Learning a Correlated Equilibrium with Perturbed Regret Minimization

Omar Boufous, Rachid El-Azouzi, Mikaël Touati, Eitan Altman, Mustapha Bouhtou

Learning a Correlated Equilibrium with Perturbed Regret Minimization

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

[1] Orange Innovation, 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 adaptive heuristic, called Correlated Perturbed Regret Minimization (CPRM) for this purpose. CPRM combines regret minimization to approach the set of correlated equilibria and a simple device suggesting actions to the players 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 CPRM is adaptive to changes in the game such as departures or arrivals of players.

Keywords: Game theory · Correlated equilibrium · Online 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 defined as an equilibrium of a game extended with a 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 each player its 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.

We consider the problem of learning a correlated equilibrium of a non-cooperative game with simple rules of behaviour known as *adaptive heuristics* (*e.g.* fictitious play, regret minimization procedures) [5]. These rules may not be rational in the sense that they may not choose the action maximizing the

player's expected utility but may lead to rational outcomes *in the long run*. 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, no adaptive heuristic implies a pointwise converge (more generally, a stabilization close) to a correlated equilibrium distribution.

The main objective of this paper is to address the latter issue by introducing a new adaptive heuristic called Correlated Perturbed Regret Minimization (CPRM) leading to such results. 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 suggestion.

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 & Mas-Colell use Blackwell's approachability theory [21] to minimize players' regrets and 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 guarantee pointwise convergence of the trajectories to equilibrium points. In [23], Greenwald et al. presented *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. This requires from every player a perfect knowledge of the game being played (payoff and actions of all players involved). In [24], Borowski et al. proposed an uncoupled learning rule with a public signal such that, in the long-run and as the perturbation term tends to zero, the process spends most of the time at the efficient coarse correlated equilibrium if feasible. It is not shown that the players' joint strategy is an efficient coarse correlated equilibrium at any time. See [5,6,8,25] for other works on learning coarse correlated equilibria. Recent contributions consider learning correlated equilibria in more general dynamic games such as [26] [27] for 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. The proposed approach seems to be limited to the considered system.

1.2 Outline

In Section 2, we define the model and provide the necessary preliminaries such as the relationship between correlated equilibria and the notion of regret used in CPRM. In Section 3 we show the dynamic of the learning algorithm. In Section 4, we evaluate and discuss numerical performances of our solution and Section 5 concludes this work 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 $\boldsymbol{x} \geq \boldsymbol{y}$ means $x_i \geq y_i$ for every i. We use calligraphic capital letters for sets and $|\mathcal{S}|$ is the cardinality of $\mathcal{S}$. The indicator function of $\mathcal{A}$ is denoted $\mathbb{1}_{\mathcal{A}}$. Table 1 summarizes other notations used in the paper.

Table 1: Table of notations.

Symbol	Meaning
$\mathbb{R}$ (resp. $\mathbb{Q}$)	Set of real (resp. rational) numbers
$\mathcal{N}$	Set of players
$\mathcal{A}$	Joint action space
$\Delta(\mathcal{A})$	Set of all probability distributions over the set of action profiles $\mathcal{A}$
$\boldsymbol{q}$	Probability distribution over action profiles in $\mathcal{A}$
$\boldsymbol{h}^t$	History of the game up to time t
$\mathcal{A}_i$	Player i's action space
$\Delta(\mathcal{A}_i)$	Set of all probability distributions over $\mathcal{A}_i$
$\boldsymbol{p}_i$	Player i's mixed strategy
$\boldsymbol{p}_i(a_i)$	Player i's probability of playing action a_i
$u_i(\cdot)$	Player i's payoff function
$d_i^t(j,k)$	Player i's average difference between payoffs of strategies j and k at t
$r_i^t(j,k)$	Player i's regret of for not choosing action k when action j is chosen at t
ε	Perturbation parameter
γ	A parameter defining the order of the perturbation ε
β	A parameter that defines the approximate correlated β-equilibrium

2.2 Model

We consider a finite non-cooperative game $G = (\mathcal{N}, (\mathcal{A}_i)_{i \in \mathcal{N}}, (u_i)_{i \in \mathcal{N}})$ such that player $i \in \mathcal{N}$ has finite set of actions $\mathcal{A}_i$. Let $a_i \in \mathcal{A}_i$ be a pure action for player i and $\boldsymbol{p_i} \in \Delta(\mathcal{A}_i)$ be a mixed strategy for player i. The set $\mathcal{A} = \prod_i \mathcal{A}_i$ is the set

of action profiles and $\mathcal{A}_{-i} = \prod_{j \neq i} \mathcal{A}_j$ is the set of action profiles for players in $\mathcal{N} \backslash \{i\}$. Players i's utility function is $u_i : \mathcal{A} \rightarrow \mathbb{R}$ such that $u_i(\boldsymbol{a}) = u(a_i, \boldsymbol{a_{-i}})$ is i's utility for action profile $\boldsymbol{a} = (a_i, \boldsymbol{a_{-i}}) \in \mathcal{A}$. Player i's utility for the mixed strategy profile $\boldsymbol{p} = (\boldsymbol{p_1}, \ldots, \boldsymbol{p_n})$ is $u_i(\boldsymbol{p}_i, \boldsymbol{p}_{-i}) = \sum_{a \in \mathcal{A}} \prod_{i \in \mathcal{N}} p_i(a_i) u_i(a_i, \boldsymbol{a_{-i}})$.

2.3 Correlated equilibria and their approximations

An approximate correlated equilibrium is a probability distribution over action profiles such that the expected utility of every player is approximately maximal. Formally, we have the following definition.

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

$$\forall i \in \mathcal{N}, \forall a_i, a_i' \in \mathcal{A}_i \quad \sum_{\boldsymbol{a_{-i}} \in \mathcal{A}_{-i}} \boldsymbol{q}(\boldsymbol{a})[u_i(a_i', \boldsymbol{a_{-i}}) - u_i(a_i, \boldsymbol{a_{-i}})] \leq \beta. \qquad (1)$$

The case $\beta = 0$ corresponds to the usual case of (an exact) correlated equilibrium [2].

As an example, consider the traffic game and involving two vehicles (players) at an intersection. At each stage, each player can either cross (action "*Go*") or wait (action "*Wait*"). When both players cross simultaneously, damage is caused by a collision, both players are penalized and incur a negative utility of $(-1, -1)$. In this game, there are two pure strategy Nash equilibria $(Wait, Go)$ and $(Go, Wait)$ and one Nash equilibrium in mixed strategies $\big((\frac{1}{2} \cdot Wait, \frac{1}{2} \cdot Go), (\frac{1}{2} \cdot Wait, \frac{1}{2} \cdot Go)\big)$. Furthermore, the probability distribution $\mathbb{P}(Go, Wait) = \mathbb{P}(Wait, Go) = 1/2$, $\mathbb{P}(Go, Go) = \mathbb{P}(Wait, Wait) = 0$ is a correlated equilibrium.

Table 2: Payoff matrix of the traffic intersection game.

	Wait	**Go**
Wait	(0, 0)	(0, 1)
Go	(1, 0)	(-1, -1)

The pairs of payoffs associated with the Nash equilibria are $(0, 1), (1, 0)$ and $(0, 0)$ respectively. Observe that pure Nash equilibria result in unfair utility vectors. Similarly, the mixed Nash equilibrium is fair but inefficient. Now assume that the game is repeated and that both players can receive recommendations on the action to play before each stage. Particularly, consider a coin toss such that:

- if player 1 observes "Head", she plays *Wait* else she plays *Go*.
- if player 2 observes, "Head" she plays *Go*, else she plays *Wait*.

In this configuration, in the long-run, the players play the profile (*Wait*, *Go*) and (*Go*, *Wait*) 50% of the time each, thus resulting in an average payoff of $(\frac{1}{2}, \frac{1}{2})$

which is more satisfactory than the three Nash equilibria and corresponds to a correlated equilibrium distribution. The idea of recommendations (*i.e.* players making decisions using private observations of a single random outcome) as well as the notion of regret that we introduce in the following paragraphs are both used in the solution we propose.

2.4 Regret minimization and correlated equilibria

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

$$\forall \boldsymbol{a} \in \mathcal{A}, \quad \boldsymbol{q}^t(\boldsymbol{a}) = \frac{1}{t}|\{\tau \leq t : \boldsymbol{a}^\tau = \boldsymbol{a}\}| \tag{2}$$

Following [6], define the matrix $D_i^t = (d_i^t(j,k))_{(j,k)\in\mathcal{A}_i\times\mathcal{A}_i}$, such that $d_i^t(j,k)$ is the average payoff difference for player i when playing action k every time j was played.

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

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

and the matrix $R_i^t = (r_i^t(j,k))_{j,k\in\mathcal{A}_i\times\mathcal{A}_i}$ such that $r_i^t(j,k)$ is the regret of player i for the action swap $j,k \in \mathcal{A}_i$.

$$\forall i \in \mathcal{N}, \forall j,k \in \mathcal{A}_i, \;\; r_i^t(j,k) = \max\{0, d_i^t(j,k)\} \tag{4}$$

The regret $r_i^t(j,k)$ is the average gain of utility player i could have obtained if he had played k instead of j every time he played j in the past t sequence of moves.

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, we deduce from the definition of $d_i^t(j,k)$ in Eq. (3b) that $r_i^t(j,k) \leq \beta$ at a given time t if and only if the empirical distribution $\boldsymbol{q}^t$ is a correlated β-equilibrium.

The Proposition 1 below shows the equivalence between low regrets and approximate correlated equilibria,

Proposition 1 ([6]) *Let $(a^t)_{t=1,2,\dots}$ be a sequence of plays (i.e., $a^t \in \mathcal{A}$ 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.*

In [6], Hart et al. proposed a regret minimizing strategy using the approachability of the negative orthant $\mathbb{R}_-^{|\mathcal{A}_i|\times|\mathcal{A}_i|}$ by the regret matrix $(r_i(j,k))_{(j,k)\in\mathcal{A}_i\times\mathcal{A}_i}$.

This learning procedure is an application of Blackwell's approachability [21] enabling player i's regret to be minimized *i.e.*

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

where $d(\cdot,\cdot)$ is the Euclidean norm. The regret minimization strategy is the following. At any given time t,

- If $R_i^t \notin \mathbb{R}_-^{|\mathcal{A}_i|\times|\mathcal{A}_i|}$, player i chooses an action according to the mixed strategy p_i^t satisfying Eq. (6).

$$\forall j,k \in \mathcal{A}_i, \quad \sum_{j\in\mathcal{A}_i} p_i^t(j) r_i^t(j,k) = p_i^t(k) \sum_{j\in\mathcal{A}_i} r_i^t(k,j). \tag{6}$$

- Else, i plays randomly (*i.e.* an action drawn from any distribution)

In the next section, we propose a perturbed variant of this process, in which players synchronize as they approach the set of correlated equilibria to stabilize the realized sequence of empirical distribution $\{\boldsymbol{q}^t\}_{t=1,2,\ldots}$.

3 Learning algorithm

3.1 Rationale

In CPRM, at time t, each player $i \in \mathcal{N}$ is characterized by an individual state x_i^t with two components. The first component is called "mood" [12] and the second is the empirical distribution $\boldsymbol{q}^t$. In one mood, she implements a regret minimizing 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 (thus contributing in stabilizing the sequence $\{\boldsymbol{q}^t\}_{t=1,2,\ldots}$).

3.2 Device

We assume a device drawing at time t an action profile $\boldsymbol{a}^t = (a_1^t, ..., a_n^t)$ with probability distribution $\boldsymbol{q}^t$ as defined in Eq. (2) and recommending to player i the component a_i^t. By assumption, the device must know $\boldsymbol{q}^t$ at t (not necessarily storing the history of play h^t) and must be able to transmit to every player her component in $\boldsymbol{a}^t$. We typically assume that it does not know the players' utility functions.

3.3 Players' states and strategies

At time t, player $i \in \mathcal{N}$ is characterized by a pair $x_i^t = (m_i^t, \boldsymbol{q}^t)$ in $\mathcal{X}_i = \{asyn, syn\} \times \Delta(\mathcal{A})$. The first component m_i^t describes the player's mood at time

t which can be synchronous (syn) of asynchronous ($asyn$). The second component is the empirical distribution of play $\boldsymbol{q}^t$ given by Eq. (2). By definition, the second component is the same for all players.

If $m_i^t = syn$, then i plays $s_i^t = a_i^t$ (sent by the device), else ($m_i^t = asyn$) player i plays a realization s_i^t of the random variable with probability distribution ξ_i^t such that ξ_i^t is the stationary distribution of matrix of regrets satisfying Eq. (6).

3.4 Players' states dynamic

Assume that at time t player i is in state $x_i^t = (m_i^t, \boldsymbol{q}^t)$ and that the profile $\boldsymbol{a}^t$ is played (some players implementing the regret minimization strategy, others following the device's suggestions depending on their moods). The state x_i^{t+1} of player i at time $t+1$ is the realization of a random variable X_i^{t+1} such that,

$$\mathbb{P}(X_i^{t+1} = (m, \boldsymbol{q}) | X_i^t = x_i^t, \boldsymbol{a}^t) = \mathbb{P}(m_i^{t+1} = m | X_i^t = x_i^t, \boldsymbol{a}^t) \times \mathbb{P}(\boldsymbol{q}^{t+1} = \boldsymbol{q} | X_i^t = x_i^t, \boldsymbol{a}^t) \quad (7)$$

where

$$\mathbb{P}(m_i^{t+1} = m | X_i^t = x_i^t, \boldsymbol{a}^t) = \begin{cases} \varepsilon^{||d_i(\boldsymbol{q}^t)||} & \text{if } m{=}syn, \text{ and } m_i^t{=}asyn \\ 1 - \varepsilon^{||d_i(\boldsymbol{q}^t)||} & \text{if } m{=}asyn, \text{ and } m_i^t{=}asyn \\ \varepsilon^{\gamma} & \text{if } m{=}asyn, \text{ and } m_i^t{=}syn \\ 1 - \varepsilon^{\gamma} & \text{if } m{=}syn, m_i^t{=}syn, \text{ and } ||d_i(\boldsymbol{q}^t)|| \leq \beta \\ 1 - \varepsilon^{\gamma} & \text{if } m{=}asyn, m_i^t{=}syn, \text{ and } ||d_i(\boldsymbol{q}^t)|| > \beta \end{cases} \quad (8)$$

where $d_i : \Delta(\mathcal{A}) \to \mathbb{R}^{|\mathcal{A}_i| \times |\mathcal{A}_i|}$ is the function defined in Eq. (3) and

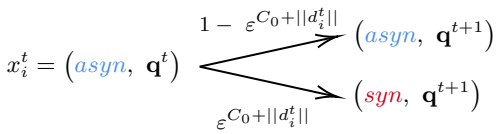
$$\mathbb{P}(\boldsymbol{q}^t = \boldsymbol{q} | X_i^t = x_i^t, \boldsymbol{a}^t) = \mathbb{1}_{\{\boldsymbol{q} = \boldsymbol{f}(t, \boldsymbol{q}^t, \boldsymbol{a}^t)\}} \quad (9)$$

where $f : \Delta(\mathcal{A}) \to \Delta(\mathcal{A})$ is the function updating the empirical distribution such that,

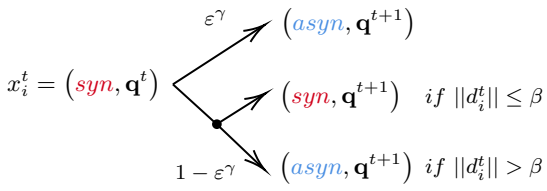
$$\forall \boldsymbol{x} \in \mathcal{A}, \quad \boldsymbol{f}(t, \boldsymbol{q}^t, \boldsymbol{a}^t)(\boldsymbol{x}) = \frac{t}{t+1} \times \boldsymbol{q}^t(\boldsymbol{x}) + \frac{1}{t+1} \mathbb{1}_{\{\boldsymbol{x} = \boldsymbol{a}^t\}} \quad (10)$$

We show below graphical representations of the mood dynamic.

– If player i is in 'mood' *asyn i.e.* $m_i^t = asyn$, she moves to mood syn with probability $\varepsilon^{C_0 + ||d_i||}$,



– If player i is in 'mood' *syn i.e.* $m_i^t = syn$, she is faced with three possible transitions. The first possibility corresponds to the case where the player experiments with probability ε^γ, without taking into account her regrets.



The two other cases deal with the situation where player i does not experiment and monitors her regrets to decide her next move. As long as the regrets are below β, player i keeps following the recommendation of the device, otherwise, the player's mood reverts back to *asyn*.

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

It can be shown that the stochastic process $\{X^t = (X_1^t, \ldots, X_n^t)\}_t$ (modeling the state dynamic and where X_i^t is the random state of player i at t) induced by CPRM is a perturbed non-homogeneous Markov process on a countable state space. Furthermore, we conjecture that if this process admits an asymptotic stationary distribution and $\gamma > n2^{n-1}(C_0 + \beta)$, then $\mathcal{S} = \{x \in \mathcal{X} : \forall i \in \mathcal{N}, m_i = syn, ||d_i(q)|| \leq \beta\}$ is the only stochastically stable set of states [31] and the sequence $\{\mathbf{q}^t\}_{t=1,2,\ldots}$ converges implying that in the long-run the players follow the suggestions of the devices drawing action profiles from a correlated β-equilibrium distribution. A detailed analysis of this process is beyond the scope of the paper.

3.5 Adaptive CPRM

The previous learning dynamic assumes a steady environment, with static game parameters, which is rarely the case in online settings, where a population of players may be dynamically changing. This characteristic is of paramount importance for many applications such as packet routing or auction systems, but it is has not yet been thoroughly investigated in the literature. For instance, the work in [32] examines a repeated n-player game in which players adapt to the environment. More specifically, each player independently leaves with a certain probability, but is immediately replaced by an arbitrary new player such that the total number of players n remains identical. In our work, the total number of players may evolve with time. We propose an adaptation of CPRM to a dynamic case to deal with arrival and departure of players. In simulations, we demonstrate the efficiency of our method on examples.

By assumption, the mediator is aware of the players involved in the game. We

describe how the probability distribution used by the mediator for recommending actions is adjusted to the play.
Let T_0 be the arrival time player $n+1$. Let b_{n+1} be an arbitrary action in $\mathcal{A}_{n+1}$. We construct the probability distribution used from T_0+1 onwards for all profile $(a_1, ..., a_n)$:

$$q^{T_0+1}(a_1, ..., a_{n+1}) = \begin{cases} q^{T_0}(a_1, ..., a_n), & \text{if } a_{n+1} = b_{n+1} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Similarly, when a player i leaves at time T_1, the probability distribution is projected from $\Delta(\prod_{i=1}^{n+1} \mathcal{A}_i)$ onto $\Delta(\prod_{i=1}^{n} \mathcal{A}_i)$ as follows:

$$q^{T_1+1}(a_1, ..., a_n) = \sum_{a_i \in \mathcal{A}_i} q^{T_1}(a_1, ..., a_{n+1}) \tag{12}$$

After each player departure or arrival at a given time T, the scaling factor t used for updating the empirical distribution is re-initialized in the following way:

$$\boldsymbol{q}^{t+1}(a) = \begin{cases} \frac{t-\tau}{t-\tau+1}\boldsymbol{q}^t(a) + \frac{1}{t-\tau+1} & \text{if } \boldsymbol{a}^t = \boldsymbol{a} \\ \frac{t-\tau}{t-\tau+1}\boldsymbol{q}^t(a) & \text{otherwise} \end{cases} \tag{13}$$

where $\tau \geq 0$ is a design parameter that is chosen depending on the desired responsiveness of the games to changes in player population. For instance, taking $\tau = T$ when a player arrives or leaves at time T essentially means that the players start a new learning phase for the new game, which is independent from the previous learning period. On the other hand, a large value of τ signifies that the updates of the probability distribution are small, thus making the game less prone to rapid changes, and resulting in a slower learning and convergence. The constant τ 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. The other parts of the algorithm remain unchanged.

4 Numerical results

In this section, we evaluate the performances of CPRM. First, we consider a simple two-players matrix game to compare our solution to other adaptive heuristics such as a regret-minimization based on Blackwell's strategy [6] and Regret Matching [6]. Then, we consider arrivals and departures of players in the game to observe how CPRM may adapt and conclude this performance evaluation on a congestion game with larger sets of actions and player-specific cost functions. For all experiments, the parameters in Table 3 were used.

4.1 Matrix games

Static setting We first consider the problem of learning a correlated equilibrium for the 3×2 matrix game shown in Table (4) admitting two mixed

Table 3: Numerical experiments parameters.

	Value	Signification
β	0.05	Approximation factor
ε	0.01	Perturbation rate
γ	5	Perturbation order
T	5×10^5	Number of iterations
C_0	1	Offset

Table 4: Payoff matrix of the 2-player game.

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

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 $(13/3, 82/7)$.

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

Fig. 1a to 2c show the evolution in time of maximal regrets and the empirical distribution over action profiles induced by Regret Matching, Blackwell's regret minimization and CPRM. In Figs. 1, we observe that regrets decrease below the threshold $\beta = 0.05$ for each dynamic. This confirms the convergence of three algorithms to the set of β-correlated equilibria (as expected from [6]). However, if both players apply the Blackwell regret minimization strategy or the Regret-Matching procedure, the regret trajectories do not stabilize implying that the empirical distribution over action profiles does not converge to a correlated equilibrium.

Fig. 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 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).

We provide an example of a sample path generated by CPRM in terms of expected utility in Fig. 3. The plot shows the evolution in time of the pairs $(\bar{u}_1(\boldsymbol{q}^t), \bar{u}_2(\boldsymbol{q}^t))$ where $\bar{u}_i(\boldsymbol{q}^t)$ is the expected utility $\bar{u}_i(\boldsymbol{q}^t) = \sum_{\boldsymbol{a} \in \mathcal{A}} \boldsymbol{q}^t(\boldsymbol{a}) u_i(\boldsymbol{a})$ for player i. The gray area represents the feasible pairs of utilities in the game. Starting from the initial action profile (A, D) with utilities $(2, 29)$, the trajectory stabilizes at a point in the vicinity of the convex hull of the two mixed Nash equilibria depicted by the red segment.

Fig. 4 shows the evolution in time of the players' moods as a scatter-plot. In the first thousands time steps, the two players are mostly asynchronous (value

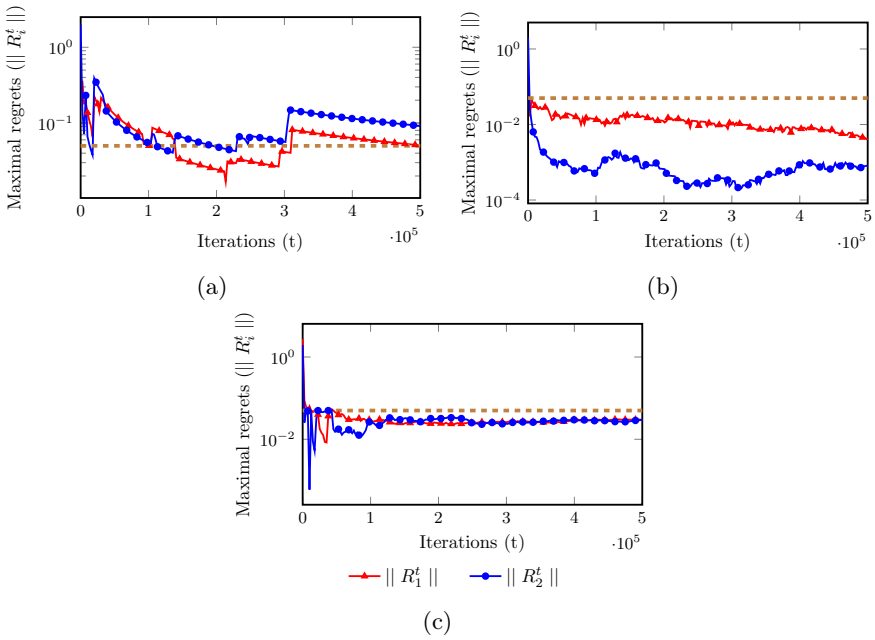


Fig. 1: Evolution of players regrets for the three algorithms: a) Regret Matching b) Blackwell-based regret minimization and c) CPRM algorithm.

"1" on the plot), thus implementing a regret minimizing strategy. Around 4.10^4 time steps, both players are synchronous, thus playing the action profile suggested by the device and drawn from $\boldsymbol{q}^t$. In this regime, asynchronous realizations 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-players game. Let $q^*(\varepsilon)$ be the correlated β-equilibrium experimentally reached with perturbation ε (last distribution in Fig.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 implemented by the latter is within a η-neighborhood (taking $\eta = 0.01$) to $q^*(\varepsilon)$. The complementary proportion of time, corresponds either to a distribution at a distance greater than η from $q^*(\varepsilon)$ or to the case where at least one player explores as a consequence of the perturbation.

Fig. 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 $q^*(\varepsilon)$. Furthermore, the plot shows that the smaller decreasing the perturbation ε implies increasing the proportion of time. This is consistent with the type of

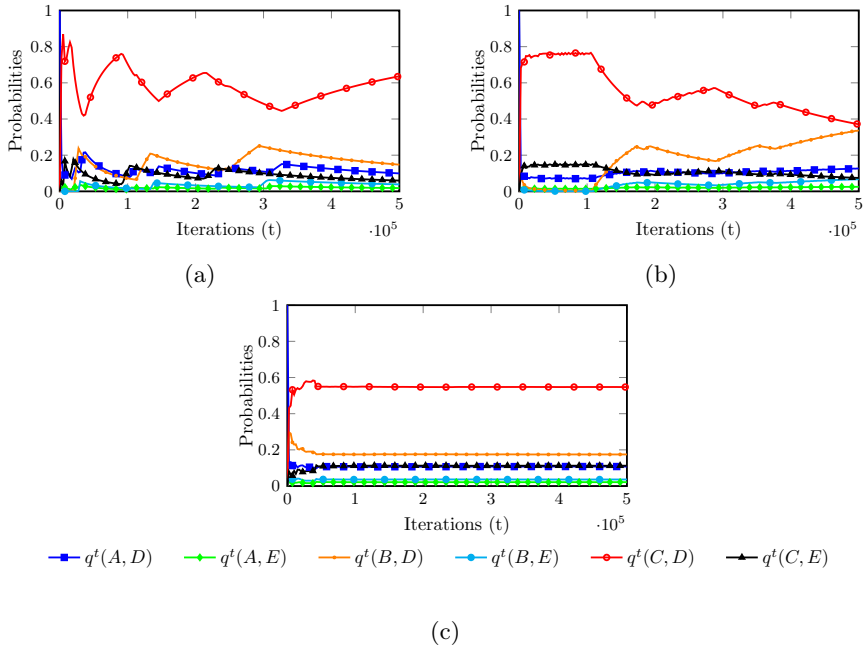


Fig. 2: Evolution of the empirical distribution over action profiles for the three algorithms: a) Regret Matching b) Blackwell's regret minimization and c) CPRM algorithm.

convergence expected from the perturbed Markov process and the conjecture stating that in the low perturbations regime the process should spend most the time close to a correlated equilibrium with synchronous players (playing profiles drawn from this distribution).

Dynamic setting Previously, we have assumed that the same stage game is played at every iteration. This is rarely the case in applications such as packet routing in networks where the set of players (or their population) may change over time. In this section, we study the flexibility of CPRM w.r.t. such updates and how the previous convergence results may be impacted as new players join or leave the game and utility functions change. CPRM cannot be used as is and must be slightly adapted to handle the arrival and departure of players (typically, players must update their regret matrices in the regret-minimization strategy ate every arrival or departure to allow for the equilibrium of the ongoing game to be approached). We do not enter the details of this second version of CPRM but show the numerical results demonstrating the potential efficiency to be studied in future works.

Assume that players start playing the game in Table 4 expanding into a three player game before evolving later on into a two-player game and eventually

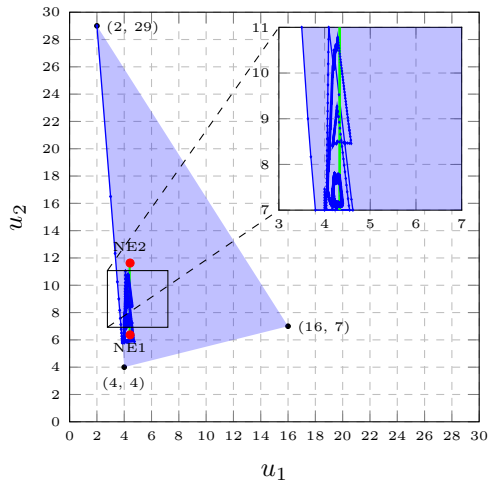


Fig. 3: Trajectory (in blue) of the expected payoffs starting from the initial action profile (A, D) and reaching a correlated β-equilibrium (close to the red segment). Utilities at mixed Nash equilibria ($MNE1$ and $MNE2$) are shown as yellow circles.

reverting back to the same three-player case afterwards as shown in Table 5. In the three-players 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 at $T_3 = 1529892$. Fig. 6a shows the evolution with 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 distributions over action profiles seem 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 [33] (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 action and action profiles. In a con-

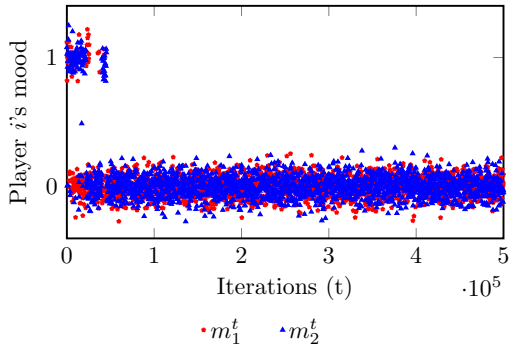


Fig. 4: Evolution of moods of the two players. Points take values 0 (mood "asyn") or 1 (mood "sync"). An artificial scattering is used to facilitate data visualization.

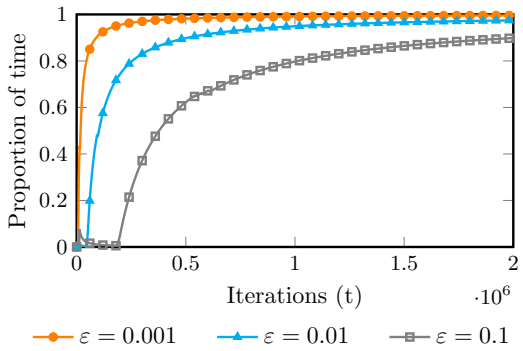


Fig. 5: Evolution in time of the proportion of time spent in the η-neighbourhood of the correlated β-equilibrium $q^*(\varepsilon)$ for $\eta = 0.01$.

gestion game with player-specific cost functions, each player selects a feasible subset of some resources and "pays" a cost defined as the sum over her selected resources of resource-based costs depending on the resource itself, the player and the total number of players selecting the resource. We consider the case where the resources are edges in a network, each player picking 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 $\mathcal{A}_\rangle$ 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 5: 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

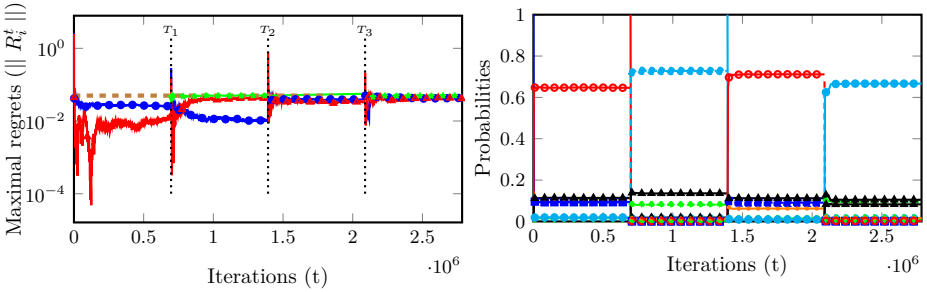
Let $f_e : \times_{i\in\mathcal{A}}\mathcal{A}_i \to \{0,\dots,N\}$ be the congestion function of edge e such that $f_e(\boldsymbol{a}) = |\{i \in \mathcal{N} : e \in a_i\}|$, i.e. the number of players using edge e. Given a strategy profile $\boldsymbol{a} \in \mathcal{A}$, player i has cost $c_i(\boldsymbol{a})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

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

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

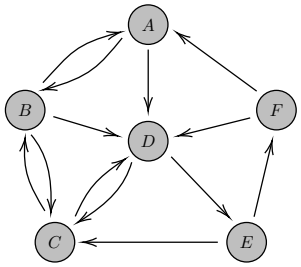
As before, we first have an interest in a constant stage game and then allow for the stage game to change because of arrival and departure of players.

Fig. 8a 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-players 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. 8b, indicating that this long-run distribution is indeed a correlated β-equilibrium distribution. Then, in the long-run, the players follow a correlated equilibrium routing policy of this game.



(a) (b)

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 is marked with horizontal dashed line on the left figure.

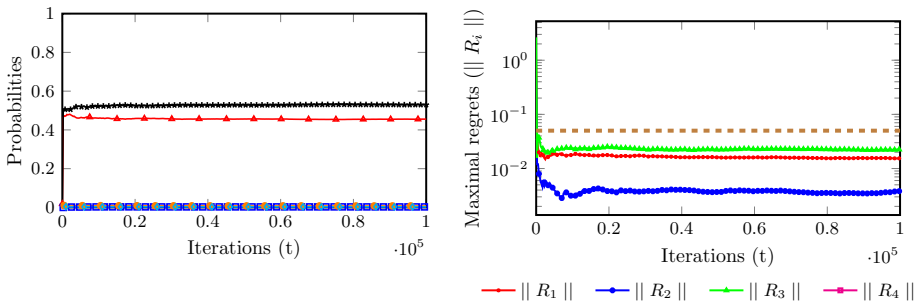


Player	**Source node**	**Destination node**
1	B	F
2	B	E
3	D	B
4	F	E

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 player 5 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. 8b 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.



(a) (b)

Fig. 8: (a) Evolution of the empirical distribution over the (five main) action profiles. (b) 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).

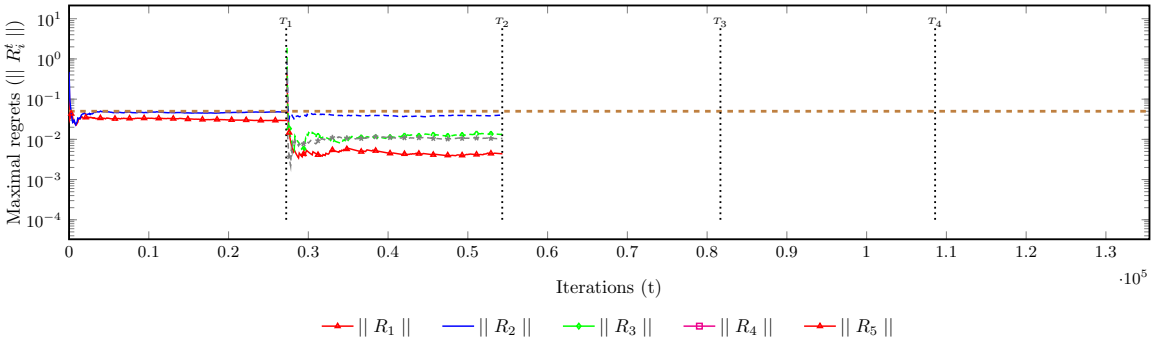


Fig. 9: Evolution of regrets with arrival and departure of players (at times indicated with vertical dotted lines). The approximate equilibrium threshold corresponds to the horizontal dashed line. Regret points that are not shown or curves correspond to very small values.

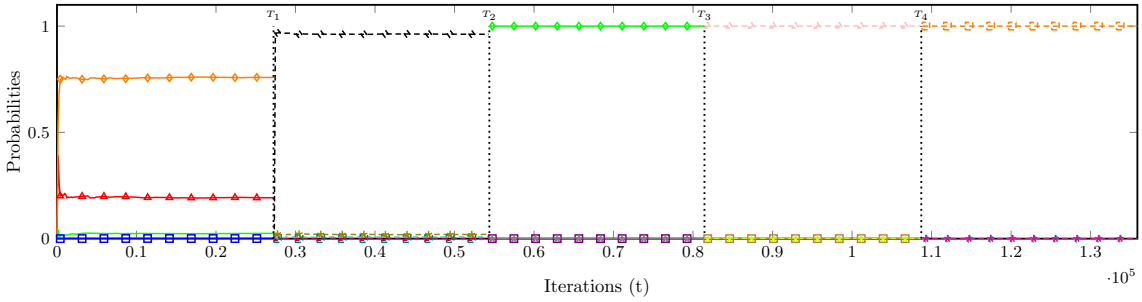


Fig. 10: Evolution of the empirical probability distribution. Only the five highest probabilities of action profiles are shown.

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 an adaptive heuristic to a correlated equilibrium distribution. We proposed a new adaptive heuristic, 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 point-wise. Additional experiments show that CPRM can be adapted to be compliant 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.

References

1. R. J. Aumann, "Subjectivity and correlation in randomized strategies," *Journal of Mathematical Economics*, vol. 1, no. 1, pp. 67–96, 1974.
2. ——, "Correlated equilibrium as an expression of bayesian rationality," *Econometrica: Journal of the Econometric Society*, pp. 1–18, 1987.
3. F. Forges, "Correlated equilibria and communication in games," *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models*, pp. 107–118, 2020.
4. R. B. Myerson, *Game theory: analysis of conflict*. Harvard university press, 1997.
5. S. Hart, "Adaptive Heuristics," *Econometrica*, vol. 73, no. 5, pp. 1401–1430, 2005.
6. S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, 2000.
7. ——, "A reinforcement procedure leading to correlated equilibrium," *Economics essays*, pp. 181–200, 2001.
8. D. P. Foster and R. V. Vohra, "Calibrated Learning and Correlated Equilibrium," *Games and Economic Behavior*, vol. 21, no. 1-2, pp. 40–55, Oct. 1997.
9. D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*, ser. MIT Press Books. The MIT Press, December 1998, vol. 1, no. 0262061945.
10. N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. USA: Cambridge University Press, 2006.
11. D. Foster and H. P. Young, "Regret testing: Learning to play nash equilibrium without knowing you have an opponent," *Theoretical Economics*, vol. 1, no. 3, pp. 341–367, 2006.
12. H. P. Young, "Learning by trial and error," *Games and economic behavior*, vol. 65, no. 2, pp. 626–643, 2009.
13. O. Boussaton, J. Cohen, J. Tomasik, and D. Barth, "On the distributed learning of nash equilibria with minimal information," in *2012 6th International Conference on Network Games, Control and Optimization (NetGCooP)*. IEEE, 2012, pp. 30–37.
14. P. Frihauf, M. Krstic, and T. Basar, "Nash equilibrium seeking in noncooperative games," *IEEE Transactions on Automatic Control*, vol. 57, no. 5, pp. 1192–1207, 2011.

15. F. Germano and G. Lugosi, "Global nash convergence of foster and young's regret testing," *Games and Economic Behavior*, vol. 60, no. 1, pp. 135–154, 2007.
16. S. Hart and A. Mas-Colell, "Uncoupled dynamics do not lead to nash equilibrium," *The American Economic Review*, vol. 93, no. 5, pp. 1830–1836, 2003.
17. J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, "Payoff-based dynamics for multiplayer weakly acyclic games," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 373–396, 2009.
18. B. S. Pradelski and H. P. Young, "Learning efficient nash equilibria in distributed systems," *Games and Economic behavior*, vol. 75, no. 2, pp. 882–897, 2012.
19. I. Ariel and Y. Babichenko, *Average testing and the efficient boundary*. Center for the study of Rationality, 2011.
20. J. R. Marden, H. P. Young, and L. Y. Pao, "Achieving pareto optimality through distributed learning," *SIAM Journal on Control and Optimization*, vol. 52, no. 5, pp. 2753–2770, 2014.
21. D. Blackwell *et al.*, "An analog of the minimax theorem for vector payoffs." *Pacific Journal of Mathematics*, vol. 6, no. 1, pp. 1–8, 1956.
22. V. Perchet and ,Université Paris-Diderot, Laboratoire de Probabilités et Modèles Aléatoires, UMR 7599, 8 place FM/13, Paris, "Approachability, regret and calibration: Implications and equivalences," *Journal of Dynamics & Games*, vol. 1, no. 2, pp. 181–254, 2014.
23. A. Greenwald, K. Hall, and R. Serrano, "Correlated q-learning," in *ICML*, vol. 3, 2003, pp. 242–249.
24. H. P. Borowski, J. R. Marden, and J. S. Shamma, "Learning to Play Efficient Coarse Correlated Equilibria," *Dynamic Games and Applications*, vol. 9, no. 1, pp. 24–46, Mar. 2019.
25. J. R. Marden, "Selecting efficient correlated equilibria through distributed learning," *Games and Economic Behavior*, vol. 106, pp. 114–133, 2017.
26. A. Celli, A. Marchesi, G. Farina, and N. Gatti, "No-regret learning dynamics for extensive-form correlated and coarse correlated equilibria," *CoRR*, vol. abs/2004.00603, 2020.
27. G. Farina, A. Celli, A. Marchesi, and N. Gatti, "Simple uncoupled no-regret learning dynamics for extensive-form correlated equilibrium," *CoRR*, vol. abs/2104.01520, 2021.
28. H. Jin, H. Guo, L. Su, K. Nahrstedt, and X. Wang, "Dynamic task pricing in multi-requester mobile crowd sensing with markov correlated equilibrium," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 1063–1071.
29. Q. Hu, Y. Nigam, Z. Wang, Y. Wang, and Y. Xiao, "A correlated equilibrium based transaction pricing mechanism in blockchain," in *2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*. IEEE, 2020, pp. 1–7.
30. L. Cigler and B. Faltings, "Reaching correlated equilibria through multi-agent learning," in *10th Conference on Autonomous Agents and Multiagent Systems AAMAS*, no. CONF, 2011.
31. H. P. Young, "The evolution of conventions," *Econometrica: Journal of the Econometric Society*, pp. 57–84, 1993.
32. T. Lykouris, V. Syrgkanis, and É. Tardos, "Learning and efficiency in games with dynamic population," in *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 2016, pp. 120–129.
33. R. W. Rosenthal, "A class of games possessing pure-strategy nash equilibria," *International Journal of Game Theory*, vol. 2, no. 1, pp. 65–67, 1973.

A Appendix: Numerical implementation of CPRM

Algorithm 1: Numerical implementation of 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 $(m_i^0)_{i\in\mathcal{N}}$, history $h^0 = (\mathbf{a}^0)$ and compute $(d_i^1)_{i\in\mathcal{N}}$
for $t = 1, 2, ...$ **do**
  /* Compute the empirical distribution of action profiles q^t */
  $\forall\ \boldsymbol{a} \in \mathcal{A}$, $\boldsymbol{q}^t(\boldsymbol{a}) \leftarrow \frac{1}{t}|\{\tau \leq t : \boldsymbol{a}^\tau = \boldsymbol{a}\}|$
  Draw an action profile $\boldsymbol{b}^t = (b_1, ..., b_n)$ from $\boldsymbol{q}^t$
  for $i \in \mathcal{N}$ **do**
    /* Play according to player i's mood & update mood */
    Draw uniformly a number 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
      Play the realization of the mixed strategy of Eq. (6)
    else
      if $\varepsilon^\gamma > var$ *or* $||d_i^t|| > \beta$ **then** $m_i^{t+1} \leftarrow$ asyn
      Play b_i
    end
  end
  /* Update the vector of the average payoff differences */
  $\forall i \in \mathcal{N}, \forall (j,k) \in \mathcal{A}_i \times \mathcal{A}_i, d_i^{t+1}(j,k) \leftarrow \frac{1}{t} \sum_{\substack{\tau \leq t \\ a_i^\tau = j}} \left[u_i\left(k, \boldsymbol{a}_{-i}^\tau\right) - u_i\left(a_i^\tau, \boldsymbol{a}_{-i}^\tau\right)\right]$
  $\boldsymbol{h}^{t+1} \leftarrow (\boldsymbol{h}^t, \boldsymbol{a}^{t+1})$
end