Learning a Correlated Equilibrium with Perturbed Regret Minimization

Omar Boufous[1(✉)], Rachid El-Azouzi[2], Mikaël Touati[1], Eitan Altman[2,3], and Mustapha Bouhtou[1]

[1] Orange, Châtillon, France
{omar.boufous,mikael.touati,mustapha.bouhtou}@orange.com
[2] LIA, University of Avignon, Avignon, France
rachid.elazouzi@univ-avignon.fr
[3] INRIA, Sophia Antipolis, Sophia Antipolis, France
eitan.altman@inria.fr

Abstract. In this paper, we consider the problem of learning a correlated equilibrium of a finite non-cooperative game and show a new learning rule, called Correlated Perturbed Regret Minimization (CPRM) for this purpose. CPRM combines regret minimization to approach the set of correlated equilibria and a simple device recommending to the players actions drawn from the empirical distribution in order to further stabilize the dynamic. Numerical experiments support the hypothesis of the pointwise convergence of the empirical distribution over action profiles to an approximate correlated equilibrium with all players following the devices' suggestions. Additional simulation results suggest that an adaptive version of CPRM can handle changes in the game such as departures or arrivals of players.

Keywords: Game theory · Correlated equilibrium · Learning

1 Introduction

Since their introduction [1,2] as a solution concept for non-cooperative games, correlated equilibria have gradually emerged as an appealing generalization of Nash equilibria. Correlated equilibria build upon the idea of correlated strategies, allowing for a non-independent randomization over actions by the players. More formally, a correlated equilibrium is an equilibrium of a game extended with an information structure defined by a probability space and a collection of player-specific events. Some of these structures known as "canonical" [3] lead to an interpretation in terms of a mediator [4] drawing an action profile according to a probability distribution and privately suggesting to each player her component. After receiving this recommendation, the player chooses her action. It was shown in [2] that the canonical structures are sufficient to generate all correlated equilibrium distributions.

E. Hyytiä and V. Kavitha (Eds.): VALUETOOLS 2022, LNICST 482, pp. 14–32, 2023.
https://doi.org/10.1007/978-3-031-31234-2_2

We consider the problem of learning a correlated equilibrium of a non-cooperative game with simple learning procedures and limited knowledge about the game (*e.g.* fictitious play, trial and error, regret-matching) [5]. Particularly, several adaptive heuristics imply the (almost sure) convergence of the empirical distribution over action profiles to the set of correlated equilibria [5–8] under relatively mild assumptions (*e.g.* every player may only know her utility function and the history of play). However, under these assumptions no learning rule implies a convergence to a correlated equilibrium distribution.

The main objective of this paper is to address the latter issue by introducing a new learning rule called Correlated Perturbed Regret Minimization (CPRM). In CPRM, players adaptively alternate between playing a regret minimization strategy to reduce the distance between the empirical distribution over action profiles and the set of correlated equilibrium distributions and following the suggestions of a device sampling from this distribution. In the long-run, the empirical distribution is expected to stabilize close to a correlated equilibrium with players following the device's suggestions.

1.1 Related Work

The majority of the literature on learning in games [9,10] studies the problem of learning pure and mixed Nash equilibria [11–16] with some contributions focusing on the convergence to equilibria satisfying properties such as Pareto efficiency [17,18] or welfare maximization [19,20].

The problem of learning correlated equilibria has received less attention in spite of a growing interest in the topic and the importance of the solution concept. In [6], Hart *et al.* propose a regret minimization strategy (using Blackwell's approachability [21]) and an adaptive heuristic called regret-matching. They show convergence of the empirical probability distribution over action profiles to the *set* of correlated equilibria. Similar guarantees are offered by calibration [8,22] but none of these procedures are known to guarantee the pointwise convergence of the trajectories to equilibrium points. In [23], Greenwald *et al.* present correlated-Q, a multi-agent reinforcement learning algorithm with an equilibrium selection feature in which a linear program is solved at each iteration to compute the polytope of correlated equilibria. Every player must know the game (utilities and sets of actions). In [24], Borowski *et al.* propose an uncoupled learning rule in which players rely on a public signal and select actions according to a perturbed process such that in the long-run, the joint strategy is a correlated equilibrium. The convergence is guaranteed if the public signal satisfies a certain condition which requires knowledge about the set of correlated equilibria. See [5,6,8,25] for other works on learning coarse correlated equilibria. Recent contributions [26,27] consider learning correlated equilibria of games in extensive form.

Finally, from an application perspective, correlated equilibria are relevant in engineering [28,29]. Particularly, [30] shows an algorithm using a correlation signal to synchronize the players' decisions so that they play a correlated equilibrium. However, the proposed approach seems to be limited to the considered system.

1.2 Outline

In Sect. 2, we define the model and provide the necessary preliminaries such as the relationship between correlated equilibria and regrets. In Sect. 3, we present the learning rule. In Sect. 4, we evaluate and discuss numerical performances of our solution. Section 5 concludes and shows possible directions of research and improvements.

2 Preliminaries

2.1 Notations

Vectors and tuples are denoted by small bold letters, matrices and random variables are denoted by capital letters. For $M \in \mathbb{R}^{m \times n}$, $M(i,j)$ denotes the entry in row i and column j and $||M|| = \max_{i,j} |M(i,j)|$. The i^{th} component of $\boldsymbol{x}$ is denoted x_i and $\boldsymbol{x} \geq \boldsymbol{y}$ iff $x_i \geq y_i$ for every i. We use calligraphic capital letters for sets, $|\mathcal{S}|$ is the cardinality of the set $\mathcal{S}$ and $\Delta(\mathcal{S})$ is the simplex on $\mathcal{S}$ interpreted as the set of probability distributions on $\mathcal{S}$. The indicator function of an event A is denoted $\mathbb{1}_A$. We denote by $d(\boldsymbol{x}, \boldsymbol{y})$ the Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{y}$.

2.2 Model

Let $G = (\mathcal{N}, (\mathcal{A}_i)_{i \in \mathcal{N}}, (u_i)_{i \in \mathcal{N}})$ be an n-player finite non-cooperative game with set of players $\mathcal{N}$. The set of actions of player i is $\mathcal{A}_i = \{1, \ldots, l_i\}$ and the set of action profiles is $\mathcal{A} = \prod_{j \in \mathcal{N}} \mathcal{A}_i$. Player i's utility function is $u_i : \mathcal{A} \to \mathbb{R}$ such that her utility for the action profile $\boldsymbol{a} \in \mathcal{A}$ is $u_i(\boldsymbol{a}) = u(a_i, \boldsymbol{a_{-i}})$. By extension, player i's expected utility for a probability distribution $\boldsymbol{q} \in \Delta(\mathcal{A})$ is given by $u_i(\boldsymbol{q}) = \sum_{\boldsymbol{a} \in \mathcal{A}} \boldsymbol{q}(\boldsymbol{a}) u_i(\boldsymbol{a})$. In particular, her utility for the mixed strategy profile $(\boldsymbol{p_1}, \ldots, \boldsymbol{p_n}) \in \prod_{j \in \mathcal{N}} \Delta(\mathcal{A}_i)$ is $u_i(\boldsymbol{p}_i, \boldsymbol{p}_{-i}) = \sum_{a \in \mathcal{A}} \prod_{j \in \mathcal{N}} \boldsymbol{p}_i(a_i) u_i(a_i, \boldsymbol{a_{-i}})$.

Assume that G is played repeatedly at discrete times $t = 1, 2, \ldots$. A history of play until time t, denoted $\boldsymbol{h}^t$, is a tuple of action profiles $\boldsymbol{h}^t = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^t) \in \prod_{\tau=1}^{t} \mathcal{A}$ where $\boldsymbol{a}^\tau$ is the action profile played at time τ.

Furthermore, assume that every player i knows her utility function u_i but not necessarily the utility function of other players and that at any time $t+1$ every player knows the history of play $\boldsymbol{h}^t$.

2.3 Correlated Equilibria

In this paper, we consider the concept of correlated equilibrium characterized by a probability distribution over action profiles, commonly interpreted as a distribution of play instructions, such that for each player, a recommended action is a best-response to the other players' actions assuming they follow their recommendations.

Definition 1 (Correlated β-equilibrium, [5]). *A probability distribution* $\boldsymbol{q} \in \Delta(\mathcal{A})$ *is a correlated* β*-equilibrium if*

$$\forall i \in \mathcal{N}, \forall j \in \mathcal{A}_i, \forall k \in \mathcal{A}_i \quad \sum_{\boldsymbol{a} \in \mathcal{A}: a_i = j} \boldsymbol{q}(\boldsymbol{a})[u_i(k, \boldsymbol{a}_{-i}) - u_i(\boldsymbol{a})] \leq \beta. \tag{1}$$

If $\beta = 0$, we have the usual definition of a correlated equilibrium [2].

As an example, consider the traffic intersection game involving two drivers arriving at an intersection. Each driver can either cross the intersection (action "*Go*") or wait (action "*Wait*"). When both drivers cross simultaneously, a collision occurs and both incur a utility of -1. Otherwise, the driver crossing gets $+1$ while the one waiting has 0. Utilities are shown in Table 1. There are two pure strategy Nash equilibria $(Wait, Go)$ with utilities $(0, 1)$, $(Go, Wait)$ with utilities $(1, 0)$ and a mixed Nash equilibrium $((1/2, 1/2), (1/2, 1/2))$ with utilities $(0, 0)$. Thus, the pure Nash equilibria result in unfair utility vectors and the mixed equilibrium is fair but inefficient. Correlated equilibria can help solving this problem by stabilizing more fair and efficient utility vectors. As an example, the probability distribution such that $\mathbb{P}(Go, Wait) = \mathbb{P}(Wait, Go) = 1/2$, $\mathbb{P}(Go, Go) = \mathbb{P}(Wait, Wait) = 0$ is a correlated equilibrium with utilities $(1/2, 1/2)$.

Table 1. Traffic intersection game.

	Wait	Go
Wait	(0, 0)	(0, 1)
Go	(1, 0)	(–1, –1)

2.4 Regret Minimization and Correlated Equilibria

Let $\boldsymbol{h}^t = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^t)$ be the history of play until time t with empirical distribution of play $\boldsymbol{q}^t$ such that,

$$\forall \boldsymbol{a} \in \mathcal{A}, \quad \boldsymbol{q}^t(\boldsymbol{a}) = \frac{1}{t} \sum_{\tau=1}^{t} \mathbb{1}_{\boldsymbol{a}^\tau = \boldsymbol{a}} \tag{2}$$

Following [6], define $D_i^t(j, k)$ the average utility difference for player i until time t when playing action k instead of j,

$$D_i^t(j, k) = \frac{1}{t} \sum_{\tau \leq t: a_i^\tau = j} \left[u_i\left(k, \boldsymbol{a}_{-i}^\tau\right) - u_i\left(j, \boldsymbol{a}_{-i}^\tau\right) \right] \tag{3a}$$

$$= \sum_{\boldsymbol{a} \in \mathcal{A}: a_i = j} \boldsymbol{q}^t(\boldsymbol{a}) \left[u_i(k, \boldsymbol{a}_{-i}) - u_i(\boldsymbol{a}) \right] \tag{3b}$$

and regret $R_i^t(j, k) = \max\{0, D_i^t(j, k)\}$ which is the average gain of utility player i could have obtained if he had played k instead of j at previous iterations.

Remark 1. For any $\beta \geq 0$, we have $R_i^t(j,k) \leq \beta$ if and only if $D_i^t(j,k) \leq \beta$. Hence, from the definition of $D_i^t(j,k)$ in Eq. (3b) we have, $R_i^t(j,k) \leq \beta$ if and only if the empirical distribution $\boldsymbol{q}^t$ is a correlated β-equilibrium.

As proved in [6], Proposition 1 below shows that regrets converging to zero is necessary and sufficient for the sequence of empirical distributions to converge to the set of correlated equilibria.

Proposition 1. *Let* $(\boldsymbol{a}^t)_{t=1,2,\ldots}$ *be a sequence of plays (*i.e. $\boldsymbol{a}^t \in \mathcal{A}$ *for all* t*) and let* $\beta \geq 0$*. Then:* $\limsup_{t\to\infty} R_i^t(j,k) \leq \beta$ *for every* $i \in \mathcal{N}$ *and every* $j,k \in \mathcal{A}_i$ *with* $j \neq k$*, if and only if the sequence of empirical distributions* q^t *converges to the set of correlated* β*-equilibria.*

Furthermore, in [6] Hart *et al.* use Blackwell's approachability result to show that if at every iteration $t+1$ player i chooses an action drawn from the mixed strategy $\boldsymbol{\xi}_i^t$ satisfying,

$$\forall j \in \mathcal{A}_i, \sum_{k\in\mathcal{A}_i} \boldsymbol{\xi}_i^t(k) R_i^t(k,j) = \boldsymbol{\xi}_i^t(j) \sum_{k\in\mathcal{A}_i} R_i^t(j,k) \tag{4}$$

then $\{R_i^t\}_t$ approaches the negative orthant $\mathbb{R}_-^{|\mathcal{A}_i|\times|\mathcal{A}_i|}$,

$$\lim_{t\to\infty} d\left(R_i^t, \mathbb{R}_-^{|\mathcal{A}_i|\times|\mathcal{A}_i|}\right) \to 0 \text{ a.s} \tag{5}$$

Thus, if all players implement such strategy, the sequence of empirical distributions converges a.s. to the set of correlated β-equilibria.

In the next section, we propose a perturbed variant of this strategy, in which players synchronize probabilistically as they approach the set of correlated equilibria to stabilize the sequence of empirical distributions $\{\boldsymbol{q}^t\}_t$.

3 Learning Rule

Assume the history of play $\boldsymbol{h}^t$ with empirical distribution $\boldsymbol{q}^t$. In CPRM (Correlated Perturbed Regret Minimization), each player implements a mood-dependent strategy [12] such that in one mood, she uses a regret minimization strategy (to decrease her regrets and contribute in decreasing the distance to the set of correlated equilibria) and in the other she plays her component of an action profile sampled from $\boldsymbol{q}^t$ by a device.

3.1 Device

Assume a device drawing at time $t+1$ an action profile from the empirical distribution $\boldsymbol{q}^t$ and recommending to each player her component. Thus, if $\boldsymbol{b}^{t+1}$ is drawn, player i receives recommendation b_i^{t+1}. This device must know $\boldsymbol{q}^t$ at $t+1$ (not necessarily storing the history of play $\boldsymbol{h}^t$), must be able to transmit her component to every player but typically cannot access any other information about the game such as utility functions (thus being "unable to compute" a correlated equilibrium of G).

3.2 Players' Strategies

At time $t+1$, player i's mood m_i^{t+1} can be synchronous (denoted *syn*) or asynchronous (denoted *asyn*). If $m_i^{t+1} = syn$, then i plays b_i^{t+1} (sent by the device, see Sect. 3.1), else $m_i^{t+1} = asyn$ and player i chooses an action using the probability distribution $\boldsymbol{\xi}_i^t$ satisfying Eq. (4).

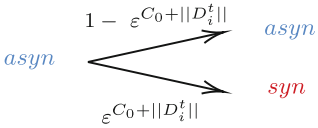
3.3 Players' Moods Dynamic

Given m_i^t and $\boldsymbol{q}^t$, player i's mood evolves according to the following transition probabilities

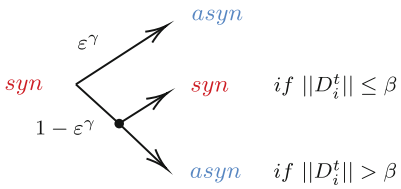
$$\mathbb{P}(m_i^{t+1}|m_i^t, \boldsymbol{q}^t) = \begin{cases} \varepsilon^{C_0+||D_i^t||} & \text{if } m_i^{t+1}{=}syn \text{ and } m_i^t{=}asyn \\ 1-\varepsilon^{C_0+||D_i^t||} & \text{if } m_i^{t+1}{=}asyn \text{ and } m_i^t{=}asyn \\ \varepsilon^{\gamma} & \text{if } m_i^{t+1}{=}asyn \text{ and } m_i^t{=}syn \\ 1-\varepsilon^{\gamma} & \text{if } m_i^{t+1}{=}syn, m_i^t{=}syn \text{ and } ||D_i^t||{\leq}\beta \\ 1-\varepsilon^{\gamma} & \text{if } m_i^{t+1}{=}asyn, m_i^t{=}syn \text{ and } ||D_i^t||{>}\beta \end{cases} \tag{6}$$

where D_i^t is induced by $\boldsymbol{q}^t$ as defined in Eq. (3b) and the scalar parameter $\varepsilon > 0$ is called perturbation or noise of the process [31]. We show below a graphical representation of the moods dynamic at time t.

– If $m_i^t = asyn$ then player i moves to mood *syn* with probability $\varepsilon^{C_0+||D_i^t||}$,



– If $m_i^t = syn$ then with probability ε^{γ} player i "experiments" and switches to *asyn* to play the regret minimization procedure. Otherwise, with probability $1-\varepsilon^{\gamma}$, if her maximum regret is lower than β she stays in mood *syn* and keeps on following the device's recommendation, else she changes to mood *asyn* to play the regret minimization procedure.



The parameter $C_0 > 0$ is introduced to control the sensitivity of the process to regrets such that the transitions probabilities from *asyn* to *syn* tend to zero as ε tends to zero. The constant γ is a parameter controlling the experimentation rate. Finally, β is a parameter defining the maximum "acceptable" value of regret for players to stay *syn* (relaxing in some sense the negative orthant condition of the Blackwell's regret minimization strategy). This parameter controls the

approximation factor of the approximate correlated equilibria (see Sect. 2.3) that are expected to be stabilized by CPRM.

It can be shown that CPRM induces a perturbed non-homogeneous Markov process on a countable state space. We conjecture that if this process admits an asymptotic stationary distribution and the parameters are such that $\gamma > n2^{n-1}(C_0 + \beta)$, then the set of states with all-synchronous players and regrets below β is the only stochastically stable set of states [31]. The latter implies that in the long-run the players follow the suggestions of the device. Moreover, we expect the sequence $\{\boldsymbol{q}^t\}_t$ to converge, implying that in the long-run the device draws action profiles from a correlated β-equilibrium distribution. A detailed analysis of this process is beyond the scope of the paper.

We report in Appendix A an implementable version of CPRM algorithm as a pseudo-code.

3.4 Adaptive CPRM

The previous model and learning rule assume that the same stage game G is played at every iteration. However, in many applications such as networks and communication systems [32], the set of players or utilities may change in time. This issue has not yet been thoroughly investigated in the literature. For instance, [33] assumes that every player may leave the game with a certain probability but is immediately replaced by a new player such that the total number of players is constant. In this section, we assume that the number of players may change in time and propose an adaptive version of CPRM, assuming that at any iteration, every player and the device (see Sect. 3.1) know the set of players in the game and the sets of action profiles.

In the following, we define an adjustment process preventing the learning phase from being re-initialized for the new stage game. First, the empirical distribution in the new stage game is updated in accordance with past plays. Second, every player plays following CPRM but using the updated empirical distribution. Third, following the latter first play in the new game, the empirical distribution is updated using a slightly modified update rule including an inertia parameter.

Let $\mathcal{N}^t$ be the set of players at time t and $\mathcal{A}^t$ be the corresponding set of action profiles. Assume that a new player l arrives at time $T \in]t, t+1[$ such that $\mathcal{N}^{t+1} = \mathcal{N}^t \cup \{l\}$ and $\mathcal{A}^{t+1} = \prod_{i \in \mathcal{N}^{t+1}} \mathcal{A}_i$. Let a_l be an arbitrary action in $\mathcal{A}_l$. To keep on relying on the properties of the empirical distribution induced by the learning rule until t, we define a new "empirical" probability distribution $\tilde{\boldsymbol{q}}^t$ in $\Delta(\mathcal{A}^{t+1})$ to be used by the players at $t+1$ such that, for any profile $\boldsymbol{a} \in \mathcal{A}$ and any action j in $\mathcal{A}_l$,

$$\tilde{\boldsymbol{q}}^t(\boldsymbol{a}, j) = \begin{cases} \boldsymbol{q}^t(\boldsymbol{a}) & \text{if } j = a_l \\ 0 & \text{else} \end{cases} \tag{7}$$

where $(\boldsymbol{a}, j)$ is the strategy profile in $\mathcal{A}^{t+1}$ such that the players in $\mathcal{N}^t$ play $\boldsymbol{a}$ and l plays j.

Assume that player l leaves at time $T \in]t, t+1[$. Then, at time $t+1$ we have $\mathcal{N}^{t+1} = \mathcal{N}^t \backslash \{l\}$, $\mathcal{A}^{t+1} = \prod_{i \in \mathcal{N}^{t+1}} \mathcal{A}_i$ and the empirical distribution must be updated to be in $\Delta(\mathcal{A}^{t+1})$. We consider the following update rule,

$$\forall \boldsymbol{a} \in \mathcal{A}^{t+1}, \tilde{\boldsymbol{q}}^t(\boldsymbol{a}) = \sum_{y \in \mathcal{A}_j} \boldsymbol{q}^t(\boldsymbol{a}, y) \tag{8}$$

Assume that the following arrival or departure occurs at $T' > t+1$, then for any integer k such that $t+1 \leq t+k \leq T'$ the empirical distribution is updated at $t+k$ such that,

$$\tilde{\boldsymbol{q}}^{t+k}(\boldsymbol{a}) = \frac{t+k-1-\tau(t)}{t+k-\tau(t)} \tilde{\boldsymbol{q}}^{t+k-1}(\boldsymbol{a}) + \frac{1}{t+k-\tau(t)} \mathbb{1}_{\{\boldsymbol{a}^{t+k-1} = \boldsymbol{a}\}} \tag{9}$$

where $0 \leq \tau(t) \leq t$ is an inertia parameter at t controlling the responsiveness of the empirical distribution and learning to changes in the game. If $k = 1$ and $\tau(t) = t$, then $\tilde{\boldsymbol{q}}^{t+1}$ equals the empirical distribution induced by a history of length one $\boldsymbol{h} = (\boldsymbol{a})$. In other words, history is (in some sense) re-initialized and the players enter a new learning period. If $\tau(t)$ is negligible w.r.t. t, then the dynamic of the empirical probability distribution is not influenced by the inertia parameter. The constant $\tau(t)$ can therefore be interpreted as the inertia of the learning with respect to changes, making it more or less responsive to arrival and departures of players in the game being played. Other steps of CPRM are left unchanged. A detailed discussion of the inertia parameter is beyond the scope of this paper.

4 Numerical Results

In this section, we evaluate the performances of CPRM and its adaptive version. First, we consider a simple two-player matrix game to compare our solution to a well-known adaptive heuristic called regret-matching and to the no-regret learning procedure based on Blackwell's approachability both described in [6]. Then, we consider arrivals and departures of players in the game to observe how the adaptive version of CPRM performs. Finally, we conclude with a congestion game with larger sets of actions and player-specific cost functions [34]. For all experiments, the parameters in Table 2 were used.

4.1 Matrix Games

Constant Stage Game. We first consider the problem of learning a correlated equilibrium for the 3×2 matrix game shown in Table 3 admitting two mixed Nash equilibria $((1/12, 0, 11/12), (5/6, 1/6))$ and $((3/14, 11/14, 0), (5/6, 1/6))$ with respective utilities $(13/3, 73/12)$ and $(13/3, 82/7)$.

Table 2. Simulation Parameters.

	Value	Signification
β	0.05	Approximation factor
ε	0.01	Perturbation rate
γ	5	Experimentation rate
T	5×10^5	Number of iterations
C_0	1	Offset constant
τ	100	Inertia parameter

Table 3. Utility matrix of the two-player game.

	D	**E**
A	(2, 29)	(16, 7)
B	(4, 7)	(6, 13)
C	(4, 4)	(6, 6)

We consider the evolution in time ($0 \leq t \leq T$) of the empirical probability distribution $\{\boldsymbol{q}^t\}_t$, maximal regrets $\{(||R_i^t||)_{i\in\mathcal{N}}\}_t$ and players' moods $\{(m_i^t)_{i\in\mathcal{N}}\}_t$.

Figures 1a to 2c show the evolution in time of maximal regrets and the empirical distribution over action profiles induced by regret-matching, Blackwell's procedure for regret minimization and CPRM. In Fig. 1, we observe that regrets decrease below the threshold $\beta = 0.05$ for each learning procedure. For regret-matching, this occurs around 2×10^6 rounds, hence the larger simulation horizon compared to the two other algorithms. This empirically supports the convergence of the three algorithms to the set of β-correlated equilibria (shown in [6] for the Blackwell's procedure and regret-matching). However, if both players apply Blackwell's regret minimization or regret-matching procedures, the regret trajectories do not stabilize implying that the empirical distribution over action profiles does not converge pointwise.

Figure 1c shows that the regrets induced by CPRM stabilize below the target threshold (even if not converging to zero), which confirms that the empirical distribution approaches the set of correlated equilibrium distributions and may converge. Furthermore, Fig. 2c shows very stable trajectories for the probabilities of each action profile, thus supporting the hypothesis of convergence. This is not the case for the trajectories induced by the regret-matching procedure on Fig. 2a or Blackwell's regret minimization strategy on Fig. 2b which do not stabilize on the graphs and at even larger timescales (not shown).

Figure 3 shows the evolution in time of the pairs $\boldsymbol{u}(\boldsymbol{q}^t) = (u_1(\boldsymbol{q}^t), u_2(\boldsymbol{q}^t))$ where $u_i(\boldsymbol{q}^t)$ is the expected utility $u_i(\boldsymbol{q}^t) = \sum_{\boldsymbol{a}\in\mathcal{A}} \boldsymbol{q}^t(\boldsymbol{a}) u_i(\boldsymbol{a})$ for player i. For the sake of clarity of the figure, the plot displays one point every hundred points in the sample path, *i.e.* $\big(\boldsymbol{u}(\boldsymbol{q}^1), \boldsymbol{u}(\boldsymbol{q}^{101}), ...\big)$. The gray area represents the feasible pairs of utilities in the game with all possible probability distributions over action profiles. Figure 3a shows that starting from the initial action profile (A, D) with utilities $(2, 29)$, the trajectory stabilizes at a point in the vicinity of

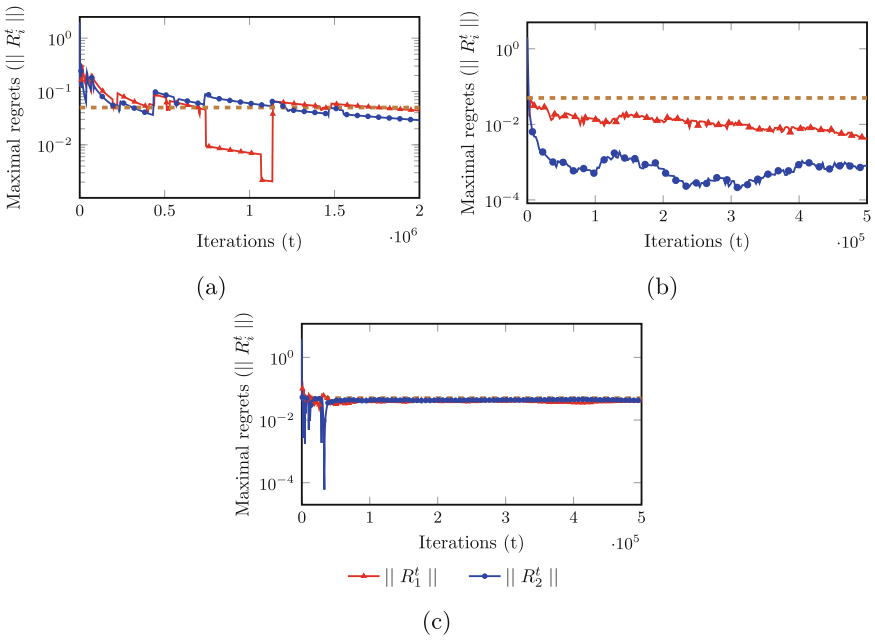


Fig. 1. Evolution of players regrets for the three algorithms: a) Regret-matching b) Blackwell-based regret minimization and c) CPRM algorithm.

the convex hull of the two mixed Nash equilibria. Figure 3b shows the trajectory approaching a correlated β-equilibrium at a smaller scale.

Figure 4 shows the evolution in time of the players' moods as a scatter-plot. In the first 4×10^4 iterations, the two players are mostly asynchronous, thus implementing a regret minimization strategy. Beyond 4×10^4, both players are synchronous, thus playing the action profile suggested by the device and drawn from $\boldsymbol{q}^t$. In this regime, asynchronous realizations (not visible on the graph for the given simulation horizon) typically come from the fact that players "explore" regardless of their regrets due to the perturbation ε^γ in the dynamic.

Fraction of Time Spent in a Correlated β-Equilibrium. In this section, we consider the impact of the perturbation on the long-run behaviour of CPRM for the previous two-player game. Let $\boldsymbol{q}^*(\varepsilon)$ be the correlated β-equilibrium experimentally reached with perturbation ε (last distribution in Figure 2c). In Fig. 5, we show the fraction of time the players are synchronous (thus following the suggestions of the device) and the empirical probability distribution is within a η-neighborhood (taking $\eta = 0.01$) to $\boldsymbol{q}^*(\varepsilon)$. The complementary proportion of time, either corresponds to a distribution at a distance greater than η from $\boldsymbol{q}^*(\varepsilon)$ or to the case where at least one player explores as a consequence of the perturbation.

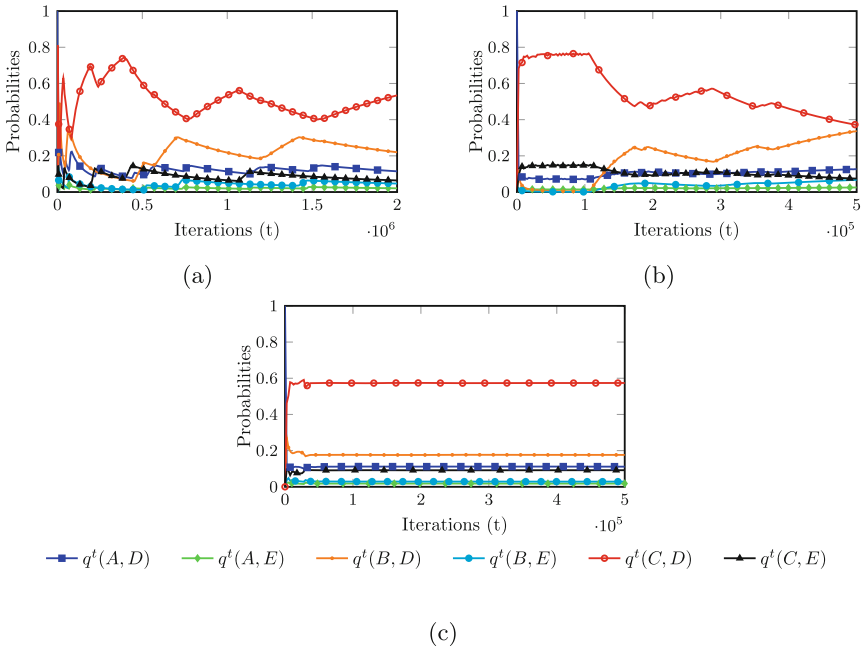


Fig. 2. Evolution of the empirical distribution over action profiles for the three algorithms: a) Regret-matching b) Blackwell's regret minimization and c) CPRM algorithm.

Figure 5 is an experimental evidence of the existence of a long-run regime such that both players follow the suggestions drawn from the correlated β-equilibrium $\boldsymbol{q}^*(\varepsilon)$. Furthermore, the plot shows that the smaller the perturbation ε the greater the proportion of time spent in the vicinity of an approximate correlated equilibrium. This is consistent with the type of convergence expected from the perturbed Markov process and the conjecture stating that in the low perturbations regime, with probability close to 1, players are synchronous and the empirical distribution converges to an approximate correlated equilibrium.

Arrivals and Departures of Players. Previously, we have assumed that the same stage game is played at every iteration. In this section, we consider the numerical performances of the adaptive version of CPRM (see Sect. 3.4) and observe how the previous convergence results may be impacted as new players join or leave the game and utility functions change.

Assume that players start playing the game in Table 3 expanding into a three player game before evolving later on into a two-player game and eventually reverting back to the same three-player case afterwards as shown in Table 4. In the three-player game, the first two players keep their original sets of actions while the third player chooses the matrix (X or Y). The first new player joins the game at $T_1 = 509583$ and leaves at $T_2 = 1019541$ while the second arrives

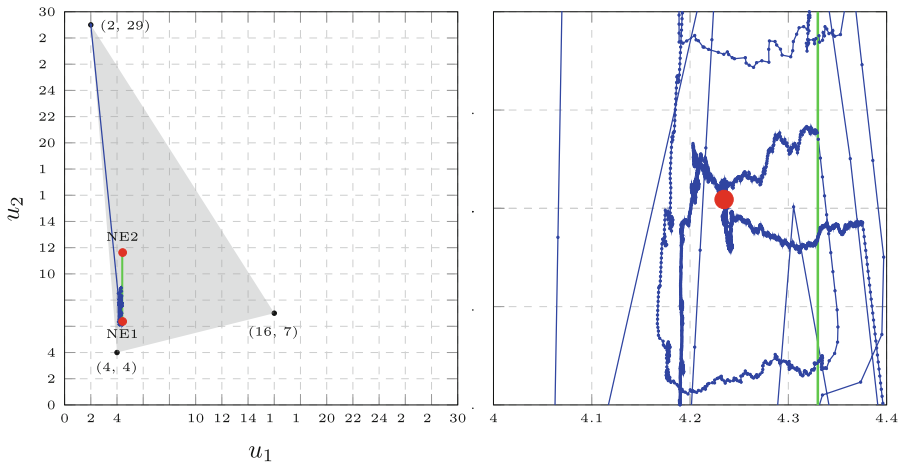


Fig. 3. (a) Trajectory (in blue) of the expected utilities starting from the initial action profile (A, D). Utilities at mixed Nash equilibria ($NE1$ and $NE2$) are shown in red. (b) Zoom-in showing the trajectory approaching a correlated β-equilibrium (red point) close to convex hull of Nash equilibria (in green). (Color figure online)

at $T_3 = 1529892$. Figure 6a shows the evolution in time of the maximal regrets of the players while Fig. 6b shows the evolution of probabilities for each profile. The arrival and departure of a player perturbs other players' regrets (red and blue curves correspond to the initial two players). It appears in Fig. 6a that for each game, the regrets are stabilized below the threshold β (dashed line) on the corresponding time interval. It can also be observed in Fig. 6b that for each game, the probability distribution over action profiles seems to converge on the corresponding time interval. These results show that CPRM may also be used in environments with arrivals and departures as long as each game is played for sufficiently long.

4.2 Congestion Game

As a final example of numerical experiment, we consider the problem of learning a correlated equilibrium in a congestion game [35] (a class of games particularly relevant w.r.t. network applications and resource allocation problems) with player-specific cost functions and larger action sets (the considered example has 108 action profiles) to test how relevant CPRM may be in this setting and its scalability with regards to the number of actions and action profiles. In a congestion game with player-specific cost functions, each player selects a feasible subset of resources and incurs a total cost defined as the sum of the costs of the chosen resources which depend on the player, each resource and the number of players using it.

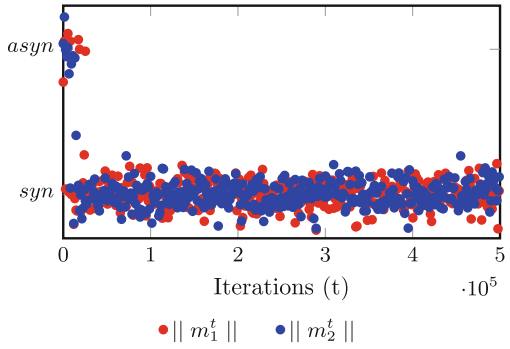


Fig. 4. Evolution of moods of the two players. An artificial scattering is used to facilitate data visualization.

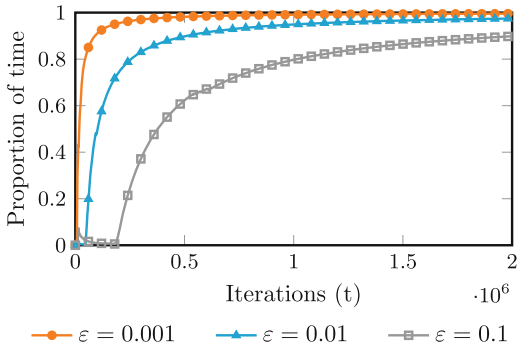


Fig. 5. Evolution in time of the proportion of time spent in the η-neighbourhood of the correlated β-equilibrium $\boldsymbol{q}^*(\varepsilon)$ for $\eta = 0.01$.

We consider the case where the resources are edges in a network and each player selects a subset of edges defining a path connecting a player-specific $(source, destination)$ pair of nodes. Formally, this game is defined by the following collection of objects,

- a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices and edges,
- a finite set $\mathcal{N} = \{1, \dots n\}$ of n players,
- for every player i, a source-destination pair $(s_i, t_i) \in \mathcal{V} \times \mathcal{V}$,
- for every player i, an action set $\mathcal{A}_i$ defined as the set of paths connecting source node s_i with target t_i,
- for every player i and every edge $e \in \mathcal{E}$, a non-decreasing delay function $d_i^e : \mathbb{N} \to \mathbb{R}$.

Table 4. Sequence of stage games considered in the dynamic case. The stage game does not necessarily evolve at every iteration.

	D	E
A	(2, 29)	(16, 7)
B	(4, 7)	(6, 13)
C	(4, 4)	(6, 6)

↓

	D	E
A	(2, 29, 2)	(16, 7, 8)
B	(4, 7, 2)	(6, 13, 0)
C	(4, 4, 1)	(6, 6, 5)

X

	D	E
A	(9, 4, 0)	(4, 1, 4)
B	(8, 0, 1)	(6, 7, 2)
C	(11, 9, 3)	(2, 0, 4)

Y

↓

	D	E
A	(2, 29)	(16, 7)
B	(4, 7)	(6, 13)
C	(4, 4)	(6, 6)

↓

	D	E
A	(2, 29, 2)	(16, 7, 8)
B	(4, 7, 2)	(6, 13, 0)
C	(4, 4, 1)	(6, 6, 5)

X

	D	E
A	(9, 4, 0)	(4, 1, 4)
B	(8, 0, 1)	(6, 7, 2)
C	(11, 9, 3)	(2, 0, 4)

Y

Let $f_e : \prod_{i \in \mathcal{N}} \mathcal{A}_i \rightarrow \{0, \ldots, n\}$ be the congestion function of edge e such that $f_e(\boldsymbol{a}) = |\{i \in \mathcal{N} : e \in a_i\}|$, *i.e.* the number of players using edge e. Given a strategy profile $\boldsymbol{a} \in \mathcal{A}$, player i has cost $c_i(\boldsymbol{a}) = \sum_{e \in a_i} d_i^e\,(f_e(\boldsymbol{a}))$.

Particularly, we consider the 4-player game with graph and pairs defined in Fig. 7, cost functions $d_i^e(x) = x$ for all $i \neq 2$, $d_i^e(x) = x^2$ for $i = 2$ and action sets,

- $\mathcal{A}_1 = \{''BCDEF'','' BDEF'','' BADEF''\}$
- $\mathcal{A}_2 = \{''BCDE'','' BDE'','' BADE''\}$
- $\mathcal{A}_3 = \{''DCB'','' DEFAB'','' DECB''\}$
- $\mathcal{A}_4 = \{''FDE'','' FADE'','' FABCDE'','' FABDE''\}$

As before, we first have an interest in a constant stage game and then allow for the stage game to change because of arrival and departure of players.

Figure 8b shows the evolution with time of the empirical distribution $\boldsymbol{q}^t$. Since we cannot show the 108 curves (one per action profile), we plot only the curves of the five action profiles with highest probabilities in the long-term. As in the previous example of the two-player matrix game, the curves support the conjectured convergence of the empirical distribution. This is to be put into perspective with the evolution of regrets shown in Fig. 8a, indicating that this long-run distribution is indeed a correlated β-equilibrium distribution. Then, in the long-run, the players follow a correlated equilibrium distribution of this game.

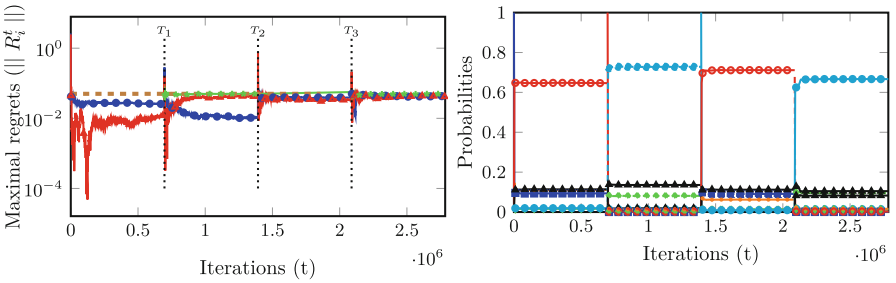
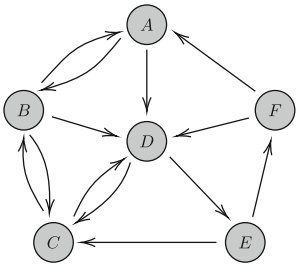


Fig. 6. Evolution of regrets (a) and the empirical distribution over action profiles (b) with arrival and departure of players (at times indicated with vertical dotted lines). The approximate equilibrium threshold β is marked with horizontal dashed line on the left figure.



Player	Source node	Destination node
1	B	F
2	B	E
3	D	B
4	F	E

Fig. 7. Network graph of the game (left) and source-destination nodes of each player (right).

To conclude, we assume that some players join or leave the congestion game. As commonly considered in network applications, we assume stochastic departures and arrivals following a Poisson process with rate $\lambda = 1/27236$. We show the results for a realization of this process such that a fifth player with pair (B, D) arrives at $T_1 = 54377$ and players 3, 5 and 4 leave at respectively $T_2 = 81434$, $T_3 = 108702$ and $T_4 = 135882$ as shown in Fig. 9. As expected, in the interval $0 \leq t \leq T_1$, the regret curves are similar to the case without arrivals and departures of Fig. 8a as the game being played in the considered time frame is the same.

It can be observed from Fig. 10 that in the third, fourth and last phase, the correlated β-equilibrium played is an approximate pure Nash equilibrium as only one profile is played with a probability close to 1. In all cases, regrets in Fig. 9 remain below the approximation threshold β.

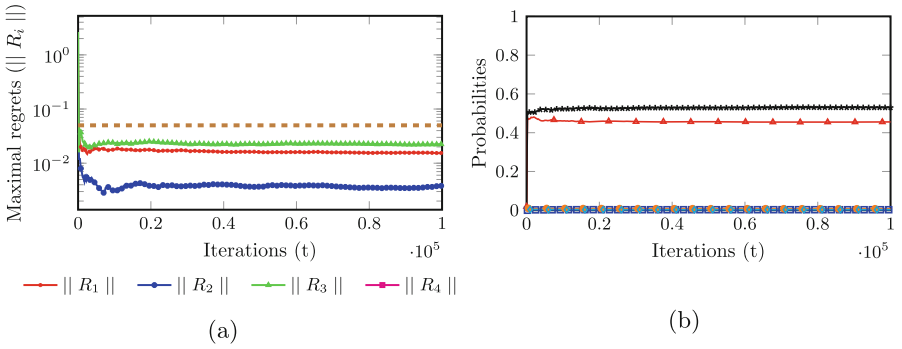


Fig. 8. (a) Evolution of the maximum regret for each player (curves of players 3 and 4 are not plotted because of low regrets and the logarithmic scale). (b) Evolution of the empirical distribution over the (five main) action profiles.

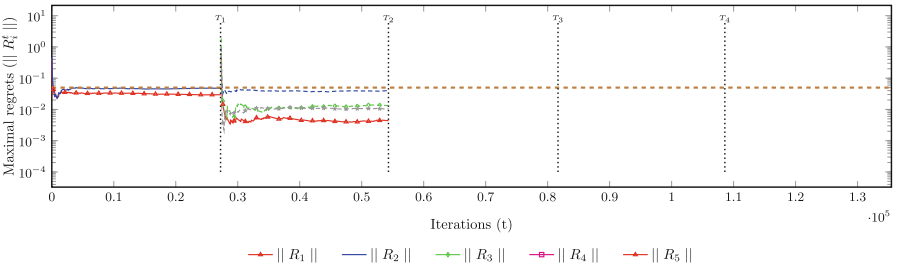


Fig. 9. Evolution of regrets with arrival and departure of players (at times indicated with vertical dotted lines). The approximate equilibrium threshold corresponds to the horizontal dashed line. Regrets with very small values are not displayed.

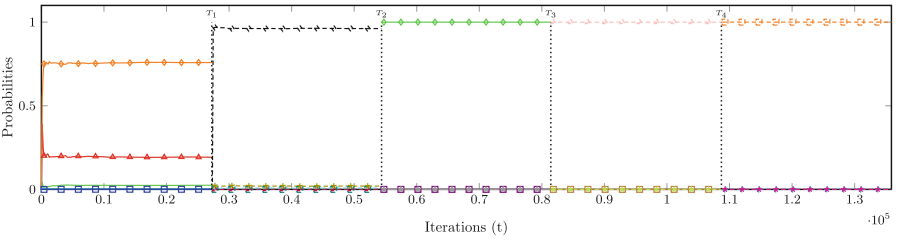


Fig. 10. Evolution of the empirical probability distribution. Only the five highest probabilities of action profiles are shown.

5 Conclusion

In this paper, we considered the problem of learning a correlated equilibrium in finite non-cooperative games with a particular focus on the open problem of convergence of the empirical probability distribution (over action profiles) induced by a learning rule to a correlated equilibrium distribution. We proposed a new

learning rule, called CPRM, combining regret minimization to approach the set of correlated equilibria and a simple device drawing samples from the empirical distribution. Numerical experiments support the conjecture that approximate correlated equilibrium distributions (the approximation factor being a parameter of the dynamic) with all players following the devices' suggestions are the only stochastically stable states and that the empirical distribution converges pointwise. Additional experiments show that CPRM can be adapted to comply with a time-varying game (*e.g.* arrivals and departures of players, changing utilities). In future research, we plan to prove the conjecture to confirm the results obtained in this paper.

A Appendix: Numerical Implementation of CPRM

Algorithm 1: Correlated Perturbed Regret Minimization (CPRM)

Let $G = (\mathcal{N}, (\mathcal{A}_i)_{i\in\mathcal{N}}, (u_i)_{i\in\mathcal{N}})$, $\varepsilon > 0$, $\beta > 0$, $C_0 > 0$, $\gamma \geq n2^{n-1}(\beta + C_0)$
Initialize moods $(m_i^1)_{i\in\mathcal{N}}$, history $h^1 = (\boldsymbol{a}^1)$, empirical distribution $\boldsymbol{q}^1$ and $(D_i^1)_{i\in\mathcal{N}}$
for *t=1, 2,...* **do**
  Draw an action profile $\boldsymbol{b}^{t+1} = (b_1, ..., b_n)$ from $\boldsymbol{q}^t$
  for $i \in \mathcal{N}$ **do**
    /* Play according to player i's mood & update mood */
    Draw uniformly in $[0, 1]$: $var \leftarrow Uniform(0, 1)$
    if $m_i^t = asyn$ **then**
      if $\varepsilon^{||D_i^t||} > var$ **then**
        $m_i^{t+1} \leftarrow$ syn
      end
      Play a realization c_i of the mixed strategy in Eq. (4)
      $a_i^{t+1} \leftarrow c_i$
    else
      if $\varepsilon^{\gamma} > var$ **then**
        $m_i^{t+1} \leftarrow$ asyn
      else
        if $||D_i^t|| > \beta$ **then**
          $m_i^{t+1} \leftarrow asyn$
        end
      end
      $a_i^{t+1} \leftarrow b_i$
    end
  end
  /* Update the empirical distribution */
  $\boldsymbol{q}^{t+1}(\boldsymbol{a}) \leftarrow \frac{t}{t+1}\boldsymbol{q}^t(\boldsymbol{a}) + \frac{1}{t+1}\mathbb{1}_{\{\boldsymbol{a}^{t+1}=\boldsymbol{a}\}}, \forall \boldsymbol{a} \in \mathcal{A}$
  /* Update the vector of the average utility differences */
  $\forall i \in \mathcal{N}, \forall j \in \mathcal{A}_i, \forall k \in \mathcal{A}_i, D_i^{t+1}(j,k) \leftarrow \sum_{\substack{\boldsymbol{a}\in\mathcal{A} \\ a_i=j}} \boldsymbol{q}^{t+1}\left[u_i\left(k, \boldsymbol{a}_{-i}\right) - u_i\left(a_i, \boldsymbol{a}_{-i}\right)\right]$
end

References

1. Aumann, R.J.: Subjectivity and correlation in randomized strategies. J. Math. Econ. **1**(1), 67–96 (1974)
2. Aumann, R.J.: Correlated equilibrium as an expression of bayesian rationality. Econometrica: J. Econometric Soc. **55**, 1–18 (1987)
3. Forges, F.: Correlated equilibria and communication in games. Complex Soc. Behav. Syst.: Game Theory Agent-Based Models, 107–118 (2020)
4. Myerson, R.B.: Game Theory: Analysis of Conflict. Harvard University Press, Cambridge (1997)
5. Hart, S.: Adaptive heuristics. Econometrica **73**(5), 1401–1430 (2005)
6. Hart, S., Mas-Colell, A.: A simple adaptive procedure leading to correlated equilibrium. Econometrica **68**(5), 1127–1150 (2000)
7. Hart, S., Mas-Colell, A.: A reinforcement procedure leading to correlated equilibrium. Econ. Essays, 181–200 (2001)
8. Foster, D.P., Vohra, R.V.: Calibrated learning and correlated equilibrium. Games Econom. Behav. **21**(1–2), 40–55 (1997)
9. Fudenberg, D., Levine, D.K.: The Theory of Learning in Games. The MIT Press, vol. 1, no. 0262061945 (1998)
10. Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press, USA (2006)
11. Foster, D., Young, H.P.: Regret testing: learning to play Nash equilibrium without knowing you have an opponent. Theor. Econ. **1**(3), 341–367 (2006)
12. Young, H.P.: Learning by trial and error. Games Econom. Behav. **65**(2), 626–643 (2009)
13. Boussaton, O., Cohen, J., Tomasik, J., Barth, D.: On the distributed learning of Nash equilibria with minimal information. In: 2012 6th International Conference on Network Games, Control and Optimization (NetGCooP). IEEE, pp. 30–37 (2012)
14. Frihauf, P., Krstic, M., Basar, T.: Nash equilibrium seeking in noncooperative games. IEEE Trans. Autom. Control **57**(5), 1192–1207 (2011)
15. Germano, F., Lugosi, G.: Global Nash convergence of foster and young's regret testing. Games Econom. Behav. **60**(1), 135–154 (2007)
16. Hart, S., Mas-Colell, A.: Uncoupled dynamics do not lead to Nash equilibrium. Am. Econ. Rev. **93**(5), 1830–1836 (2003)
17. Marden, J.R., Young, H.P., Arslan, G., Shamma, J.S.: Payoff-based dynamics for multiplayer weakly acyclic games. SIAM J. Control. Optim. **48**(1), 373–396 (2009)
18. Pradelski, B.S., Young, H.P.: Learning efficient Nash equilibria in distributed systems. Games Econom. Behav. **75**(2), 882–897 (2012)
19. Ariel, I., Babichenko, Y.: Average Testing and the Efficient Boundary. Center for the study of Rationality (2011)
20. Marden, J.R., Young, H.P., Pao, L.Y.: Achieving pareto optimality through distributed learning. SIAM J. Control. Optim. **52**(5), 2753–2770 (2014)
21. Blackwell, D., et al.: An analog of the minimax theorem for vector payoffs. Pac. J. Math. **6**(1), 1–8 (1956)
22. Perchet, V.: Université Paris-Diderot, Laboratoire de Probabilités et Modèles Aléatoires, UMR 7599, 8 place FM/13, Paris, "Approachability, regret and calibration: Implications and equivalences," J. Dyn. Games, vol. 1, no. 2, pp. 181–254 (2014)
23. Greenwald, A., Hall, K., Serrano, R.: Correlated q-learning. In: ICML, vol. 3, pp. 242–249 (2003)

24. Borowski, H.P., Marden, J.R., Shamma, J.S.: Learning to play efficient coarse correlated equilibria. Dyn. Games Appl. **9**(1), 24–46 (2019)
25. Marden, J.R.: Selecting efficient correlated equilibria through distributed learning. Games Econom. Behav. **106**, 114–133 (2017)
26. Celli, A., Marchesi, A., Farina, G., Gatti, N.: No-regret learning dynamics for extensive-form correlated and coarse correlated equilibria. CoRR, vol. abs/2004.00603 (2020)
27. Farina, G., Celli, A., Marchesi, A., Gatti, N.: Simple uncoupled no-regret learning dynamics for extensive-form correlated equilibrium. CoRR, vol. abs/2104.01520 (2021)
28. Jin, H., Guo, H., Su, L., Nahrstedt, K., Wang, X.: Dynamic task pricing in multi-requester mobile crowd sensing with markov correlated equilibrium. In: IEEE INFOCOM 2019-IEEE Conference on Computer Communications. IEEE, pp. 1063–1071 (2019)
29. Hu, Q., Nigam, Y., Wang, Z., Wang, Y., Xiao, Y.: A correlated equilibrium based transaction pricing mechanism in blockchain. In: 2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). IEEE, pp. 1–7 (2020)
30. Cigler, L., Faltings, B.: Reaching correlated equilibria through multi-agent learning. In: 10th Conference on Autonomous Agents and Multiagent Systems AAMAS, no. CONF (2011)
31. Young, H.P.: The evolution of conventions. Econometrica: J. Econometric Soc. pp. 57–84 (1993)
32. Dolan, R.J.: Incentive mechanisms for priority queuing problems. Bell J. Econ. 421–436 (1978)
33. Lykouris, T., Syrgkanis, V., Tardos, É.: Learning and efficiency in games with dynamic population. In: Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete algorithms. SIAM, pp. 120–129 (2016)
34. Ackermann, H., Röglin, H., Vöcking, B.: Pure Nash equilibria in player-specific and weighted congestion games. Theoretical Comput. Sci. **410**(17), 1552–1563 (2009)
35. Rosenthal, R.W.: A class of games possessing pure-strategy Nash equilibria. Internat. J. Game Theory **2**(1), 65–67 (1973)