Courtesy as a Means to Anti-coordinate

Panayiotis Danassis, Boi Faltings
Artificial Intelligence Laboratory (LIA), École Polytechnique Fédérale de Lausanne (EPFL)
{panayiotis.danassis,boi.faltings}@epfl.ch

ABSTRACT

In this paper, we investigate the problem of anti-coordination under rationality constraints. This includes role allocation, task assignment, resource allocation, etc. Inspired by human behavior, we propose a framework (CA^{3}NONY) that enables fast convergence to efficient and fair allocations based on a simple convention of courtesy. We prove that following such convention induces a strategy which constitutes an approximate subgame-perfect equilibrium of the repeated allocation game with discounting. Simulation results highlight the effectiveness of CA^{3}NONY as compared to state-of-the-art bandit algorithms, since it achieves more than two orders of magnitude faster convergence, higher efficiency, fairness, and average payoff.

KEYWORDS

Multiagent learning; Noncooperative games: theory & analysis

1 INTRODUCTION

In multi-agent systems, agents are often called upon to implement a joint plan in order to maximize their rewards. Typically, a joint plan consists of two distinct elements: coordination, where agents are required to take the same action [27] [13], and anti-coordination, where the goal is to take distinct actions [14] [12] [15] [7]. This paper studies anti-coordination in repeated allocation games. This includes role allocation (e.g. teammates during a game), task assignment (e.g. employees of a factory), resource allocation (e.g. wireless bandwidth (channels) for IoT devices, parking spaces and/or charging stations for autonomous vehicles) etc. In many real world applications of the aforementioned scenarios the agents are indifferent to which role/task/resource they attain, as long as they receive one (e.g. wireless frequencies). Thus, this paper focuses on achieving a game-theoretically stable anti-coordination between 'indifferent' agents, leaving individual preferences for future work. Beyond the scope of repeated allocation games, the proposed framework can also be applied as a negotiation protocol in one-shot interactions. E.g. self-driving vehicles attempting to get a parking space can utilize such protocol in a simulated environment with message exchange.

A central coordinator who possesses complete information can recommend an action to each agent. Yet, an omniscient central coordinator is not always available, and in real-world applications with partial observability agents might not be willing to trust such recommendations. Moreover, inter-agent interactions might need to take place in an *ad-hoc* fashion, with previously un-encountered collaborators [29]. Planning in such environments becomes even more challenging. Part of this difficulty stems from the lack of responsiveness and/or communication between the participants. Typical ad-hoc approaches (e.g. Monte Carlo algorithms, Bayesian learning, bandit algorithms etc.) tend to require too many rounds to converge to be feasible in dynamic environments and real-life applications. Yet, humans are able to routinely coordinate in such an ad-hoc fashion.

One key concept that facilitates human ad-hoc coordination is the use of *conventions* [18]. Behavioral conventions are a fundamental part of human societies, yet they have not appeared meaningfully in empirical modeling of multi-agent systems. Inspired by human behavior, we propose the adoption of a simple convention based on *courtesy*. Courtesy is a moral virtue, an obligation that arises by the social nature of humans. Society demands that an individual should conduct himself in consideration of others. This allows for fast convergence, albeit it is not game theoretically sound; people adhere to it due to social pressure. Such problems become even more severe in situations with scarcity of resources. Under such conditions, courtesy breaks down and in the name of self-preservation people exhibit urgency and competitive behavior [16]. Thus, to satisfy our rationality constraint (i.e. self-interested agents do not have an incentive to deviate) in an artificial system, we need a deterrent mechanism.

In this paper we present a framework that reproduces courtesy, using monitoring and punishments as a deterrent mechanism. Our focus is on repeated allocation games with discounting. The proposed framework, CA^{3}NONY (**C**ontextual **A**nti-coordination in **A**d-hoc **Anony**mous games), achieves *fast convergence*, and high *efficiency* and *fairness* while being game theoretically sound. The agents follow a simple convention of courtesy, based on the decentralized allocation algorithm of [12]. Coupled to that, we introduce a monitoring scheme on the resource side to ensure equality among the participants and make adhering to the convention a rational choice. The main contributions of this paper are:

- Introduction of an anti-coordination framework (CA^{3}NONY) which consists of a *courteous convention* and a *monitoring scheme*.
- Proof that under such a framework, the use of the courteous convention induces strategies that constitute an approximate *subgame-perfect equilibrium*.
- Comparison to state-of-the-art bandit algorithms.

The rest of the paper is organized as follows. Section 2 situates CA^{3}NONY in the literature, Section 3 presents the proposed framework, and Section 4 provides simulation results. Finally, Section 5 concludes the paper.

2 RELATED WORK

In ad-hoc multi-agent coordination the goal is to design autonomous agents that achieve high flexibility and efficiency in a setting that admits no prior coordination between the participants [29]. Typical scenarios include the use of Monte Carlo algorithms [4], Bayesian learning [1], or bandit algorithms [9], [5]. Traditionally, pure ad-hoc approaches suffer from slow learning (e.g. [9]), which makes pure

ad-hoc coordination a very ambitious goal for real-life applications. In this paper we propose a middle-ground approach. Inspired by human ad-hoc coordination, we incorporate prior knowledge in the form of simple *conventions*. The coordination can still be considered ad-hoc as it is not pre-programmed, rather it involves learning. This allows for faster convergence to efficient and fair allocations compared to pure ad-hoc approaches.

A convention is defined as a customary, expected and self-enforcing behavioral pattern [32] [18]. In multi-agent systems, there are two scopes through which we study conventions. First, a convention can be considered as a behavioral rule, designed and agreed upon ahead of time or decided by a central authority [28] [31]. Second, a convention may emerge from within the system itself through repeated interactions between the participants [22] [31]. The proposed courteous convention falls on the first category. It is incorporated as prior knowledge to the agents, and, as proven in Section 3.5, it is self-enforcing since the induced by the convention strategies constitute an approximate subgame-perfect equilibrium of the repeated allocation game.

An alternative way to model the anti-coordination problem is as a multi-armed bandit problem [2], since the agents only receive partial (bandit) feedback. In multi-armed bandit problems an agent is given a number of arms (resources) and at each time-step has to decide which arm to pull to get the maximum expected reward. Bandit (or no-regret) algorithms typically minimize the total regret of each agent, which is the difference between the expected received payoff and the payoff of the best strategy in hindsight. As such, they satisfy our rationality constraint since they constitute an approximate correlated or coarse correlated equilibrium [24] [26]. Bandit algorithms have been successfully applied in many scenarios in recent years, like in 5G wireless networks [20] or (IoT)-driven cell networks [21]. Nevertheless, the studied problem presents many challenges: there is no stationary distribution (adversarial rewards), all agents are able to learn (similar to recursive modeling), and yielding gives a reward of 0 (desirable option for minimizing regret, but not in respect to fairness). Due to their ability to learn from partial feedback, bandit algorithms would be the natural choice for a pure ad-hoc approach.

Game theoretic equilibria are desirable (since they satisfy the rationality constraint), but hard to obtain. Deciding whether an anti-coordination (anonymous) game has a pure Nash equilibrium (NE) is NP-complete [8]. Furthermore, allocation games often admit undesirable equilibria: pure NE which are efficient but not fair, or mixed-strategy NE which are fair but not efficient [12]. Hence, iterative best-response algorithms are not satisfactory. On the other hand, an optimal correlated equilibrium (CE) of an anonymous game may be found in polynomial time [25]. However, in a multi-agent scenario, we are mostly interested in repeated interactions; agents who are able to learn and end up converging to an equilibrium. Moreover, we are interested in information-restrictive learning rules (i.e. completely uncoupled [30]), where each agent is only aware of his own history of action/reward pairs. Such an approach was applied in [12] to design a distributed algorithm for reaching efficient and fair CE in wireless channel allocation games. Yet, while the algorithm reaches an equilibrium in polynomial number of steps, cooperation to achieve this state is not rational. A self-interested agent could keep accessing a resource forever, until everyone else backs off. In this paper, we build upon the ideas of Cigler and Faltings and develop an anti-coordination strategy that constitutes an approximate subgame-perfect equilibrium, i.e. cooperation with the algorithm is a best-response strategy at each sub-game of the original stage game, given any history.

Finally, a generalization of anti-coordination games, called dispersion games, was described in [15]. In a dispersion game, agents are able to choose from several actions, favoring the one that was chosen by the smallest number of agents (analogous to minority games [10]). The authors in [15] define a maximal dispersion outcome as an outcome where no agent can switch to an action chosen by fewer agents. The agents themselves do not have any particular preference for the attained equilibrium. Contrary to that, we are interested in achieving an efficient and fair outcome. Expanding the studied techniques to tackle dispersion games, and therefore non-binary utilities, remains open for future research.

3 THE CA^{3}NONY FRAMEWORK

In this section, we present CA^{3}NONY , an anti-coordination framework for repeated allocation games with discounting (δ). The framework is applicable to any role, task, resource, etc. allocation scenario. For simplicity hereafter we will refer only to resources.

3.1 The Repeated Allocation Game

Let a 'resource' be any element that can be successfully assigned to only one agent at a time. At each time-step, $\mathcal{N} = \{1, \dots, N\}$ agents try to access $\mathcal{R} = \{1, \dots, R\}$ identical and indivisible resources, where possibly $N \gg R$. The set of available actions is denoted as $\mathcal{A} = \{Y, A_1, \dots, A_R\}$, where Y refers to yielding and A_r refers to accessing resource r. We assume that access to a resource is slotted and of equal duration. A successful access yields a positive payoff, while no access has a payoff of 0. If more than one agent access a resource simultaneously, a collision occurs and the colliding parties incur a cost $\zeta < 0$. Thus, the agents only receive a binary feedback of success or failure. Let a_n denote agent n's action, and $a_{-n} = \times_{\forall n' \in \mathcal{N} \setminus \{n\}} a_{n'}$ the joint action for the rest of the agents. The payoff function is defined as:

$$u_n(a_n, a_{-n}) = \begin{cases} 0, & \text{if } a_n = Y \\ 1, & \text{if } a_n \neq Y \wedge a_i \neq a_n, \forall i \neq n \\ \zeta, & \text{otherwise} \end{cases} \tag{1}$$

Conforming to real-world scenarios, we assume that each agent n is only aware of his own history of action/reward pairs, $\mathcal{H}_n^\tau = \{(\alpha_n^t, u_n(\alpha_n^t, \alpha_{-n}^t))_{\forall t \leq \tau}\}$.

Finally, we assume that the agents can observe side information from their environment at each time-step t. We call this side information context (e.g. time, date etc.). The agents utilize this context as a common signal in their decision-making process, a means to learn and anti-coordinate their actions. Let $\mathcal{K} = \{1, \dots, K\}$ denote the context space. We do not assume any a priori relation between the context space and the problem. The only constraints are that the values should repeat periodically, and satisfy $K = \lceil N/R \rceil$.

Algorithm 1 Pseudo-code of the CA^{3}NONY framework.

Require: $\forall n \in \mathcal{N}$ initialize g_n u.a.r. in $\mathcal{R}$.
Require: $\forall n \in \mathcal{N}$ allocate $c_n \leftarrow m$ of artificial cash (AC).
Require: Set fee $f_r \leftarrow m, \forall r \in \mathcal{R}$
for $k_t \in \mathcal{K}$ **do**
  Agents observe context k_t.
  if $g_n(k_t) = A_r$ **then**
    Agent n accesses resource r and
    pays access fee of f_r AC.
    if Collision(r) **then**
      Set $g_n(k_t) \leftarrow Y$ with prob. $p_{backoff} > 0$.
      Agent n gets reimbursed f_r AC.
    else
      Agent n gets reimbursed $(1-\xi)f_r$ AC.
    end if
  else if $g_n(k_t) = Y$ **then**
    Agent n monitors random resource $r \in \mathcal{R}$.
    if Free(r) **then**
      Set $g_n(k_t) \leftarrow A_r$ with probability 1.
    end if
  end if
end for
Set fee $f'_r \leftarrow (1-\xi)f_r, \forall r \in \mathcal{R}$

3.2 Adopted Convention

The adopted convention is based on the cooperative allocation algorithm of [12]. Each agent n has a strategy $g_n : \mathcal{K} \rightarrow \mathcal{A}$ that determines a resource to access at time-step t after having observed context k_t. The strategy is initialized uniformly at random in $\mathcal{R}$. If $g_n(k_t) = A_r$, then agent n accesses resource r. Otherwise, if $g_n(k_t) = Y$, the agent does not access a resource but instead chooses uniformly at random a resource r to monitor for activity. If it is free, then the agent updates $g_n(k_t) \leftarrow A_r$.

In [12], agents back-off probabilistically in case of a collision (set $g_n(k_t) \leftarrow Y$ with probability $p_{n_{backoff}}$). In such a setting, it is possible to reach a symmetric subgame-perfect equilibrium. But in order to actually play it, the agents need to be able to calculate it. It is not always possible to obtain the closed form of the back-off probability distribution of each resource. Furthermore, a self-interested agent could stubbornly keep accessing a resource forever, until everyone else backs off (also known as 'bully' strategy [19]).

Instead, we adopted a simple convention where agents are being *courteous*, i.e. if there is a collision, the colliding agents will back-off with some constant positive probability: $p_{n_{backoff}} = p > 0, \forall n \in \mathcal{N}$. Being courteous though, does not satisfy the rationality constraint. However, a uniform distribution of resources is socially optimal (i.e. fair allocations maximize the social welfare). Hence, if we introduce quotas to the use of resources and punishments upon violating them, courtesy induces rational strategies. In the following sections we introduce a monitoring scheme and prove that the resulting strategy constitutes an approximate subgame-perfect equilibrium.

3.3 Rationality

In order to ensure the proposed convention's rationality, the agents must be assured that they will eventually be successful, i.e. we must provide safeguards against the monopolization of resources. In this section we present a decentralized, self-regulated monitoring scheme based on artificial currency (which is used solely as an internal mechanism). The monitoring scheme deters agents from monopolizing resources to the point that each agent can access a resource only for one context value out of K. In order to be able to enforce such a scheme we need to employ *monitoring Authorities* (MA) at each resource. Initially all the agents that 'buy-in' are issued the same amount $m \in \mathbb{R}$ of artificial cash (AC). This amount also corresponds to the initial fee for every resource f_r. To allow access to resource r, the MA responsible for that resource charges f_r units of AC, and monitors the event. If there was a successful access, the MA reimburses the amount of $(1-\xi)f_r$ AC to the accessing agent, where $\xi \rightarrow 0 \in \mathbb{R}$ is a commission fee. Otherwise, the MA reimburses the full amount of f_r AC to the participating agents, so that they are able to try again for a different context value. The charging and reimbursement can be performed anonymously using the underlying architecture of any decentralized digital currency scheme (e.g. [23] [11]). Finally, after each period of context values, the MAs lower the fee to $f'_r \leftarrow (1-\xi)f_r, \forall r \in \mathcal{R}$. The pseudo-code of CA^{3}NONY is provided in Algorithm 1.

A valid monitoring scheme for our framework must prohibit the monopolization of resources. To see why this is the case for the proposed decentralized version, we can consider the following. After every successful access, the amount of AC that an agent possesses drops below the access fee of a resource. Hence, a rational agent will only access one resource for one context value. Waiting for the fee to drop to the point that $f_r = m/2$ is not a rational behavior since, assuming $\xi \rightarrow 0$, the number of iterations required to allow accessing two resources at the same time will reach ∞. At that point the rest of the agents will have reached a correlated equilibrium and the adversarial agent will not have an incentive to access an additional resource, besides the one that corresponds to him, since it would result in a collision. If, due to implementation constraints, we can not select a small ξ, the MAs can change the artificial currency every I periods, invalidating the old one and again making such strategy irrational.

3.3.1 Punishments. Along with the monitoring scheme, it is necessary to put punishments into effect for situations where agents are able to access resources without paying. E.g. in a wireless scenario, if an agent transmits to some other than the designated channel, then his packets will no longer be relayed. In other words, punishments are application specific, and only needed in applications where the resources are publicly available.

3.4 Rate of convergence

Theorem 3.1. *In a repeated allocation game with N agents and R resources the expected number of steps before Algorithm 1 converges to a correlated equilibrium is given by Equation 2.*

$$O\left(N\left(\log\left(\left\lceil\frac{N}{R}\right\rceil\right)+1\right)(\log(N)+R)\right) \tag{2}$$

Proof. The employed monitoring authorities do not affect the computational complexity of Algorithm 1, since their time complexity cost is $O(1)$, i.e. independent of the input parameters (N, R). The adopted learning rule is based on the cooperative allocation

algorithm of [12]. Theorems 12 and 13 of [12] prove that for N agents, $R \geq 1$, $K \geq 1$, and back-off probability $0 < p < 1$, the expected number of steps before the learning algorithm converges to a correlated equilibrium is:

$$O\left((K\log(K)+2K)\,R\frac{1}{1-p}\left(\frac{1}{p}\log(N)+R\right)+1\right) \tag{3}$$

.

For a constant back-off probability and $K = \lceil N/R \rceil$, Equation 3 gives the required bound. □

3.5 Courtesy Pays Off

In this section we prove that if the agents back-off with a constant positive probability $p_{backoff} > 0$, then Algorithm 1 induces a strategy that is *almost rational*; no agent can improve his payoff by more than $\epsilon > 0$ (ϵ-equilibrium).

Let $U_n^\tau(\sigma,\delta) = \sum_{t=0}^{\tau}\delta^t u_n(a_n^t, a_{-n}^t)$ denote the cumulative payoff of agent n that follows strategy σ up to time-step τ in a repeated allocation game with discounting (where $\delta \in (0,1)$ is the discount factor). We prove the following theorem.

Theorem 3.2. *Suppose that in a repeated allocation game with discounting (δ) the agents who collide back-off with a constant probability $p_{backoff} > 0$. Let σ_p denote the aforementioned strategy (courteous strategy). Let $\mathbb{E}(U_n^\infty(\sigma_p,\delta))$ be the expected cumulative payoff for each agent in this case and $\mathbb{E}(U_n^\infty(\sigma_*,\delta))$ be the expected payoff of the best-response strategy σ_*. Then $\forall\epsilon > 0, \exists\delta_0, 0 < \delta_0 < 1$ such that $\forall\delta, \delta_0 \leq \delta < 1$:*

$$\mathbb{E}(U_n^\infty(\sigma_p,\delta)) > (1-\epsilon)\mathbb{E}^*(U_n^\infty(\sigma_*,\delta))$$

Proof. Suppose that an agent could access a resource at every time-step. His payoff would be $1+\delta+\delta^2+\delta^3+\cdots = \frac{1}{1-\delta}$ for $|\delta| < 1$. But, the introduced monitoring scheme prohibits from monopolizing resources, i.e. each agent can only access a resource for his corresponding context value. As a result, the best-response strategy's payoff for some δ is bounded by:

$$\mathbb{E}(U_n^\infty(\sigma_*,\delta)) \leq \frac{1}{K}\frac{1}{1-\delta} \tag{4}$$

When agents adopt the courteous convention, in each round until they converge to a correlated equilibrium, they receive a payoff between $\zeta < 0$ (collision cost) and 1. After convergence, their expected payoff is:

$$\delta^\tau + \delta^{\tau+K} + \delta^{\tau+2K} + \cdots = \sum_{i=0}^{\infty}\delta^{\tau+iK} = \frac{\delta^\tau}{1-\delta^K}$$

where τ is the number of steps it took them to converge. Hence the convention induced strategy's payoff is at least:

$$\mathbb{E}(U_n^\infty(\sigma_p,\delta)) \geq \sum_{\tau=0}^{\infty} Pr[\text{conv. in } \tau \text{ steps}]\cdot\left(\zeta\frac{1-\delta^\tau}{1-\delta}+\frac{\delta^\tau}{1-\delta^K}\right)$$

We can define a random variable X such that $X = \tau$ if the algorithm converges after exactly τ steps. Since δ^x is a convex function we have that $\mathbb{E}(\delta^x) \geq \delta^{\mathbb{E}(x)}$, therefore:

$$\mathbb{E}(U_n^\infty(\sigma_p,\delta)) \geq \zeta\frac{1-\delta^{\mathbb{E}(X)}}{1-\delta}+\frac{\delta^{\mathbb{E}(X)}}{1-\delta^K} \tag{5}$$

By dividing Equation 5 by Equation 4 we get:

$$\frac{\mathbb{E}(U_n^\infty(\sigma_p,\delta))}{\mathbb{E}(U_n^\infty(\sigma_*,\delta))} \geq \frac{\zeta(1-\delta^{\mathbb{E}(X)})(1-\delta)^{-1}+\delta^{\mathbb{E}(X)}(1-\delta^K)^{-1}}{K^{-1}(1-\delta)^{-1}} \tag{6}$$

$\mathbb{E}(X)$ does not depend on δ. Moreover, $\delta^{\mathbb{E}(X)}$ is continuous in δ, monotonous, and $\lim_{\delta\to1^-}\delta^{\mathbb{E}(X)} = 1$. Thus, we can take the limit of Equation 6 as $\delta \to 1^-$, which equals:

$$\lim_{\delta\to1^-}\frac{\mathbb{E}(U_n^\infty(\sigma_p,\delta))}{\mathbb{E}(U_n^\infty(\sigma_*,\delta))} = 1$$

□

In order to guarantee rationality, the discount factor δ must be close to 1 since, as δ gets closer to 1, the agents do not care whether they access now or in some future round. Since the proposed monitoring scheme guarantees that every agent will access a resource for his corresponding context value, when $\delta \to 1$, the expected payoff for agents who are accessing a resource and for those who have not accessed a resource yet will be the same. In other words, the cost (overhead) of learning the correlated equilibrium decreases.

Assuming that the agents are indifferent in claiming a resource in a period of T_{ind} rounds ($\delta_t = 1, \forall t \leq T_{ind}$), we can use the Markov bound to prove that with high probability the proposed algorithm will converge in under T_{ind} rounds, hence satisfying the rationality constraint. We assume the agents are willing to accept quasilinear 'delay' with regard to the number of resources R, the number of agents N, and the size of the context space K, specifically:

$$T_{ind} = O\left(R\log(R)N\log(N)K\log(K)\right) \tag{7}$$

Using the Markov bound, it follows that the probability that the proposed algorithm takes more than the accepted number of steps T_{ind} to converge is:

$$Pr[\neg\text{conv. after } T_{ind}] = O\left(\frac{R+\log(N)}{N\log(N)\log(R)}\right) \tag{8}$$

Even though when used directly the Markov's inequality generally does not give very good bounds, Equation 8 proves that our algorithm converges in the required time with high probability. We can further strengthen our rationality hypothesis by using a tighter bound (e.g. Chebyshev's inequality), albeit computing the theoretical variance of the convergence time is an arduous task, thus it remains open for future work.

4 EXPERIMENTAL EVALUATION

In this section we model the resource allocation problem as a multi-armed bandit problem and provide simulation results of CA^{3}NONY 's performance in comparison to state-of-the-art, well established bandit algorithms, namely the EXP4 [3], EXP4.P [6], and EXP3 [3]. In every case we report the average value over 128 runs of the same simulation. The back-off (being courteous) probability of CA^{3}NONY is set to $p_{backoff} = 0.5$, which is the optimal value according to Equation 3. For the EXP family of algorithms, the input parameters are set to their optimal values, as prescribed in [3], and [6]. Finally, we assume a reward of 1 for a successful access, -1 if there is a collision, and 0 if the agent yielded at that time-step.

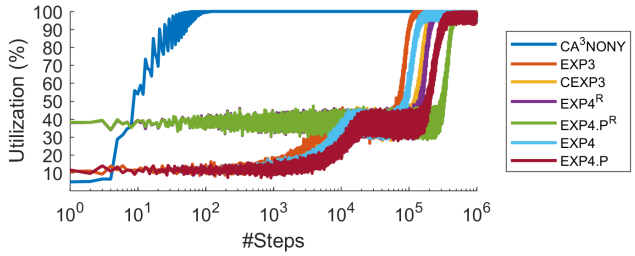

Figure 1: Utilization: $R = 4, N = 16, K = 4$ **(x-axis in log scale).**

4.1 Employed Bandit Algorithms

In our setting, the reward of each arm (resource) does not follow a fixed probability distribution (adversarial setting). Furthermore, the agents are able to observe side-information (context) at each time-step t, before making their decision. The arm that yields the highest expected reward can be different depending on the context. Hence we will focus on adversarial contextual bandit algorithms (see [33] for a survey on contextual bandits).

A typical method for solving adversarial contextual bandit problems is to model them as bandit problems with expert advice. In this method we assume a set of experts $\mathcal{M} = \{1, \ldots, M\}$ who, at each time-step t, generate a probability distribution on which arm to pull depending on the context. A no-regret algorithm in this case performs asymptotically as well as the best expert. Such an algorithm is the EXP4, which achieves a regret bound of $O(\sqrt{TR \log M})$, where T is the time horizon. Yet, the EXP4 exhibits high variance [33], hence the regret bound holds only in expectation. Beygelzimer et al. presented EXP4.P which achieves the same regret, $O(\sqrt{TR \log(M/\lambda)})$, with high probability $(1-\lambda, \forall\lambda)$ by combining the confidence bounds of UCB1 [2] and EXP4.

The computational complexity and memory requirements of the above algorithms are linear in M, making them intractable for large number of experts. In order to deal with the increased complexity in larger simulations and to ensure fairness in the presented results, we have included two restricted versions of the EXP4 algorithms, denoted as 'EXP4$^{\text{R}}$' and 'EXP4.P$^{\text{R}}$'. In these versions, the set $\mathcal{M}$ is limited to uniform correlated equilibria, the same set of equilibria that CA^{3}NONY converges to. This is equivalent to enabling the use of the monitoring scheme by the EXP4 algorithms, i.e. 'EXP4$^{\text{R}}$' and 'EXP4.P$^{\text{R}}$' utilize the employed monitoring authorities.

An alternative approach is to use a non-contextual adversarial bandit algorithm, such as the EXP3, whose weak regret is bounded by $O(\sqrt{TR \log R})$. Moreover, we can convert EXP3 to a contextual algorithm by setting up a separate instance of the EXP3 $\forall k \in \mathcal{K}$. We call this version 'CEXP3'. This results in a contextual bandit algorithm which has the edge over EXP4 from an implementation viewpoint since its running time at each time-step is $O(R)$ and its memory requirement is $O(KR)$ (CA^{3}NONY 's running time at each time-step is $O(1)$ and its memory requirement $O(K)$).

4.2 Simulation Results

4.2.1 Convergence Speed & Efficiency. We know that CA^{3}NONY converges to a correlated equilibrium which is *efficient*. If all agents follow the courteous convention of Algorithm 1, the system converges to a state where no resources remain un-utilized and there are no collisions (Theorem 13 of [12]). Furthermore, Theorem 3.1 argues for fast convergence. The former are both corroborated by Figure 1 (similar results ($> \times 10^2$ faster convergence) were acquired for $R \in \{2, 8, 16\}$ as well.). Figure 1 depicts the total utilization of resources for a simulation period of $T = 10^6$ time-steps. Note that the x-axis is in logarithmic scale. CA^{3}NONY converges significantly ($> \times 10^2$) faster than the bandit algorithms to a state of 100% efficiency. On the other hand, the bandit algorithms exhibit high variance, never achieve 100% efficiency, and are not able to handle efficiently the increase in context space size and number of resources.

Table 1: Fairness (Jain Index), $K = R, N = R \times K$.

	$R = 2$	$R = 4$	$R = 8$	$R = 16$
CA^{3}NONY	1.0000	1.0000	1.0000	1.0000
EXP3	0.5000	0.2500	0.1250	0.0625
CEXP3	0.6875	0.5905	0.5317	0.9621
EXP4(.P)$^{\text{R}}$	1.0000	0.9999	0.9880	0.9789
EXP4(.P)	0.7157	0.6206	N/A	N/A

Table 2: Average Payoff, $K = R, N = R \times K$.

	$R = 2$	$R = 4$	$R = 8$	$R = 16$
CA^{3}NONY	45.2	15.0	-6.5	-25.4
(C)EXP3	-60.7	-94.2	-99.9	-100.0
EXP4(.P)$^{\text{R}}$	-59.2	-81.0	-90.7	-95.4
EXP4(.P)	-60.5	-94.3	N/A	N/A
CA^{3}NONY	50.4	354.9	9196.3	62451.0
(C)EXP3	-66.4	-1370.0	-68410.5	-940766.0
EXP4(.P)$^{\text{R}}$	-65.1	-1186.8	-66921.8	-954283.6
EXP4(.P)	-66.4	-1375.8	N/A	N/A

4.2.2 Fairness. The usual predicament of efficient equilibria for allocation games is that they assign the resources only to a fixed subset of agents, which leads to an unfair result (e.g. an efficient PNE is for R agents to access and $N - R$ agents to yield). This is not the case for CA^{3}NONY , which converges to an equilibrium that is not just efficient but fair as well. Due to the enforced monitoring scheme, all users acquire the same amount of resources. As a measure of fairness, we will use the Jain index [17]. The Jain index exhibits a lot of desirable properties such as: population size independence, continuity, scale and metric independence, and boundedness. For an allocation game of N users, such that the n^{th} user receives an allocation of x_n, the Jain index is given by $\mathbb{J}(x) = \left|\sum_{n\in\mathcal{N}} x_n\right|^2 / N \sum_{n\in\mathcal{N}} x_n^2$. An allocation is considered fair, iff $\mathbb{J}(x) = 1$.

Table 1 presents the expected Jain Index of the evaluated algorithms at the end of the time horizon T. CA^{3}NONY converges to a fair equilibrium, achieving a Jain index of 1. The EXP4$^{\text{R}}$ and EXP4.P$^{\text{R}}$ were the fairest amongst the bandit algorithms, achieving a Jain index of close to 1. This is to be expected since the set of experts $\mathcal{M}$ is limited to the same set of equilibria that CA^{3}NONY converges to. On the other hand, the rest of the bandit algorithms (EXP4, EXP4.P, EXP3) performed considerably worse, with the EXP3

exhibiting the worst performance in terms of fairness, equal to a PNE's: $\mathbb{J}_{PNE}(x) = \frac{R^2}{NR} = \frac{1}{K} = \mathbb{J}_{EXP3}(x)$. The clustered pairs of bandit algorithms in Tables 1 & 2 exhibited $< 0.5\%$ difference, hence we included the average value of the pair.

4.2.3 Average Payoff. The average payoff corresponds to the total discounted payoff an agent would receive in the time horizon T. This is an important metric since it is an essential indicator of the algorithm's individual performance. Table 2 presents the average payoff for the studied algorithms. The discount factor was set to $\delta = 0.99$. At the top half we do not assume any indifference period, while at the bottom half we assume quasilinear indifference period ($\delta_t = 1, \forall t \leq T_{ind}$, where T_{ind} is given by Equation 7).

Once more, CA^{3}NONY significantly outperforms all the bandit algorithms. The latter have relatively similar performance, with EXP4^R and EXP4.P^R being the best amongst them. It is worth noting that adding an indifference period has a dramatic effect on the results. CA^{3}NONY achieves a large increase on average payoff, while the opposite happens for the bandit algorithms. This is because the learning rule of Algorithm 1 prohibits from accessing an already claimed resource, thus there are no collisions after the first round (of each joining player in a dynamic population) and the payoff is ≥ 0. On the contrary, bandit algorithms constantly explore (they assign a positive probability mass to every arm) which leads to collisions. In a multi-agent system where every agent learns this can have a cascading effect. The latter becomes apparent when fixing $\delta = 1$ for T_{ind} steps. The collision cost remains high for longer which, as seen by Table 2, has a significant impact on the bandit algorithms' performance.

4.2.4 Large Scale Systems. The innovation of CA^{3}NONY stems from the adoption of a simple convention which in turns allows its applicability to large scale multi-agent systems. To evaluate the latter, Figures 2 and 3 depict the convergence time for increasing number of resources R, and increasing context space size K respectively. Both graphs are in a double logarithmic scale, and the error bars represent one standard deviation of uncertainty. The total number of agents for each simulation is given in both cases by $N = R{\times}K$. Thus, the two largest simulations involve $16 \times 1024 = 16.384$ agents. Along with CA^{3}NONY , we depict the fasted (based on the previous simulations) of the bandit algorithms, namely EXP3. In both cases we acquire $\times 10^3 - \times 10^5$ faster convergence. The above validate CA^{3}NONY 's performance in both scenarios with abundance ($N \approx R$ or small K), and scarcity of resources ($N \gg R$ or large K). As depicted, CA^{3}NONY is significantly faster than the EXP3 and can gracefully handle an increasing number of resources, and large context spaces. Finally note that, in several of the simulations, EXP3 was unable to reach its convergence goal of 90% efficiency (utilization of resources) in reasonable amount of computation time (1.5×10^8 time-steps), hence the resulting gaps in EXP3's lines in Figures 2 and 3. Especially in situations with scarcity of resources the utilization was significantly lower.

5 CONCLUSION

In this paper we proposed CA^{3}NONY , an anti-coordination framework under rationality constraints. CA^{3}NONY is based on a simple convention of courtesy, which prescribes a positive back-off probability in case of a collision. Coupled with a monitoring scheme which deters the monopolization of resources, we proved that the induced strategy constitutes an approximate subgame-perfect equilibrium. We compared CA^{3}NONY to state-of-the-art bandit algorithms, namely the EXP4, EXP4.P and EXP3. Simulation results demonstrated that CA^{3}NONY outperforms these algorithms by achieving more than two orders of magnitude faster convergence, while converging to a totally fair allocation, and providing higher average payoff for the agents. The efficiency of CA^{3}NONY stems from the adoption of the human-inspired convention of courtesy. The aforementioned gains corroborate our choice and suggest that human-inspired conventions may prove beneficial in other ad-hoc coordination scenarios as well or other classes of anonymous games.

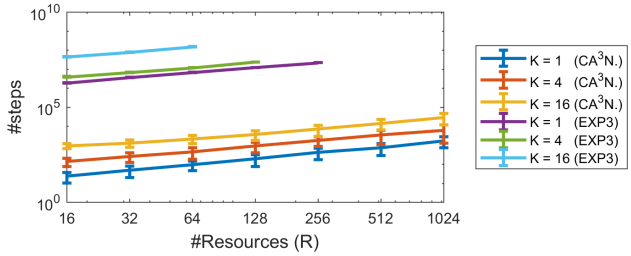

Figure 2: Convergence time for increasing number of resources R and varying context space size K, $N = R{\times}K$ (double log scale).

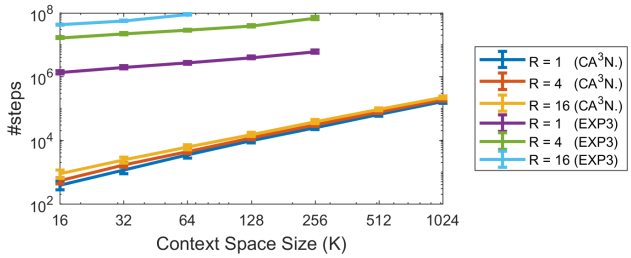

Figure 3: Convergence time for increasing context space size K and varying number of resources R, $N = R \times K$ (double log scale).

REFERENCES

[1] Stefano V Albrecht, Jacob W Crandall, and Subramanian Ramamoorthy. 2016. Belief and truth in hypothesised behaviours. *Artificial Intelligence* (2016).

[2] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47, 2 (2002), 235–256. https://doi.org/10.1023/A:1013689704352

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing* (2002).

[4] S. Barrett, A. Rosenfeld, S. Kraus, and P. Stone. 2017. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence* (2017).

[5] S. Barrett and P. Stone. 2011. Ad hoc teamwork modeled with multi-armed bandits: An extension to discounted infinite rewards. In *Proc. of 2011 AAMAS Workshop on Adaptive and Learning Agents.*

[6] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual bandit algorithms with supervised learning guarantees. In *Proc. of the Fourteenth Int. Conf. on Artificial Intelligence and Statistics.* 19–26.

[7] Yann Bramoullé, Dunia López-Pintado, Sanjeev Goyal, and Fernando Vega-Redondo. 2004. Network formation and anti-coordination games. *International*

Journal of Game Theory 33, 1 (2004), 1–19.
[8] Felix Brandt, Felix Fischer, and Markus Holzer. 2009. Symmetries and the Complexity of Pure Nash Equilibrium. *J. Comput. Syst. Sci.* 75, 3 (May 2009), 163–177. https://doi.org/10.1016/j.jcss.2008.09.001
[9] Mithun Chakraborty, Kai Yee Phoebe Chua, Sanmay Das, and Brendan Juba. 2017. Coordinated Versus Decentralized Exploration In Multi-Agent Multi-Armed Bandits. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 164–170. https://doi.org/10.24963/ijcai.2017/24
[10] Damien Challet, Matteo Marsili, Yi-Cheng Zhang, et al. 2013. Minority games: interacting agents in financial markets. *OUP Catalogue* (2013).
[11] Dimitris Chatzopoulos, Sujit Gujar, Boi Faltings, and Pan Hui. 2016. LocalCoin: An ad-hoc payment scheme for areas with high connectivity: poster. In *Proc. of the 17th ACM Int. Symposium on Mobile Ad Hoc Networking and Computing*. ACM.
[12] Ludek Cigler and Boi Faltings. 2013. Decentralized anti-coordination through multi-agent learning. *Journal of Artificial Intelligence Research* 47 (2013), 441–473.
[13] Russell Cooper. 1999. *Coordination Games*. Cambridge University Press. https://doi.org/10.1017/CBO9780511609428
[14] Panayiotis Danassis and Boi Faltings. 2018. Learning in Ad-hoc Anti-coordination Scenarios. https://www.aaai.org/ocs/index.php/SSS/SSS18/paper/view/17485
[15] Trond Grenager, Rob Powers, and Yoav Shoham. 2002. Dispersion games: general definitions and some specific learning results. In *AAAI/IAAI*.
[16] Shipra Gupta and James W. Gentry. 2016. The behavioral responses to perceived scarcity âĂŞ the case of fast fashion. *The International Review of Retail, Distribution and Consumer Research* (2016). https://doi.org/10.1080/09593969.2016.1147476 arXiv:https://doi.org/10.1080/09593969.2016.1147476
[17] Raj Jain, Dah-Ming Chiu, and W. Hawe. 1998. A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems. *CoRR* cs.NI/9809099 (1998). http://arxiv.org/abs/cs.NI/9809099
[18] David Lewis. 2008. *Convention: A philosophical study*. John Wiley & Sons.
[19] Michael L. Littman and Peter Stone. 2002. *Implicit Negotiation in Repeated Games*. Springer Berlin Heidelberg, 393–404. https://doi.org/10.1007/3-540-45448-9_29
[20] Setareh Maghsudi and Ekram Hossain. 2015. Multi-armed Bandits with Application to 5G Small Cells. *CoRR* abs/1510.00627 (2015). http://arxiv.org/abs/1510.00627
[21] Setareh Maghsudi and Ekram Hossain. 2016. Distributed Cell Association for Energy Harvesting IoT Devices in Dense Small Cell Networks: A Mean-Field Multi-Armed Bandit Approach. *CoRR* abs/1605.00057 (2016). http://arxiv.org/abs/1605.00057
[22] Mihail Mihaylov, Karl Tuyls, and Ann Nowé. 2014. A decentralized approach for convention emergence in multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 28, 5 (2014), 749–778.
[23] Satoshi Nakamoto. 2008. Bitcoin: A peer-to-peer electronic cash system. (2008).
[24] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. 2007. *Algorithmic game theory*. Vol. 1. Cambridge University Press Cambridge.
[25] Christos H. Papadimitriou and Tim Roughgarden. 2008. Computing Correlated Equilibria in Multi-player Games. *J. ACM* 55, 3, Article 14 (Aug. 2008), 29 pages. https://doi.org/10.1145/1379759.1379762
[26] Tim Roughgarden. 2016. *Twenty Lectures on Algorithmic Game Theory* (1st ed.). Cambridge University Press, New York, NY, USA.
[27] Thomas C Schelling. 1960. The strategy of conflict. *Cambridge, Mass* (1960).
[28] Yoav Shoham and Moshe Tennenholtz. 1995. On social laws for artificial agent societies: off-line design. *Artificial Intelligence* (1995). https://doi.org/10.1016/0004-3702(94)00007-N
[29] P. Stone, G. A. Kaminka, S. Kraus, and J. S. Rosenschein. 2010. Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination. In *Proc. of the Twenty-Fourth Conf. on Artificial Intelligence*.
[30] Mohammad Sadegh Talebi. 2013. Uncoupled Learning Rules for Seeking Equilibria in Repeated Plays: An Overview. *CoRR* abs/1310.5660 (2013). http://arxiv.org/abs/1310.5660
[31] A. Walker and M. J. Wooldridge. 1995. Understanding the Emergence of Conventions in Multi-Agent Systems. In *ICMAS95*. http://groups.lis.illinois.edu/amag/langev/paper/walker95understandingThe.html
[32] H Peyton Young. 1996. The economics of convention. *The Journal of Economic Perspectives* (1996).
[33] Li Zhou. 2015. A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326* (2015).