Nash equilibria.*

(b) *If* $G^1 = G^2$*, 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.*

Observe that Theorems 2.7 and 2.8 imply that all the conclusions of Theorem 3.6 above hold for our procedure.

Some insights on the proof of Theorem 3.6. We believe that it is interesting to sketch the proof of Theorem 3.6. For that purpose, we need to introduce some notions that will be useful later on.

Let S be a finite set and M an irreducible stochastic matrix over S with invariant measure π. For a function $f : S \to \mathbb{R}$, the variance and the energy of f are defined, respectively, as

$$\operatorname{var}(f) = \sum_{s\in S} \pi(s) f^2(s) - \left(\sum_{s\in S} \pi(s) f(s) \right)^2,$$

$$\mathcal{E}(f,f) = \frac{1}{2} \sum_{s,r\in S} (f(s) - f(r))^2 M(s,r)\pi(s).$$

Definition 3.7. *Let* M *be a stochastic irreducible matrix over the finite set* S *and* π *be its unique invariant measure.*

(i) *The* spectral gap *of* M *is defined by*

$$\chi(M) = \min\left\{ \frac{\mathcal{E}(f,f)}{\operatorname{var}(f)} \,:\, \operatorname{var}(f) \neq 0 \right\}.$$

(ii) *The* pseudoinverse *matrix of* M *is the unique matrix* $Q \in \mathbb{R}^{|S|\times|S|}$ *such that* $\sum_{r\in S} Q(s,r) = 0$ *for every* $s \in S$*, which satisfies the Poisson's equation*

$$Q(I - M) = (I - M)Q = I - \Pi, \tag{3.5}$$

where Π *is the matrix defined as* $\Pi(s,r) = \pi(r)$ *for every* $s, r \in S$ *and* I *denotes the identity matrix.*

For a matrix $Q \in \mathbb{R}^{|S|\times|S|}$ and a vector $U \in \mathbb{R}^{|S|}$, set $|Q| = \max_{s,r} |Q(s,r)|$ and $|U| = \max_s |U(s)|$.

We want to apply Theorem 3.4 with $H(s) = (\delta_{s^1}, \delta_{s^2})$. Recall that v_n^i is the empirical frequency of play of player i. Thus, the random variable $z_n = v_n$ is given by

$$v_n = \frac{1}{n} \sum_{m=1}^{n} \left(\delta_{s_n^1}, \delta_{s_n^2} \right) = \left(v_n^1, v_n^2 \right).$$

Therefore, the evolution of v_n is described by

$$v_{n+1} - v_n = \frac{1}{n+1}(\mu_n - v_n + W_{n+1}),$$

where $\mu_n = \sum_{s\in S} \tilde{\pi}_n(s) H(s) = (\tilde{\pi}_n^1, \tilde{\pi}_n^2)$ and

$$\tilde{W}_{n+1} = (\delta_{s_n^1}, \delta_{s_n^2}) - \mu_n = \left(\delta_{s_n^1} - \tilde{\pi}_n^1, \delta_{s_n^2} - \tilde{\pi}_n^2 \right).$$

We first provide a sketch of the proof that (3.3) holds for the sequence $(\tilde{W}_n)_n$. Afterward, we will verify that the set-valued map BR is adapted to $(v_n, \mu_n)_n$ and will conclude by applying Theorem 3.4.

Consequences (a) and (b) for games follow from the fact that the set of Nash equilibria is an attractor for the best-response dynamics in the relevant classes of games. We will omit this part of the proof, since the same argument will be used in section 4.2.

Let $\tilde{Q}_n^i$ be the pseudoinverse of $\tilde{M}_n^i$. Benaïm and Raimond prove that if, for $i \in \{1, 2\}$,

$$\begin{aligned} \lim_{n\to+\infty} \frac{|\tilde{Q}_n^i|^2 \ln(n)}{n} &= 0, \\ \lim_{n\to+\infty} |\tilde{Q}_{n+1}^i - \tilde{Q}_n^i| &= 0, \\ \lim_{n\to+\infty} |\tilde{\pi}_{n+1}^i - \tilde{\pi}_n^i| &= 0 \end{aligned} \tag{3.6}$$

almost surely, then (3.3) holds for $(\tilde{W}_n)_n$.

Proposition 3.4 in Benaïm and Raimond [7] shows that the norm of $\tilde{Q}_n^i$ can be controlled as a function of the spectral gap $\chi(\tilde{M}_n^i)$. If in addition the constants A^i are sufficiently small, then (3.6) holds.

Finally, since $\beta_n^i \to +\infty$, we have that if $(v_n^1, v_n^2) \to (v^1, v^2)$, then $\tilde{\pi}_n^i \to \overline{\pi}^i[v^{-i}]$, where, for all $s \in S^i$

$$\overline{\pi}^i[v^{-i}](s) = \frac{\pi_0^i(s)\mathbb{1}_{\{s\in \operatorname{argmax}_r G^i(r,v^{-i})\}}}{\sum_{s'\in S^i} \pi_0^i(s')\mathbb{1}_{\{s'\in \operatorname{argmax}_r G^i(r,v^{-i})\}}} \in \mathrm{BR}^i(v^{-i}).$$

This implies that map BR is adapted to $(v_n, \mu_n)_n$, and the proof is finished.

4. Proof of the main results. There are two key aspects which highlight the difference between the proof of Theorem 2.7 and the proof of Theorem 3.6. First, to show that the noise sequence (defined in (4.1) below) satisfies condition (3.3), we cannot directly use condition (3.6). Second, the proof that BR is adapted to $(v_n, \mu_n)_n$ is considerably more involved. In contrast to the approach for the MFP procedure, the invariant measure π_n^i of matrix M_n^i depends on state variable R_n^i, which is updated, in turn, using π_{n-1}^i. To overcome these difficulties, we develop a more general approach, which is presented in the appendix.

In what follows, we present an extended sketch of the proof of Theorem 2.7. The proof of Theorem 2.8 will follow as a consequence.

4.1. Proof of Theorem 2.7. We aim to apply Theorem 3.4. Let $\Sigma = \Delta(S^1) \times \Delta(S^2)$. We take $V_n = (\delta_{s_n^1}, \delta_{s_n^2})$ and $\mu_n = (\pi_n^1, \pi_n^2)$. As before, let $v_n = (v_n^1, v_n^2)$. Then we have

$$v_{n+1} - v_n = \frac{1}{n+1}\left(\mu_n - v_n + \overline{W}_{n+1}\right),$$

where

$$\overline{W}_{n+1} = \left(\overline{W}^1_{n+1}, \overline{W}^2_{n+1}\right) = (\delta_{s^1_{n+1}} - \pi_n^1, \delta_{s^2_{n+1}} - \pi_n^2). \tag{4.1}$$

We need to verify that two conditions hold. First, we have to prove that

$$\epsilon\left(\overline{W}_{n+1}/(n+1), T\right) \to 0$$

almost surely for all $T > 0$. Proposition A.6(ii) provides proof of this.

Second, we need to verify that the best-response correspondence BR is adapted to $(v_n, \mu_n)_n$. As we will see, this problem basically amounts to showing that vector R_n^i becomes a good asymptotic estimator of vector $G^i(\cdot, v_n^{-i})$.

Fix $i \in \{1,2\}$ and $s \in S^i$. Lemma A.3 shows that for a sufficiently large n, $(\gamma_{n+1}^i(s))^{-1} = (n+1)\pi_n^i(s)$ for any $s \in S^i$. Therefore, from the definition of R_n^i and without any loss of generality, we have

$$\begin{aligned} R_{n+1}^i(s) - R_n^i(s) &= \frac{1}{(n+1)\pi_n^i(s)} \left[\mathbb{1}_{\{s_{n+1}^i = s\}} G^i(s, s_{n+1}^{-i}) - \mathbb{1}_{\{s_{n+1}^i = s\}} R_n^i(s) \right], \\ &= \frac{1}{(n+1)\pi_n^i(s)} \Big[\pi_n^i(s) \left(G^i(s, \pi_n^{-i}) - R_n^i(s) \right) \\ &\qquad + \left(\mathbb{1}_{\{s_{n+1}^i = s\}} G^i(s, s_{n+1}^{-i}) - \pi_n^i(s) G^i(s, \pi_n^{-i}) \right) \\ &\qquad + R_n^i(s) \left(\pi_n^i(s) - \mathbb{1}_{\{s_{n+1}^i = s\}} \right) \Big]. \end{aligned}$$

Hence,

$$R_{n+1}^i(s) - R_n^i(s) = \frac{1}{n+1} \left[G^i(s, \pi_n^{-i}) - R_n^i(s) + W_{n+1}^i(s) \right], \tag{4.2}$$

where for convenience we set $W_{n+1}^i(s) = W_{n+1}^{i,1}(s) + W_{n+1}^{i,2}(s)$ with

$$W_{n+1}^{i,1}(s) = \frac{R_n^i(s)}{\pi_n^i(s)} \left(\pi_n^i(s) - \mathbb{1}_{\{s_{n+1}^i = s\}} \right) \text{ and} \tag{4.3}$$

$$W_{n+1}^{i,2}(s) = \frac{1}{\pi_n^i(s)} \left(\mathbb{1}_{\{s_{n+1}^i = s\}} G^i(s_{n+1}^1, s_{n+1}^2) - \pi_n^i(s) G^i(s, \pi_n^{-i}) \right). \tag{4.4}$$

Propositions A.6(i) and A.7 prove that, almost surely and for any $T > 0$, $\epsilon(W_{n+1}^{i,1}(s)/(n+1), T) \to 0$ and $\epsilon(W_{n+1}^{i,2}(s)/(n+1), T) \to 0$, respectively.

Recall that $U_n^i = G^i(\cdot, v_n^{-i})$. Naturally, the evolution of vector U_n^i can be written as

$$U_{n+1}^i - U_n^i = \frac{1}{n+1} \left(G^i(\cdot, \pi_n^{-i}) - U_n^i + W_{n+1}^{i,3} \right), \tag{4.5}$$

where $W_{n+1}^{i,3} = G^i(\cdot, s_{n+1}^{-i}) - G^i(\cdot, \pi_n^{-i})$. Again, Proposition A.6(iii) shows that for all $T > 0$, $\epsilon(W_{n+1}^{i,3}/(n+1), T) \to 0$ almost surely.

Remark 4.1. Note that (4.5) does not hold if there are three players or more, since the maps $v^{-i} \mapsto G^i(s^i, v^{-i})$ are no longer linear in that case.

We define $\zeta_n^i = R_n^i - G^i(\cdot, v_n^{-i}) = R_n^i - U_n^i$. Equations (4.2) and (4.5) show that the evolution of the sequence $(\zeta_n^i)_n$ can be recast as

$$\zeta_{n+1}^i - \zeta_n^i = \frac{1}{n+1} \left[-\zeta_n^i + \mathcal{W}_{n+1}^i \right],$$

where $\mathcal{W}_{n+1}^i = W_{n+1}^{i,1} + W_{n+1}^{i,2} - W_{n+1}^{i,3}$, and each component of $W_{n+1}^{i,1}$ and $W_{n+1}^{i,2}$ is defined by (4.3) and (4.4), respectively.

Collecting all the analysis above, we conclude that $\epsilon(\mathcal{W}_{n+1}^i/(n+1), T) \to 0$ almost surely for all $T > 0$.

Based on the fact that sequence $(\zeta_n^i)_n$ is bounded (see Lemma A.3) and on standard results from stochastic approximation theory, the limit set of the sequence $(\zeta_n^i)_n$ is almost surely an ICT set of the ordinary differential equation $\dot{\zeta} = -\zeta$ which admits the set $\{0\}$ as a global attractor.

Therefore, for $i \in \{1,2\}$, $R_n^i - G^i(\cdot, v_n^{-i}) \to 0$ as $n \to +\infty$ almost surely.

Now let us assume that

$$(v_{n_k}^1, v_{n_k}^2) \to (v^1, v^2) \in \Sigma, \text{ and } (\pi_{n_k}^1, \pi_{n_k}^2) \to (\pi^1, \pi^2) \in \Sigma$$

for a subsequence $(n_k)_k$.

For $i \in \{1,2\}$, let $r \notin \operatorname{argmax}_{s'} G^i(s', v^{-i})$ and take $\hat{s} \in S^i$ such that $G^i(r, v^{-i}) < G^i(\hat{s}, v^{-i})$. Since $R_n^i - G^i(\cdot, v_n^{-i}) \to 0$, there exists $\varepsilon > 0$ and $k_0 \in \mathbb{N}$ such that, for any $k \geq k_0$, $R_{n_k}^i(r) < R_{n_k}^i(\hat{s}) - \varepsilon$, so that, for k sufficiently large,

$$\pi_{n_k}^i(r) \leq \frac{\pi_0^i(r)}{\pi_0^i(\hat{s})} \exp\left[\beta_{n_k}^i (R_{n_k}^i(r) - R_{n_k}^i(\hat{s}))\right] \leq \frac{\pi_0^i(r)}{\pi_0^i(\hat{s})} \exp(-\beta_{n_k}^i \varepsilon).$$

Then, $\pi^i(r) = 0$ and we have proved that $\pi^i \in \mathrm{BR}^i(v^{-i})$, which implies that the set-valued map BR is adapted to $(v_n, \mu_n)_n$. ∎

4.2. Proof of Theorem 2.8. For all three points, the result follows from an application of Theorem 2.7.

Consider the variable $z_n = (v_n^1, v_n^2, \overline{g}_n^1, \overline{g}_n^2)$, where $\overline{g}_n^i = n^{-1} \sum_{m=1}^n g_m^i$ is the average realized payoff for player $i \in \{1,2\}$. Recall that the evolution of $\overline{g}_n^i$ can be written as

$$\begin{aligned} \overline{g}_{n+1}^i - \overline{g}_{n+1}^i &= \frac{1}{n+1}\left(g_{n+1}^i - \overline{g}_n^i\right) \\ &= \frac{1}{n+1}\left(G^i(\pi_n^i, \pi_n^{-i}) - \overline{g}_n^i + W_{n+1}^{i,4}\right), \end{aligned}$$

where $W_{n+1}^{i,4} = G^i(s_{n+1}^i, s_{n+1}^{-i}) - G^i(\pi_n^i, \pi_n^{-i})$.

Let $\mathbf{G}$ be the convex hull in $\mathbb{R}^2$ of the set

$$\{(G^1(s,r), G^2(s,r)) \,:\, s \in S^1\,,\, r \in S^2\}$$

and let $\Sigma = \Delta(S^1) \times \Delta(S^2) \times \mathbf{G}$. We define the set-valued map $\mathbf{C} : \Sigma \to \Sigma$ such that $\mathbf{C}(z)$ is given by

$$\left\{(\alpha^1, \alpha^2, \gamma) : \alpha^1 \in \mathrm{BR}^1(v^2), \alpha^2 \in \mathrm{BR}^2(v^1), \gamma = (G^1(\alpha^1, \alpha^2), G^2(\alpha^1, \alpha^2))\right\}$$

for $z = (v^1, v^2, \overline{g}^1, \overline{g}^2) \in \Sigma$, and we consider the differential inclusion

$$\dot{z} \in -z + \mathbf{C}(z). \tag{4.6}$$

Let $\overline{\mu}_n = (\pi_n^1, \pi_n^2, (G^1(\pi_n^1, \pi_n^2), G^2(\pi_n^1, \pi_n^2))$. From Theorem 2.7, the map $\mathbf{C}$ is adapted to $(z_n, \overline{\mu}_n)$. Proposition A.7(ii) shows that $\epsilon(W_{n+1}^{i,4}/(n+1), T)$ goes to zero almost surely for all fixed $T > 0$. Therefore, by writing the evolution of z_n in the same manner as for v_n before, we can conclude that the limit set of the sequence $(z_n)_n$ is an ICT set of the differential inclusion (4.6).

Zero-sum games. Hofbauer and Sorin [24] (by exhibiting an explicit Lyapunov function) show that the set of Nash equilibria is a global attractor for the differential inclusion (BRD). Hence, if we denote by g_* the value of the game, a direct consequence is that

$$\{(v^1, v^2, g^1, g^2) \,:\, v^1 \ \in \mathrm{BR}^1(v^2)\,,\, v^2 \in \mathrm{BR}^2(v^1)\,,\, (g^1, g^2) = (g_*, -g_*))\}$$

is a global attractor for (4.6). Therefore, $(v_n)_n$ converges to the set of Nash equilibria and $\overline{g}_n^1$ converges to the value of the game.

Potential games. In the same spirit as above, Φ is a Lyapunov function for the differential inclusion (4.6) (see [6, Theorem 5.5]). Since, in our case, the payoff functions are linear in all variables, Propositions 3.27 and 3.28 in Benaïm, Hofbauer, and Sorin [6] imply that $(v_n)_n$ converges almost surely to a connected component of Nash equilibria on which the potential Φ is constant. In particular, if $G^1 = G^2$, let G^* be the value of G on the limit set of $(v_n)_n$. Then, $\lim_n G(v_n^1, v_n^2) = G^*$. Therefore, by definition of $\mathbf{C}$, we also have $\lim_n \overline{g}_n^1 = G^*$.

$2 \times N$ games. Our result follows from the fact that any trajectory of the best-response dynamics converges to the set of Nash equilibria in this case (see Berger [9]). ∎

Appendix A. Technical results. While Assumption 2.5 is used here, in fact, the proofs are written in such a way that they can be easily extended to the case where the less stringent Assumption 2.11 is considered on the sequences $(\beta_n^i)_n$.

A.1. A general result. Returning to the framework of section 3.1, we consider a discrete time stochastic process $(s_n, M_n)_n$, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which is equipped with a nondecreasing sequence of σ-algebras $(\mathcal{F}_n)_n$. The process $(s_n, M_n)_n$ takes values in $S \times \mathcal{M}(S)$ and satisfies Assumption 3.1. Let Σ be a compact convex nonempty set which is assumed, for simplicity, to be contained in $\mathbb{R}^{|S|}$. It will become clear that the argument extends to the case of arbitrary euclidean spaces.

As before, let $H : S \to \Sigma$, $V_n = H(s_n)$ and μ_n be defined by (3.1). The pseudoinverse matrix of M_n is denoted by Q_n (see (3.5)). The following technical proposition will be key to our main result.

PROPOSITION A.1. *Let $(\varepsilon_n)_n$ be a real random process which is adapted to $(\mathcal{F}_n)_n$. Let us assume that, almost surely,*

(i) $|\varepsilon_n||Q_n| \leq n^a$ *for* $a < 1/2$ *and* n *large,*
(ii) $|Q_n||\varepsilon_n - \varepsilon_{n-1}| \to 0$ *,*
(iii) $|\varepsilon_n|\,(|Q_{n+1} - Q_n| + |\pi_{n+1} - \pi_n|) \to 0$.

Let $W_{n+1} = \varepsilon_n\,(V_{n+1} - \mu_n)$. Then, for all $T > 0$, $\epsilon\,(W_{n+1}/(n+1), T) \to 0$, almost surely as n goes to infinity.

Proof. Let c be a positive constant that may change from line to line. In a similar manner as in the proof of [7, Theorem 2.6], we can decompose the noise term as follows:

$$\begin{aligned}
\frac{1}{n+1} W_{n+1} &= \frac{\varepsilon_n}{n+1}(V_{n+1} - \mu_n), \\
&= \frac{\varepsilon_n}{n+1}(H(s_{n+1}) - \mu_n), \\
&= \frac{\varepsilon_n}{n+1}\left(H(s_{n+1}) - \sum_{s \in S} \pi_n(s) H(s)\right).
\end{aligned}$$

For a matrix $A \in \mathbb{R}^{|S|\times|S|}$ and for any $r \in S$, let $A[r]$ be the rth line of A. Let us identify the function H with the matrix $\mathbf{H}$ where, for each $r \in S$, $\mathbf{H}[r] = H(r)$. Notice that, by definition of the matrix Π_n (see (3.5)), we have that $\Pi_n\mathbf{H}[r] = \mu_n$ for every $r \in S$. Therefore, we can write

$$\begin{aligned}\frac{1}{n+1}W_{n+1} &= \frac{\varepsilon_n}{n+1}\left((I-\Pi_n)\mathbf{H}\right)[s_{n+1}]\\ &= \frac{\varepsilon_n}{n+1}\left((Q_n - M_nQ_n)\mathbf{H}\right)[s_{n+1}],\\ &= \frac{\varepsilon_n}{n+1}\left((Q_n\mathbf{H})[s_{n+1}] - (M_nQ_n\mathbf{H})[s_{n+1}]\right),\\ &= \sum_{j=1}^{4} u_n^j,\end{aligned}$$

where the second identity follows from the definition of the pseudoinverse matrix, and

$$\begin{aligned}u_n^1 &= \frac{\varepsilon_n}{n+1}\left((Q_n\mathbf{H})[s_{n+1}] - (M_nQ_n\mathbf{H})[s_n]\right),\\ u_n^2 &= \frac{\varepsilon_n}{n+1}(M_nQ_n\mathbf{H})[s_n] - \frac{\varepsilon_{n-1}}{n}(M_nQ_n\mathbf{H})[s_n],\\ u_n^3 &= \frac{\varepsilon_{n-1}}{n}(M_nQ_n\mathbf{H})[s_n] - \frac{\varepsilon_n}{n+1}(M_{n+1}Q_{n+1}\mathbf{H})[s_{n+1}],\\ u_n^4 &= \frac{\varepsilon_n}{n+1}(M_{n+1}Q_{n+1}\mathbf{H})[s_{n+1}] - \frac{\varepsilon_n}{n+1}(M_nQ_n\mathbf{H})[s_{n+1}],\\ &= \frac{\varepsilon_n}{n+1}\left(M_{n+1}Q_{n+1} - M_nQ_n\right)\mathbf{H}[s_{n+1}].\end{aligned}$$

Since $\mathbb{E}((Q_n\mathbf{H})[s_{n+1}] \mid \mathcal{F}_n) = (M_nQ_n\mathbf{H})[s_n]$, the random process u_n^1 is a martingale difference and

$$\left\|u_n^1\right\| \le c\frac{|\varepsilon_n||Q_n|}{n+1}.$$

The exponential martingale inequality (see equation (18) in Benaïm [3]) gives that, for all $K > 0$,

$$\mathbb{P}(\epsilon(u_n^1, T) \ge K) \le c\exp\left(\frac{-K^2}{c\sum_{j=n}^{m(\tau_n+T)}\varepsilon_j^2|Q_j|^2/j^2}\right).$$

By assumption we have that, almost surely and for j large enough, $|\varepsilon_j||Q_j| \le j^a$ for $a < 1/2$, so that

$$c\sum_{j=n}^{m(\tau_n+T)}\frac{\varepsilon_j^2|Q_j|^2}{j^2} \le c\frac{1}{n^{1-2a}}\sum_{j=n}^{m(\tau_n+T)}\frac{1}{j} \le c\frac{T+1}{n^{1-2a}},$$

by definition of $m(t)$. Therefore,

$$\mathbb{P}(\epsilon(u_n^1, T) \ge K) \le c\exp\left(-\frac{K^2}{T+1}n^{1-2a}\right).$$

Finally, from the fact that $a < 1/2$, we have $\sum_{n\ge1}\mathbb{P}(\epsilon(u_k^1, T) \ge K) < +\infty$ for all $K > 0$, and the Borel–Cantelli lemma implies that $\epsilon(u_n^1, T) \to 0$ almost surely.

For the second term,

$$\begin{aligned}
\epsilon(u_n^2,T) &\le c \sum_{j=n}^{m(\tau_n+T)} |Q_j| \left| \frac{\varepsilon_{j-1}}{j} - \frac{\varepsilon_j}{j+1} \right|, \\
&= c \sum_{j=n}^{m(\tau_n+T)} |Q_j| \left| \frac{(j+1)\varepsilon_{j-1} - j\varepsilon_j}{j(j+1)} \right|, \\
&= c \sum_{j=n}^{m(\tau_n+T)} |Q_j| \left| \frac{\varepsilon_{j-1} - \varepsilon_j}{j} + \frac{\varepsilon_j}{j(j+1)} \right|, \\
&\le c \left[\sup_{j\ge n} |Q_j||\varepsilon_j - \varepsilon_{j-1}| + \sup_{j\ge n} \frac{|Q_j||\varepsilon_j|}{j+1} \right] \sum_{j=n}^{m(\tau_n+T)} \frac{1}{j}, \\
&\le c \left[\sup_{j\ge n} |Q_j||\varepsilon_j - \varepsilon_{j-1}| + \sup_{j\ge n} \frac{|Q_j||\varepsilon_j|}{j+1} \right] (T+1),
\end{aligned}$$

by definition of $m(t)$. Hence, from assumptions (i) and (ii), we conclude that $\epsilon(u_n^2,T)$ goes to zero almost surely.

Now for u_n^3, by cancellation of successive terms,

$$\begin{aligned}
\epsilon(u_n^3,T) &= \frac{\varepsilon_{n-1}}{n}(M_nQ_n\mathbf{H})[s_n] - \frac{\varepsilon_{m(\tau_n+T)-1}}{m(\tau_n+T)}(M_{m(\tau_n+T)}Q_{m(\tau_n+T)}\mathbf{H})[s_{m(\tau_n+T)}], \\
&\le 2\sup_{j\ge n} \frac{|Q_j||\varepsilon_{j-1}|}{j},
\end{aligned}$$

which implies, by (i), that $\epsilon(u_n^3,T) \to 0$ almost surely.

For the fourth term, recall that $M_nQ_n = Q_n - I + \Pi_n$ for all $n \in \mathbb{N}$. Therefore, we can write

$$u_n^4 = \frac{\varepsilon_n}{n+1} \left(Q_{n+1} - Q_n - (\Pi_{n+1} - \Pi_n) \right) \mathbf{H}[s_{n+1}].$$

Hence,

$$\begin{aligned}
\epsilon(u_n^4,T) &\le c \sum_{j=n}^{m(\tau_n+T)} \frac{1}{j} \left[\sup_{j\ge n} |\varepsilon_j||Q_{j+1} - Q_j| + |\varepsilon_j||\pi_{j+1} - \pi_j| \right], \\
&\le c(T+1) \left[\sup_{j\ge n} |\varepsilon_j||Q_{j+1} - Q_j| + |\varepsilon_j||\pi_{j+1} - \pi_j| \right].
\end{aligned}$$

Assumption (iii) implies that $\epsilon(u_n^4,T) \to 0$ almost surely, as n goes to infinity. ∎

A.2. Stability. The following lemma is a trivial consequence of the recursive definition of the vector R_n^i and the fact that $\gamma_n^i(s) \in]0,1]$.

LEMMA A.2. *For any $i \in \{1,2\}$, $s \in S^i$, $n \in \mathbb{N}$, we have $R_n^i(s) \in [-K^i, K^i]$, where*

$$K^i = \max \left\{ \max_{s,r\in S^i} |R_0^i(s) - R_0^i(r)|, \max_{s\in S^i} \max_{s^{-i},r^{-i}\in S^{-i}} |G^i(s,s^{-i}) - G^i(s,r^{-i})| \right\}. \tag{A.1}$$

The following result states that, without loss of generality, we can suppose that the step size $\gamma_n^i(s)$ is equal to $(n\pi_{n-1}^i(s))^{-1}$ for all $s \in S^i$.

LEMMA A.3. *Let $\alpha \in]0,1[$. There exists $n_0(\alpha) \in \mathbb{N}$ (which only depends on α, R_0^i, the payoff functions G^i, and the vanishing sequence $(A_n^i)_n$) such that, for any*

$n \geq n_0(\alpha)$ *and* $s \in S^i$, $\pi_n^i(s) \geq n^{-\alpha}$. *In particular, there exists* $n_0 \in \mathbb{N}$ *such that, for any* $n \geq n_0$ *and* $s \in S^i$, $(\gamma_n^i(s))^{-1} = n\pi_{n-1}^i(s)$.

Proof. Let $\alpha \in]0,1[$ and $\alpha' \in]0,\alpha[$. Choose $n_0 \in \mathbb{N}$ such that, for any $n \geq n_0$, $2K^i A_n^i \leq \alpha'$, where K^i is defined in (A.1), and take $r_n \in S^i$ such that $R_n^i(r_n) = \max_r R_n^i(r)$. Then, for any $s \in S^i$,

$$\pi_n^i(s) = \frac{\pi_0^i(s)\exp\left(A_n^i \ln(n)(R_n^i(s) - R_n^i(r_n))\right)}{\pi_0^i(r_n) + \sum_{r \neq r_n} \pi_0^i(r)\exp\left(A_n^i \ln(n)(R_n^i(r) - R_n^i(r_n)\right)}$$
$$\geq \pi_0^i(s)\exp\left(-2A_n^i \ln(n)K^i\right) \geq \min_r \pi_0^i(r)\exp(-\alpha' \ln(n)) \geq \min_r \pi_0^i(r) n^{-\alpha'}.$$

Without loss of generality, we can assume that n_0 is large enough so that $\min_r \pi_0^i(r)$ $n^{-\alpha'} \geq n^{-\alpha}$. This concludes the proof of the first point. In particular, there exists $n_0 \in \mathbb{N}$ such that, for any $n \geq n_0$, $(n+1)\pi_n^i(s) > 1$, which proves the second point. ∎

A.3. Analysis of the noise sequences. Let us fix $i \in \{1,2\}$ and let χ_n^i be the spectral gap of the matrix $M_n^i = M[\beta_n^i, R_n^i]$, i.e.,

$$\chi_n^i = \min\left\{\frac{\mathcal{E}_n^i(f,f)}{\mathrm{var}_n^i(f)} \,:\, \mathrm{var}_n^i(f) \neq 0\right\},$$

where

$$\mathrm{var}_n^i(f) = \sum_{s \in S^i} \pi_n^i(s) f^2(s) - \left(\sum_{s \in S^i} \pi_n^i(s) f(s)\right)^2,$$
$$\mathcal{E}_n^i(f,f) = \frac{1}{2}\sum_{s,r \in S^i} (f(s) - f(r))^2 M_n^i(s,r)\pi_n^i(s).$$

The following result is a direct consequence of results of Holley and Stroock [25].

LEMMA A.4. *There exists a positive constant* c *such that, for a sufficiently large* $n \in \mathbb{N}$

$$c\exp(-2K^i\beta_n^i) \leq \chi_n^i,$$

where K^i *is defined in Lemma* A.2.

Proof. By [25, Lemma 2.7], for sufficiently large n, $\chi_n^i \geq c\exp(-\beta_n^i \mathbf{m}_n)$, where

$$\mathbf{m}_n = \max_{s,r \in S^i}\left\{\min_{\gamma \in \Gamma}\max_{s' \in \gamma} R_n^i(s') - R_n^i(s) - R_n^i(r) + \min_{s' \in S^i} R_n^i(s')\right\},$$

and Γ is the set of every path from s to r on the graph that represents the action set of player i. Now it is clear that $\mathbf{m}_n \leq 2K^i$, by Lemma A.2. ∎

LEMMA A.5. *Under Assumption* 2.5, *the following holds, almost surely, as* $n \to +\infty$. *Given* $s \in S^i$,

(i) $\dfrac{|Q_n^i|}{n^a \pi_n^i(s)^b} \longrightarrow 0$ *for any* $a > 0, b > 0$,

(ii) $\dfrac{|Q_{n+1}^i - Q_n^i| n^{1-\alpha}}{\pi_n^i(s)} \longrightarrow 0$ *and* $\dfrac{|\pi_{n+1}^i - \pi_n^i| n^{1-\alpha}}{\pi_n^i(s)} \longrightarrow 0$ *for any* $\alpha > 0$.

Proof. Let c be a general positive constant that may change from line to line.

(i) The first inequality in [7, Proposition 3.4] (based on estimations obtained by Saloff-Coste [35]) reads in this case, for $n \in \mathbb{N}$ and $s, s' \in S^i$,

$$|Q_n^i(s,s')| \leq \frac{1}{\chi_n^i}\left(\frac{\pi_n^i(s')}{\pi_n^i(s)}\right)^{1/2} \leq \frac{1}{\chi_n^i}(\pi_n^i(s))^{-1/2}. \tag{A.2}$$

Let $a > 0$ and $b > 0$. By Lemma A.4, $(\chi_n^i)^{-1} \leq c^{-1} n^{2K^i A_n^i}$. Pick $\frac{a}{b+1/2} > \alpha > 0$. There exists $n_0(\alpha)$ such that, for any $n \geq n_0$, for any $s \in S^i$, $\pi_n^i(s) \geq n^{-\alpha}$. Therefore, for sufficiently large n,

$$\frac{|Q_n^i|}{n^a \pi_n^i(s)^b} \leq c^{-1}\frac{n^{2K^i A_n^i + \alpha/2}}{n^a n^{-b\alpha}} = c^{-1} n^{2K^i A_n^i + \alpha(1/2+b)-a}.$$

Thus, the conclusion follows from the fact that $\alpha(1/2+b) - a < 0$ and $\lim_n A_n^i = 0$.

(ii) Let $\alpha > 0$. Recall that $M_n^i = M^i[\beta_n^i, R_n^i]$. Therefore,

$$\begin{aligned} |M_{n+1}^i - M_n^i| \leq & \left|M^i[\beta_{n+1}^i, R_{n+1}^i] - M^i[\beta_n^i, R_{n+1}^i]\right| \\ & + \left|M^i[\beta_n^i, R_{n+1}^i] - M^i[\beta_n^i, R_n^i]\right|. \end{aligned}$$

A simple application of the mean value theorem on the functions

$$\beta \to M^i[\beta, R] \text{ and } R \to M^i[\beta, R]$$

yields, respectively,

$$\left|M^i[\beta_{n+1}^i, R_{n+1}^i] - M^i[\beta_n^i, R_{n+1}^i]\right| \leq c\frac{A_n^i}{n}$$

and

$$\left|M^i[\beta_n^i, R_{n+1}^i] - M^i[\beta_n^i, R_n^i]\right| \leq c\beta_n^i |R_{n+1}^i - R_n^i|,$$

By Lemma A.3, and since $|R_{n+1}^i - R_n^i| \leq \max_{s\in S^i} c\gamma_{n+1}^i(s)$, we have that

$$\left|M_{n+1}^i - M_n^i\right| \leq \frac{1}{n^{1-\alpha/4}}$$

for sufficiently large n. Analogously, recalling that $\pi_n^i = \pi^i[\beta_n^i, R_n^i]$, we have

$$|\pi_{n+1}^i - \pi_n^i| \leq \frac{1}{n^{1-\alpha/4}} \tag{A.3}$$

for sufficiently large n. Recall that, from part (i), $|Q_n^i| \leq n^{\alpha/8}$ for sufficiently large n. Also, $\pi_n^i(s) \geq n^{-\alpha/4}$. Using the last inequality in the proof of [7, Proposition 3.3],

$$|Q_{n+1}^i - Q_n^i| \leq c\left(|Q_{n+1}^i||Q_n^i||M_{n+1}^i - M_n^i| + |Q_n^i||\pi_{n+1}^i - \pi_n^i|\right),$$

we have that

$$\begin{aligned} \frac{|Q_{n+1}^i - Q_n^i| n^{1-\alpha}}{\pi_n^i(s)} &\leq c\frac{n^{1-\alpha}}{n^{-\alpha/4}}\left(|Q_{n+1}^i||Q_n^i||M_{n+1}^i - M_n^i| + |Q_n^i||\pi_{n+1}^i - \pi_n^i|\right) \\ &\leq \frac{1}{n^{\alpha/8}}, \end{aligned}$$

almost surely for sufficiently large n. ∎

The following two propositions establish all the results on the noise terms that we need in the proof of Theorem 2.7 (cf. section 4).

PROPOSITION A.6. *Suppose that Assumption* 2.5 *holds and let* $i \in \{1,2\}$.

(i) *For* $s \in S^i$, *let*

$$W^{i,1}_{n+1}(s) = \frac{R^i_n(s)}{\pi^i_n(s)}\left(\mathbb{1}_{\{s^i_{n+1}=s\}} - \pi^i_n(s)\right) \in \mathbb{R}.$$

Then, for all $T > 0$, $\epsilon(W^{i,1}_{n+1}(s)/(n+1), T) \to 0$ *almost surely as* n *goes to infinity.*

(ii) *Let*

$$\overline{W}^i_{n+1} = \delta_{s^i_{n+1}} - \pi^i_n \in \mathbb{R}^{|S^i|}.$$

Then, for all $T > 0$, $\epsilon(\overline{W}^{i,1}_{n+1}/(n+1), T) \to 0$ *almost surely as* n *goes to infinity.*

(iii) *Let*

$$W^{i,3}_{n+1} = G^i(\cdot, s^{-i}_{n+1}) - G^i(\cdot, \pi^{-i}_n) \in \mathbb{R}^{|S^i|}.$$

Then, for all $T > 0$, $\epsilon(W^{i,3}_{n+1}/(n+1), T) \to 0$ *almost surely as* n *goes to infinity.*

Proof. We prove part (i) in detail. Given that the arguments are very similar, the remaining proofs are omitted.

We apply Proposition A.1 with $S = S^i$, $\Sigma = \Delta(S^i)$, $s_n = s^i_n$, $M_n = M^i_n$, $\pi_n = \pi^i_n$ and $H(r) = \delta_r$ for all $r \in S^i$. Therefore, in this case $\mu_n = \pi^i_n$ and $V_{n+1} = \delta_{s^i_{n+1}}$. We also put $\varepsilon_n = R^i_n(s)/\pi^i_n(s)$.

From the fact that R^i_n is bounded, it is easy to see that points (i) and (ii) of Lemma A.5, respectively, imply assumptions (i) and (iii) of Proposition A.1. To confirm that assumption (ii) holds, it suffices to compute

$$\begin{aligned}|Q^i_n||\varepsilon_n - \varepsilon_{n-1}| &= \frac{|Q^i_n||\pi^i_n(s)(R^i_n(s) - R^i_{n-1}(s)) + R^i_n(s)(\pi^i_{n-1}(s) - \pi^i_n(s))|}{\pi^i_n(s)\pi^i_{n-1}(s)} \\ &\le c|Q^i_n|n^{-1+\alpha}\end{aligned}$$

by definition of R^i_n, Lemma A.3, and (A.3) for sufficiently large n and any $\alpha > 0$. Hence, by Lemma A.5, $|Q^i_n||\varepsilon_n - \varepsilon_{n-1}|$ goes to zero almost surely as n goes to infinity. By using Proposition A.1, we show that $\epsilon(\mathcal{U}^i_{n+1}/(n+1), T)$ goes to zero almost surely for any $T > 0$, where

$$\mathcal{U}^i_{n+1} = \frac{R^i_n(s)}{\pi^i_n(s)}\left(\delta_{s^i_{n+1}} - \pi^i_n\right) \in \mathbb{R}^{|S^i|}.$$

The result follows from the fact that the sth component of the vector $\mathcal{U}^i_{n+1}$ is equal to $W^{i,1}_{n+1}(s)$. ∎

PROPOSITION A.7. *Suppose that Assumption* 2.5 *holds and let us fix* $i \in \{1,2\}$.

(i) *For* $s \in S^i$, *let*

$$W^{i,2}_{n+1}(s) = \frac{1}{\pi^i_n(s)}\left(\mathbb{1}_{\{s^i_{n+1}=s\}} G^i(s^1_{n+1}, s^2_{n+1}) - \pi^i_n(s) G^i(s, \pi^{-i}_n)\right) \in \mathbb{R}.$$

Then, for all $T > 0$, $\epsilon(W^{i,2}_{n+1}/(n+1), T) \to 0$ *almost surely as* n *goes to infinity.*

(ii) *Let*

$$W^{i,4}_{n+1} = G^i(s^i_{n+1}, s^{-i}_{n+1}) - G^i(\pi^i_n, \pi^{-i}_n) \in \mathbb{R}.$$

Then, for all $T > 0$, $\epsilon(W^{i,4}_{n+1}/(n+1), T) \to 0$ *almost surely as* n *goes to infinity.*

Proof.

(i) For the sake of clarity, let us set $i = 1$. Again, we use Proposition A.1, where in this case, $S = S^1 \times S^2$, $\Sigma \subseteq \mathbb{R}^{|S^1|}$ is defined by

$$\left\{ \sum_{s^2 \in S^2} \sigma^2(s^2) G^1(\cdot, s^2) \, : \, \sum_{s^2 \in S^2} \sigma^2(s^2) = 1 \text{ and } \sigma^2(s^2) \geq 0 \text{ for all } s^2 \in S^2 \right\}.$$

Also, $s_n = (s^1_n, s^2_n)$, $M_n = M^1_n \otimes M^2_n$, $\pi_n = \pi^1_n \otimes \pi^2_n$ and $H : S^1 \times S^2 \to \Sigma$, where $H(s^1, s^2) = \delta_{s^1} G^1(s^1, s^2)$ for all $(s^1, s^2) \in S^1 \times S^2$. Notice that in this case δ is the Kronecker's delta function taking values in $\Delta(S^1)$. Therefore, $\mu_n = (\mu_n(s^1))_{s^1 \in S^1}$ with $\mu_n(s^1) = \pi^1_n(s^1) G^1(s^1, \pi^2_n)$ and $V_{n+1} = (V_{n+1}(s^1))_{s^1 \in S^1}$, where

$$V_{n+1}(s^1) = \mathbb{1}_{\{s^1_{n+1} = s^1\}} G^1(s^1_{n+1}, s^2_{n+1}) = \mathbb{1}_{\{s^1_{n+1} = s^1\}} G^1(s^1, s^2_{n+1}).$$

We also set in this case $\varepsilon_n = 1/\pi^1_n(s)$. Let Q_n be the pseudoinverse matrix of the stochastic matrix M_n. It is easy to see that the spectral gap of M_n verifies that

$$\chi(M_n) = \chi(M^1_n \otimes M^2_n) = \min\{\chi(M^1_n), \chi(M^2_n)\} = \min\{\chi^1_n, \chi^2_n\}.$$

By using inequality (A.2) for the matrix Q_n and the fact that $\pi_n(s^1, s^2) = \pi^1_n(s^1)\pi^2_n(s^2) \geq n^{-\alpha}$ for any $\alpha > 0$ and sufficiently large n, we can obtain exactly the same conclusions as in Lemma A.5 for Q_n and π_n. Hence, as in the proof of Proposition A.6, we deduce that sequences $(\varepsilon_n)_n$ and $(Q_n)_n$ verify assumptions (i)–(iii) of Proposition A.1.

Therefore, we have that $\epsilon(\mathcal{U}^i_{n+1}/(n+1), T)$ goes to zero almost surely for any $T > 0$ where, for $s^1 \in S^1$,

$$\begin{aligned} \mathcal{U}^i_{n+1}(s^1) &= \frac{1}{\pi^1_n(s)} \left(\mathbb{1}_{\{s^1_{n+1} = s^1\}} G^1(s^1, s^2_{n+1}) - \pi^1_n(s^1) G^1(s^1, \pi^2_n) \right) \\ &= \frac{1}{\pi^1_n(s)} \left(\mathbb{1}_{\{s^1_{n+1} = s^1\}} G^1(s^1_{n+1}, s^2_{n+1}) - \pi^1_n(s^1) G^1(s^1, \pi^2_n) \right). \end{aligned}$$

The conclusion follows taking $s^1 = s$ in the equation above.

(ii) The proof of this part also follows from Proposition A.1, taking as Σ a sufficiently large compact set in $\mathbb{R}$, $s_n = (s^1_n, s^2_n)$, $M_n = M^1_n \otimes M^2_n$, $\pi_n = \pi^1_n \otimes \pi^2_n$ and $H : S^1 \times S^2 \to \Sigma$, where $H(s^1, s^2) = G^i(s^1, s^2)$. Therefore, $\mu_n = G^i(\pi^i_n, \pi^{-i}_n)$ and $\varepsilon_n = 1$ for all $n \in \mathbb{N}$. Finally, using the same argument as in part (i), we prove that the assumptions (i)–(iii) hold and we can conclude. ∎

Acknowledgments. This paper was initially motivated by a question from Drew Fudenberg and Satoru Takahashi. The authors would like to thank the Aix-Marseille School of Economics for inviting M. Bravo to work on this project. Both authors are

indebted to Sylvain Sorin for useful discussions, and for inviting M. Faure to Jussieu (Paris 6) in the early stages of this work. Also the authors are grateful to Jérôme Renault, Fabien Gentsbittel, and an anonymous referee for the help improving the presentation of the model.

REFERENCES

[1] J.P. Aubin and A. Cellina, *Differential Inclusions: Set-Valued Maps and Viability Theory*, Springer, Berlin, 1984.
[2] A.W. Beggs, *On the convergence of reinforcement learning*, J. Econom. Theory, 122 (2005), pp. 1–36.
[3] M. Benaïm, *Dynamics of stochastic approximation algorithms*, in Séminaire de Probabilités, XXXIII, Lecture Notes in Math. 1709, Springer, Berlin, 1999, pp. 1–68.
[4] M. Benaïm and M. Faure, *Stochastic approximations, cooperative dynamics and supermodular games*, Ann. Appl. Probab., 22 (2012), pp. 2133–2164.
[5] M. Benaïm and M.W. Hirsch, *Mixed equilibria and dynamical systems arising from fictitious play in perturbed games*, Games Econom. Behav., 29 (1999), pp. 36–72.
[6] M. Benaïm, J. Hofbauer, and S. Sorin, *Stochastic approximations and differential inclusions*, SIAM J. Control Optim., 44 (2005), pp. 328–348.
[7] M. Benaïm and O. Raimond, *A class of self-interacting processes with applications to games and reinforced random walks*, SIAM J. Control Optim., 48 (2010), pp. 4707–4730.
[8] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer, Berlin, 1990.
[9] U. Berger, *Fictitious play in $2 \times n$ games*, J. Econom. Theory, 120 (2005), pp. 134–154.
[10] T. Börgers and R. Sarin, *Learning through reinforcement and replicator dynamics*, J. Econom. Theory, 77 (1997), pp. 1–14.
[11] M. Bravo, *An Adjusted Payoff-Based Procedure for Normal Form Games*, preprint; also available online from http://arxiv.org/abs/1106.5596.
[12] G.W. Brown, *Iterative solution of games by fictitious play*, in Activity Analysis of Production and Allocation, Wiley, New York, 1951, pp. 374–376.
[13] R. Cominetti, E. Melo, and S. Sorin, *A payoff-based learning procedure and its application to traffic games*, Games Econom. Behav., 70 (2010), pp. 71–83.
[14] I. Erev and A.E. Roth, *Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria*, Amer. Econom. Rev., 88 (1998), pp. 848–81.
[15] M. Faure and G. Roth, *Stochastic approximations of set-valued dynamical systems: Convergence with positive probability to an attractor*, Math. Oper. Res., 35 (2010), pp. 624–640.
[16] D. Fudenberg and D.M. Kreps, *Learning mixed equilibria*, Games Econom. Behav., 5 (1993), pp. 320–367.
[17] D. Fudenberg and D.K. Levine, *The Theory of Learning in Games*, MIT Press, Cambridge, MA, 1998.
[18] I. Gilboa and A. Matsui, *Social stability and equilibrium*, Econometrica, 59 (1991), pp. 859–867.
[19] B. Hajek, *Cooling schedules for optimal annealing*, Math. Oper. Res., 13 (1988), pp. 311–329.
[20] J.C. Harsanyi, *Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points*, Internat. J. Game Theory, 2 (1973), pp. 1–23.
[21] S. Hart and A. Mas-Colell, *A simple adaptive procedure leading to correlated equilibrium*, Econometrica, 68 (2000), pp. 1127–1150.
[22] S. Hart and A. Mas-Colell, *A reinforcement procedure leading to correlated equilibrium*, in Economics Essays: A Festschrift for Werner Hildebrand, Springer, Berlin, 2001, pp. 181–200.
[23] J. Hofbauer and W.H. Sandholm, *On the global convergence of stochastic fictitious play*, Econometrica, 70 (2002), pp. 2265–2294.
[24] J. Hofbauer and S. Sorin, *Best response dynamics for continuous zero-sum games*, Discrete Contin. Dyn. Syst. Ser. B, 6 (2006), pp. 215–224.
[25] R. Holley and D. Stroock, *Simulated annealing via Sobolev inequalities*, Comm. Math. Phys., 115 (1988), pp. 553–569.
[26] E. Hopkins, *Two competing models on how people learn in games*, Econometrica, 70 (2002), pp. 2141–2166.
[27] E. Hopkins and M. Posch, *Attainability of boundary points under reinforcement learning*, Games Econom. Behav., 53 (2005), pp. 110–125.

[28] H.J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, Berlin, 2003.
[29] D.S. Leslie and E.J. Collins, *Individual Q-learning in normal form games*, SIAM J. Control Optim., 44 (2005), pp. 495–514.
[30] D.S. Leslie and E.J. Collins, *Generalised weakened fictitious play*, Games Econom. Behav., 56 (2006), pp. 285–298.
[31] K. Miyasawa, *On the Convergence of the Learning Process in a* 2×2 *Non-Zero-Sum Two-Person Game*, Technical report, DTIC, 1961.
[32] D. Monderer and L.S. Shapley, *Fictitious play property for games with identical interests*, J. Econom. Theory, 68 (1996), pp. 258–265.
[33] M. Posch, *Cycling in a stochastic learning algorithm for normal form games*, J. Evol. Econ., 7 (1997), pp. 193–207.
[34] J. Robinson, *An iterative method of solving a game*, Ann. of Math., 54 (1951), pp. 296–301.
[35] L. Saloff-Coste, *Lectures on finite Markov chains*, in Lectures Notes in Math., Springer, Berlin, 1997, pp. 301–413.