Learning in Ad-Hoc Anti-Coordination Scenarios

Panayiotis Danassis, Boi Faltings
Artificial Intelligence Laboratory (LIA), École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
Email: {panayiotis.danassis, boi.faltings}@epfl.ch

Abstract

We present a brief overview of learning dynamics for anti-coordination in ad-hoc scenarios. Specifically, we consider multi-armed bandit algorithms, reinforcement learning, and symmetric strategies for the repeated resource allocation game. In a multi-agent system with dynamic population where every agent is able to learn, the anti-coordination problem exhibits unique challenges. Thus, it is essential for the success of a joint plan that the agents can quickly and robustly learn their optimal behavior. In this work we will focus on convergence rate, efficiency, and fairness in the final outcome.

1 Introduction

In multi-agent systems, most scenarios require coordination on the same value which involves solving the consensus problem, a well-studied problem in distributed computing (Coulouris, Dollimore, and Kindberg 2005). However, there are also many situations where agents are required to choose distinct actions as in 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. This is called *anti-coordination*. Figure 1 provides an illustrative example. For simplicity, we focus on resource allocation scenarios, although the considered learning models can be applied in any analogous anti-coordination scenario.

Anti-coordination in multi-agent systems presents many unique challenges. First, it requires agents to take different actions while facing the same problem. Hence, we need agents that are able to learn to behave *differently* in the presence of (possibly) identical agents while having similar preferences across their available actions. An autonomous vehicle would prefer the route with the least traffic, an IoT device would prefer the higher bandwidth channel, a bidding agent participating in multiple auctions would prefer the one with the fewer participants, etc. Nevertheless, in order to achieve high efficiency, we need some agents to take less desirable actions. An added challenge is ensuring fairness in the final outcome, i.e. make sure that those agents are not exploited,

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

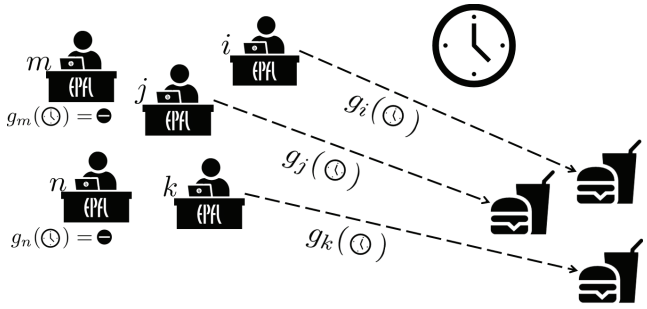

Figure 1: Every day N employees have lunch at a cafeteria which accommodates R patrons, thus their goal is to anti-coordinate their lunch breaks. Each of them has a strategy ($g_n, \forall n \in \{1, \ldots, N\}$) for selecting their lunch time. All employees have similar preferences (e.g. have lunch between 12p.m. - 2p.m., find an empty seat etc.). Each time they attempt to have a lunch break, they update their strategy based on their personal feedback of success or failure.

and ensuring that self-interested, rational agents are not able to manipulate the algorithm to maximize their utility. Furthermore, in real world applications agents tend to receive only partial feedback; i.e. each agent is only aware of his own history of action/reward pairs. Hence, we require completely uncoupled learning rules and agents that are capable of achieving high efficiency and fast convergence in such information-restrictive settings. Finally, intra-agent interactions might need to take place in an ad-hoc fashion, which brings forth the need for robust agents that are able to coordinate with previously un-encountered participants (Stone et al. 2010). However, planning in such environments becomes even more challenging. Part of this difficulty stems from the lack of responsiveness and/or communication between the participants.

Little work has been done in anti-coordination problems as compared to classical coordination scenarios. Mapping anti-coordination to the consensus problem results in an exponential expansion of the solution space. Hence, special effort is required from a learning perspective. In this paper we present a brief comparative overview of multi-agent learning paradigms applicable to the anti-coordination setting. The rest of the paper is organized as follows. Section

2 provide a formal definition of the repeated resource allocation (anti-coordination) problem, Section 3 presents the evaluated multi-agent learning models, and finally, Section 4 concludes the paper.

2 Preliminaries

2.1 The Repeated Resource Allocation Problem

In this section we formally define the repeated resource allocation problem. The goal for the agents is to maximize their discounted cumulative payoff. We refer to a 'resource' as 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. The set of available actions is denoted as $\mathcal{A} = \{Y, A_1, \dots, A_R\}$, where Y refers to yielding, while 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 accesses a resource simultaneously, a collision occurs and the colliding parties incur a cost $\zeta < 0$. The payoff function is defined by Equation 1, where a_n denotes 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.

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

In accordance to real-world phenomena we furthermore assume that the agents receive only partial feedback of success or failure; i.e. each agent n is only aware of his own history of action/reward pairs, $\mathcal{H}_n^t = \{(\alpha_n^\tau, u_n(\alpha_n^\tau, \alpha_{-n}^\tau))_{\forall \tau \leq t}\}$. The payoff matrix of the stage-game of a simple 1-resource, 2-agents, repeated resource allocation game is presented in Figure 2.

Finally, we assume that the agents can observe side information (context) from their environment at each time-step t (e.g. time, date etc. in the example of Figure 1), before taking their action. 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 constraint is that the context values should repeat periodically. In this work we assume that the context is a set of random integers. The motivation behind the introduction of the context space will become apparent in the following section. In short, we want to achieve high efficiency and fairness. In anti-coordination games with completely uncoupled learning rules such a goal is hard to attain since the aforesaid rules do not allow for correlation between the agents. The introduction of a common signal (such as the proposed context) resolves that issue.

2.2 Solution Concepts

In this section we examine possible game theory[1] solution concepts of the repeated resource allocation game, focusing on the following two axes:

[1] See (Nisan et al. 2007) for an introduction to game theory.

	Y	A
Y	0, 0	0, 1
A	1, 0	ζ, ζ

Figure 2: Resource allocation game, $R = 1, N = 2$. Two agents want to access a single resource. Both of them have two actions, either to yield (Y), and get a payoff of 0, or access (A). If only one of the agents accesses the resource, he gets a payoff of 1. But if both of them access the resource at the same time, they collide and both incur a cost $\zeta < 0$.

i *Efficiency*: Percentage of utilized resources after convergence (alternatively, social welfare).

ii *Ex-post Fairness*: Equality of allotted resources after convergence (alternatively, ex-post expected payoff).

As a measure of fairness, we will use the Jain index (Jain, Chiu, and Hawe 1998). The Jain index exhibits a lot of desirable properties such as: population size independence, continuity, scale and metric independence, and boundedness. For a resource allocation game of N users, such that the n^{th} user receives an (expected) allocation of $w_n \geq 0$ resources, the Jain index is given by Equation 2. This equation measures the equality of allocation $\mathbf{w} = (w_1, \dots, w_N)^\top$. An allocation is considered fair, iff $\mathbb{J}(\mathbf{w}) = 1$.

$$\mathbb{J}(\mathbf{w}) = \frac{\left|\sum_{n \in \mathcal{N}} w_n\right|^2}{N \sum_{n \in \mathcal{N}} w_n^2} \quad (2)$$

Resource allocation games often admit undesirable equilibria; asymmetric pure Nash equilibria (PNE) which are efficient but not fair, or symmetric mixed-strategy Nash equilibria (MNE) which are fair but not efficient. For example, the set of asymmetric PNE corresponds to R agents accessing while $N - R$ yield. This results to 100% efficiency, but $\mathbb{J}_{PNE}(\mathbf{w}) = \frac{R^2}{NR} = \frac{R}{N}$. In the symmetric MNE, each agent decides to access with probability $Pr[\mathcal{A} \setminus \{Y\}] = \min\left\{R\left(1 - \sqrt[N-1]{\frac{|\zeta|}{1+|\zeta|}}\right), 1\right\}$ and then chooses which resource to access uniformly at random (Cigler 2013)). The latter results to expected $\mathbb{J}_{MNE}(\mathbf{w}) = 1$, but 0% expected efficiency (assuming small number of resources, R). As such, the aforementioned equilibria are rather undesirable. We can overcome the previously mentioned drawbacks using the notion of correlated equilibria (Aumann 1974).

Correlated equilibria (CE) are a superset of Nash equilibria. They allow for dependencies amongst the the agents' probability distributions, thus the optimization takes place on the joint action space. Correlated equilibria are desired solution concepts in resource allocation games, as they allow for efficient and fair solutions by avoiding positive probability mass on less desirable outcomes. Moreover, an optimal correlated equilibrium for resource allocation games may be found in polynomial time (Papadimitriou and Roughgarden 2008). Subsequently, 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. In a multi-agent scenario we are interested in agents who are able to *learn*; adapt their strategies and converge to an equilibrium. In order to be able to reach richer solution concepts, like correlated equilibria, the agents require a common signal upon which they can learn to anti-coordinate their actions. Hence the introduction of the environmental context, proposed in Section 2.1.

3 Overview of Learning Approaches

In this section we will outline potential multi-agent learning approaches for tackling the anti-coordination problem. We will examine bandit algorithms, reinforcement learning algorithms, and finally, symmetric equilibria for the repeated resource allocation game. We will focus on bimatrix (2-agents, 1-resource) games since, in spite of their simple form, they present many challenges in multi-agent learning scenarios (Littman and Stone 2002).

3.1 Ad-hoc Coordination & Multi-armed Bandit Algorithms

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 (Stone et al. 2010). Typical scenarios include the use of Monte Carlo algorithms (Barrett et al. 2017), Bayesian learning (Albrecht, Crandall, and Ramamoorthy 2016), or bandit algorithms (Chakraborty et al. 2017), (Barrett and Stone 2011). Traditionally, ad-hoc approaches suffer from slow learning, which makes ad-hoc coordination a very ambitious goal for real-life applications. Due to their ability to learn from partial feedback, bandit algorithms would be the natural choice for solving the anti-coordination problem in an ad-hoc setting.

In multi-armed bandit problems an agent is given a number of arms 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. Additionally, they satisfy incentive constraints for rational agents since they constitute an approximate correlated or coarse correlated equilibrium (Nisan et al. 2007). 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 which might be a desirable option for minimizing regret, but not in respect to fairness.

To better understand these limitations, we evaluate three state-of-the-art, well established adversarial bandit algorithms, namely the EXP3 (Auer et al. 2002), the EXP4 (Auer et al. 2002), and the EXP4.P (Beygelzimer et al. 2011). The last two belong to a variant of multi-armed bandits, called contextual bandits[2], that is, at each time-step t, they can exploit the observed context $k_t \in \mathcal{K}$ before making their decision. As such, the chosen arm can be different depending on the context. Moreover, the EXP4.P combines the confidence bounds of UCB1 (Auer, Cesa-Bianchi, and Fischer 2002) with the EXP4 to achieve the same regret as EXP4 but with high probability. Figure 3 depicts the total utilization of resources for the 1-resource, 2-agents allocation game of Figure 2. The x-axis is in logarithmic scale, and the reported values are the average over 128 runs of the same simulation. The input parameters for the EXP family of algorithms are set to their optimal values, as prescribed in (Auer et al. 2002), and (Beygelzimer et al. 2011), assuming time horizon of $T = 10^5$ time-steps[3]. As depicted, all of the evaluated algorithms take a significant number of time-steps to reach a high utilization state, never achieve 100% efficiency, and exhibit high variance.

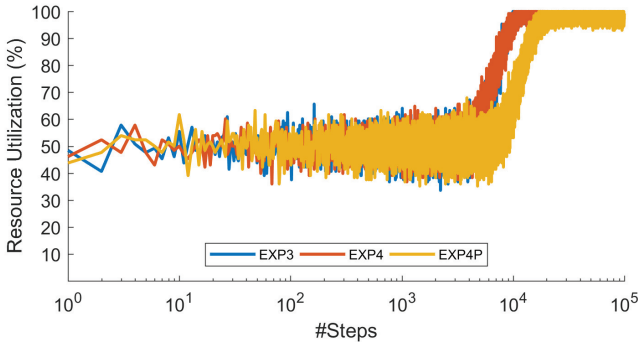

Figure 3: Resource utilization over time achieved by the employed bandit algorithms in the 1-resource, 2-agents allocation game of Figure 2 (x-axis in logarithmic scale).

Along with efficiency, we are interested in the fairness of the final outcome. Being able to achieve both is of the utmost importance for the adoption of such learning paradigms in real-world applications. The evaluated bandit algorithms exhibit considerably low fairness, specifically: $\mathbb{J}_{EXP3}(\mathbf{w}) = 0.50$, $\mathbb{J}_{EXP4}(\mathbf{w}) = 0.76$, $\mathbb{J}_{EXP4.P}(\mathbf{w}) = 0.73$. As a matter of fact, EXP3's achieved fairness is equal to that of an unfair asymmetric PNE: $\mathbb{J}_{PNE}(\mathbf{w}) = \frac{R}{N} = 0.5 = \mathbb{J}_{EXP3}(\mathbf{w})$. The contextual bandits performed somewhat better but, considering the simplicity of the evaluated example, not good enough. This leads to suggest that the evaluated contextual bandit algorithms are unable to handle the large policy space of anti-coordination games.

3.2 Reinforcement Learning & Replicator Dynamics

Closely related to the bandit algorithms of Section 3.1 is reinforcement learning. Reinforcement learning is based on the concept of learning through the interactions with the environment. An agent takes an action, observes some feedback from the environment, and updates his policy so as to maximize some notion of cumulative reward. The most eminent example of such an algorithm is Q-learning (Watkins

[2]See (Zhou 2015) for a survey on contextual bandits.

[3]Note the high sensitivity to the input parameter ($\gamma \in (0, 1]$), which is another crucial shortcoming of the studied bandit algorithms in ad-hoc scenarios.

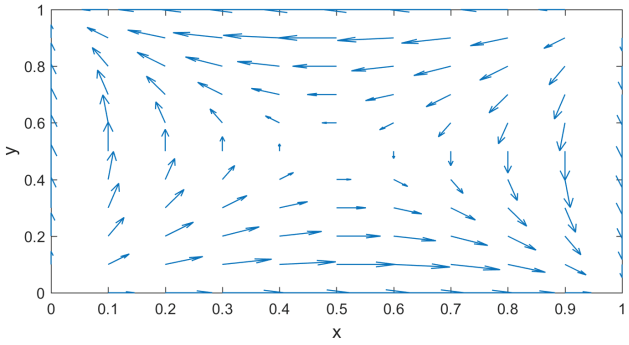

Figure 4: The replicator dynamics, plotted in the unit simplex, for the 1-resource, 2-agents allocation game of Figure 2. x denotes the first agent's probability of playing the first action (Y), while y denotes the second agent's probability of playing the first action (Y). The probabilities of playing the second action (A) are $1-x$ and $1-y$ respectively.

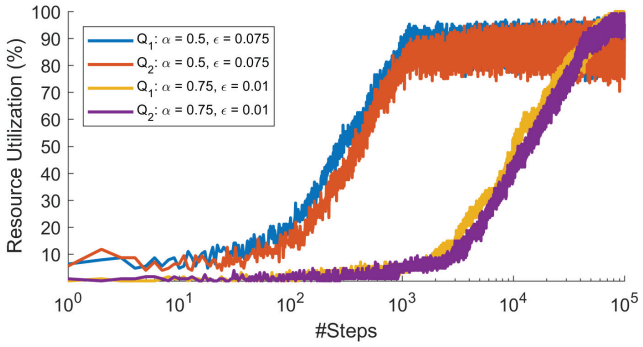

Figure 5: Resource utilization over time achieved by Q-learning in the 1-resource, 2-agents allocation game of Figure 2 (x-axis in logarithmic scale).

and Dayan 1992) which solves Bellman's optimality equation (Bellman 2013) using an iterative approximation procedure. A detailed taxonomy of multi-agent reinforcement learning algorithms can be found in (Busoniu, Babuska, and De Schutter 2008).

There is a formal relationship between reinforcement learning and the replicator dynamics of evolutionary game theory (Bloembergen et al. 2015), hence reinforcement learning algorithms can satisfy our incentive constraints. Evolutionary game theory (EGT)[4] differs from classical game theory in that it focuses on the dynamics of the learning process (strategy change). In a multi-agent system in which agents adapt their behavior in response to strategic interactions with other agents, evolutionary game theory provides a solid mechanism to analyze and understand it (Tuyls and Parsons 2007). Evolutionary game theory is built around the replicator equations:

$$\dot{x_i} = x_i \left[f_i(\mathbf{x}) - \phi(\mathbf{x}) \right] \tag{3}$$

Equation 3 describes the evolution of a population ($\mathbf{x}$) of individuals (x_i) over time, or alternatively (and more befitting to multi-agent learning), the evolution of an agent's strategy $\mathbf{x} = (x_1, \ldots, x_R)^\top$. In the latter interpretation, the population share of each type ($x_i : 0 \leq x_i \leq 1, \forall i$) represents the probability of selecting action a_i, $f_i(\mathbf{x})$ is the fitness (utility) of action a_i, $\phi(\mathbf{x}) = \sum_j x_j f_j(\mathbf{x})$ is the weighted average fitness, and $\dot{x_i} = dx_i/dt$. For the the two agent game of Figure 2, we can rewrite Equation 3 for the strategy vector of the first agent $\mathbf{x}$ as:

$$\dot{x_i} = x_i \left[(U\mathbf{y})_i - \mathbf{x}^\top U \mathbf{y} \right] \tag{4}$$

where U is the payoff matrix (similar for $\mathbf{y}$).

Finding the optimal policy in a multi-agent system where all agents learn simultaneously is inherently more complex. Each agent is faced with a moving-target learning problem.

[4]See (Gintis 2000) for an introduction to EGT.

Changes in the policy of one agent can affect the rewards and thus have a cascading effect on the optimal policies of the others. Furthermore, just as with the bandit algorithms, the adaptation of such dynamics in real-world multi-agent problems requires fairness guarantees. An insight to the quality of the final allocation can be provided by examining the replicator dynamics (Equation 4) of the simple 1-resource, 2-agents allocation game of Figure 2, depicted in Figure 4. As seen by the plot, the two evolutionary stable strategies are the two unfair asymmetric PNE, (Y, A) and (A, Y). Moreover, Figure 5 depicts the total utilization of resources of two Q-learning approaches (the reported values are the average over 128 runs of the same simulation). The Q_1 approach uses the context as its state, while the Q_2 approach uses both the context and the former action as the state. The intuition behind Q_2 is to enable the learning of a possibly more fair multi-step best respond, i.e. investigate the possibility of learning a correlated equilibrium where the two agents alternate between accessing and yielding. The Q table is updated according to Equation 5:

$$Q(s,a) = \alpha(u + \delta \max_{a'} Q(s',a')) + (1-\alpha)Q(s,a) \tag{5}$$

where α is the learning rate, δ the discount factor, and s, s', a, u the state, next state, action, and utility (reward) respectively. Both approaches select their actions according to an ϵ-greedy policy (as in (Littman and Stone 2002)), i.e. in state s, with probability ϵ they choose a random action, while with probability $1-\epsilon$ they take action $\arg\max_a Q(s,a)$. The algorithm's performance is highly sensitive to the aforementioned parameters. We have identified two interesting scenarios, presented in Figure 5. Setting $\alpha = 0.75$ and $\epsilon = 0.01$ results in higher efficiency and lower variance, but lower fairness ($\mathbb{J}_{Q_1}(\mathbf{w}) = 0.64$, and $\mathbb{J}_{Q_2}(\mathbf{w}) = 0.83$). On the other hand, $\alpha = 0.5$ and $\epsilon = 0.075$ results in lower efficiency and higher variance (due to the increased randomness), but higher fairness ($\mathbb{J}_{Q_1}(\mathbf{w}) = 0.82$, and $\mathbb{J}_{Q_2}(\mathbf{w}) = 0.89$). The above are true for both approaches (Q_1, and Q_2).

The aforementioned results of Figure 5 suggest that by incorporating a larger state space (i.e. using the common context and the former action) we can achieve better results than

the replicator dynamics indicated. Given a broad enough state space, Q-learning can learn a multi step best response (Littman and Stone 2002). Nevertheless, in both cases, both approaches require a significant number of time-steps to reach a high utilization state. As such, reinforcement learning in anti-coordination scenarios faces similar shortcomings as bandit algorithms, albeit it seems to achieve higher fairness in the evaluated example. Furthermore, it is worth noting that basic reinforcement learning algorithms like Q-learning, compute quantity values for each possible state or state-action pair. As mentioned, mapping anti-coordination to the consensus problem results in an exponential expansion of the solution space, thus in an exponential increase of the computational and memory complexity for the reinforcement learning algorithms as well. The latter constitutes such approaches infeasible for real-world applications.

Instantiations of a correlated equilibrium can be achieved via reinforcement learning. One example is Correlated Q-learning (Greenwald, Hall, and Serrano 2003), albeit it requires the sharing of Q-tables amongst the agents. The latter necessitates either to allow full observability, or a central planner, neither of which is feasible in ah-hoc scenarios.

3.3 Symmetric Strategies & The Price of Anonymity

The two agent resource allocation game of Figure 2 is an inherently symmetric game, yet the only efficient Nash equilibria are asymmetric; one agent yields while the other accesses, achieving 100% efficiency. Asymmetric equilibria of symmetric games are undesirable for two reasons. First, they are unfair and second they require possibly identical agents to differentiate their actions (and thus learning rules). The symmetric MNE (access with probability $\frac{1}{|\mathcal{C}|+1}$) on the other hand achieves 0% efficiency. The Price of Anonymity (Cigler and Faltings 2014) allows us to measure the degradation of the system's efficiency (social welfare) due to the requirement of symmetry imposed by anonymity. In an anonymous game agents do not distinguish between other agents, i.e. agents have different utilities but an agent's utility depends only on its own strategy and the number of other agents that chose the same strategy, and not on their identities (Nisan et al. 2007). The Price of Anonymity is the ratio between the optimal social payoff of any (possibly asymmetric) equilibrium and the expected social payoff of the worst symmetric equilibrium. In this example, the price of anonymity is infinite. Nevertheless, it is possible to have solution concepts that are symmetric and efficient by making use of correlated equilibria (Aumann 1974).

Cigler and Faltings developed a symmetric learning rule for reaching an efficient and fair correlated equilibrium of the repeated resource allocation game (Cigler and Faltings 2013). By exploiting the history of their interactions along with the environmental context as a correlation mechanism, the agents are able to learn to coordinate their accesses. Each agent n has a strategy $g_n : \mathcal{K} \rightarrow \{0\} \cup \mathcal{R}$ which maps context to resources. As the algorithm progresses, agents who have successfully accessed a resource ($u_n(a_n, a_{-n}) = 1$) for a given context value $k \in \mathcal{K}$ will continue to access the same resource every time they observe the same context k. Agents who have not accessed a resource for a given context value k will not attempt to access an occupied resource. If there is a collision, the colliding parties will back-off with probability $p_n^{backoff}$. Algorithm 1 provides the pseudo-code of the allocation algorithm.

Algorithm 1 Pseudo-code of (Cigler and Faltings 2013).

Require: $\forall n \in \mathcal{N}$ initialize g_n u.a.r. in $\mathcal{R}$.
1: Agents observe context $k_t \in \mathcal{K}$.
2: **if** $g_n(k_t) > 0$ **then**
3: Agent n accesses resource $r \leftarrow g_n(k_t)$.
4: **if** Collision(r) **then**
5: Set $g_n(k_t) \leftarrow 0$ with probability $p_n^{backoff}$.
6: **end if**
7: **else if** $g_n(k_t) = 0$ **then**
8: Agent n monitors random resource $r \in \mathcal{R}$.
9: **if** Free(r) **then**
10: Set $g_n(k_t) \leftarrow r$ with probability 1.
11: **end if**
12: **end if**

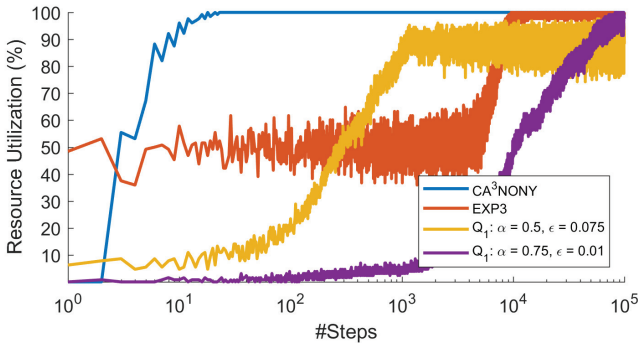

Figure 6: Resource utilization over time of CA^{3}NONY vs. EXP3, Q_1, and Q_2 in the 1-resource, 2-agents allocation game of Figure 2 (x-axis in logarithmic scale).

Algorithm 1 is only applicable in cooperative scenarios. A self-interested agent could stubbornly keep accessing a resource forever, until everyone else backs off (also known as 'bully' strategy (Littman and Stone 2002)[5]). There exist equilibrium back-off probabilities, but in order to actually play them, the agents need to be able to calculate them. It is not always possible to obtain the closed form of the back-off probability distribution of each resource. We have build upon the ideas of (Cigler and Faltings 2013) and proposed instead the adoption of a human-inspired convention of courtesy, which prescribes a constant positive back-off probability in case of collision ($p_n^{backoff} = p > 0, \forall n \in \mathcal{N}$). Coupled with a bookkeeping scheme and punishments for deviating agents, we have proven that adhering to the algorithm is a best-response strategy at each sub-game of the original stage game, given any history of the play. The developed an anti-coordination framework (CA^{3}NONY (Danas-

[5]Such strategies similarly affect Q-learning (Littman and Stone 2002) and bandit algorithms.

sis and Faltings 2018)) still follows to the simple learning rule of Algorithm 1, which allows for fast convergence and its applicability to large scale multi-agent systems.

To verify its performance, Figure 6 depicts the total utilization of resources for the simple 1-resource, 2-agents allocation game of Figure 2, while Figure 7 compares the convergence time of CA^{3}NONY to the fastest of the presented algorithms (EXP3, Q_1, and Q_2) for increasing number of resources R ($N = 2 \times R$). In every case we report the average value over 128 runs of the same simulation. Note that in the first graph, the x-axis is in logarithmic scale, while the second graph is in double logarithmic scale and the error bars represent one standard deviation of uncertainty. For the second simulation (Figure 7), we chose a high enough time horizon ($= 10^8$) to facilitate EXP3 in achieving the convergence criterion ($\geq 90\%$ efficiency) in larger simulations ($R > 64$). Nevertheless, it was unable to do so for $R > 256$, hence the gaps in the EXP3's lines in Figure 7. For the same reason (again regarding Figure 7), we set Q_1 and Q_2's parameters as $\alpha = 0.75$ and $\epsilon = 0.001$. The high learning rate and low randomness were necessary, otherwise Q_1 and Q_2 were unable to reach high utilization. As depicted, CA^{3}NONY is significantly faster than both the bandit and Q-learning algorithms, exhibits lower variance, and can gracefully handle increasing number of resources. In addition to being efficient, CA^{3}NONY converges to a fair allocation $\mathbb{J}_{\text{CA}^3\text{NONY}}(\mathbf{w}) = 1$. Fairness plays an important role, especially in scenarios with scarcity of resources. If the final allocation is fair, rational agents will be more willing to adhere to the protocol and wait for their turn. Under low fairness, the competition between rational agents is increased, which in turn slows down convergence. In Figure 7, the two Q-learning approaches (especially Q_1) might look appealing from the perspective of scalability, but both result in considerably low fairness (lower on average than an unfair PNE). For any number of resources, $\mathbb{J}_{Q_1}(\mathbf{w}) \in [0.45, 0.52]$, with a mean value of 0.48, while $\mathbb{J}_{Q_2}(\mathbf{w}) \in [0.37, 0.48]$, with a mean value of 0.44. Thus, both Q-learning approaches converge to a situation similar to an unfair PNE. In repeated games though, rational agents might not be willing to concede to a PNE (as in the 'bully' strategy of (Littman and Stone 2002)). Finally, CA^{3}NONY provides higher average payoff for the agents (45.09 for CA^{3}NONY vs. -50.54 for the EXP3, -79.03 for Q_1, and -84.08 for Q_2 in the scenario of Figure 6, assuming collision cost $\zeta = -1$), which is an essential indicator of the algorithms individual performance. The latter constitute CA^{3}NONY a promising framework for real-life applications.

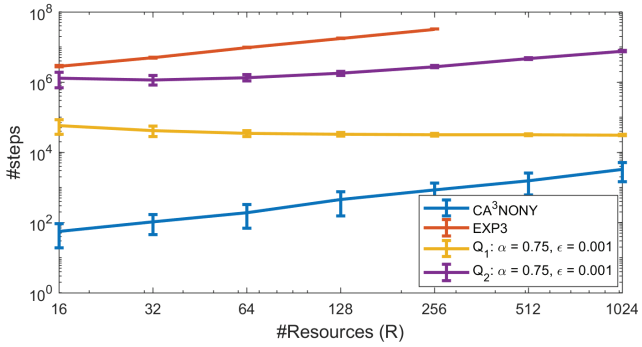

Figure 7: Convergence time of CA^{3}NONY vs. EXP3, Q_1, and Q_2 for increasing number of resources R, $N = 2 \times R$ (double logarithmic scale).

4 Conclusion

The relevance of anti-coordination in multi-agent scenarios stems from the need of sharing (possibly) indivisible, limited resources. The curse of dimensionality encompassing the mapping of anti-coordination problems to the classical consensus problem along with the non-stationarity arisen from the simultaneous learning of all the participants make achieving a desirable outcome even more challenging. Furthermore, contrary to coordination problems which are typically encountered in cooperative settings, anti-coordination deals mostly with self-interested, rational agents. Rational agents are able to manipulate the algorithm to maximize their own utility, which brings forth the need for developing algorithms resilient to such manipulations. Ultimately, anti-coordination boils down to incentivizing participants to systematically and consistently adopt less desirable actions, albeit in a way that ensures high efficiency and fairness in the final outcome.

In this paper, we presented a brief overview of multi-agent learning dynamics for the anti-coordination problem, to increase interest and motivate research in the area. We focused on satisfying incentive constraints, efficiency, fairness and convergence speed. Specifically, we examined bandit algorithms, reinforcement learning, and symmetric strategies for the repeated resource allocation game. We demonstrated that most of the classical, well-established multi-agent learning techniques suffer from slow convergence rate and/or poor fairness. An exception to that is CA^{3}NONY , an anti-coordination framework based on the human-inspired convention of courtesy. Contrary to the aforementioned approaches, CA^{3}NONY is able to reach efficient and fair allocations in polynomial time. Moreover, adhering to the protocol constitutes a rational strategy. The latter suggests that human-inspired conventions may prove beneficial in other ad-hoc coordination scenarios as well. An interesting future direction would be to combine well-established multi-agent learning techniques with simple conventions (e.g. allowing others to acquire a resource first (courtesy convention), or maintaining the acquired resource after convergence) for solving more complex anti-coordination problems.

Finally, a generalization of anti-coordination games, called dispersion games, was described in (Grenager, Powers, and Shoham 2002). 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 (Challet et al. 2013)). In (Grenager, Powers, and Shoham 2002) the agents 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, would be another interest-

ing avenue for future research.

References

Albrecht, S. V.; Crandall, J. W.; and Ramamoorthy, S. 2016. Belief and truth in hypothesised behaviours. *Artificial Intelligence* 235:63–94.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing* 32(1):48–77.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47(2):235–256.

Aumann, R. J. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1(1):67 – 96.

Barrett, S., and Stone, P. 2011. Ad hoc teamwork modeled with multi-armed bandits: An extension to discounted infinite rewards. In *Proceedings of 2011 AAMAS Workshop on Adaptive and Learning Agents*, 9–14.

Barrett, S.; Rosenfeld, A.; Kraus, S.; and Stone, P. 2017. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence* 242:132–171.

Bellman, R. 2013. *Dynamic programming*. Courier Corporation.

Beygelzimer, A.; Langford, J.; Li, L.; Reyzin, L.; and Schapire, R. 2011. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 19–26.

Bloembergen, D.; Tuyls, K.; Hennes, D.; and Kaisers, M. 2015. Evolutionary dynamics of multi-agent learning: A survey. *J. Artif. Int. Res.* 53(1):659–697.

Busoniu, L.; Babuska, R.; and De Schutter, B. 2008. A comprehensive survey of multiagent reinforcement learning. *Trans. Sys. Man Cyber Part C* 38(2):156–172.

Chakraborty, M.; Chua, K. Y. P.; Das, S.; and Juba, B. 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.

Challet, D.; Marsili, M.; Zhang, Y.-C.; et al. 2013. Minority games: interacting agents in financial markets. *OUP Catalogue*.

Cigler, L., and Faltings, B. 2013. Decentralized anti-coordination through multi-agent learning. *Journal of Artificial Intelligence Research* 47:441–473.

Cigler, L., and Faltings, B. 2014. Symmetric subgame-perfect equilibria in resource allocation. *J. Artif. Int. Res.* 49(1):323–361.

Cigler, L. 2013. *Multi-Agent Learning for Resource Allocation Problems*. Ph.D. Dissertation, École Polytechnique Fédérale de Lausanne.

Coulouris, G. F.; Dollimore, J.; and Kindberg, T. 2005. *Distributed systems: concepts and design*. pearson education.

Danassis, P., and Faltings, B. 2018. A courteous learning rule for ad-hoc anti-coordination. *arXiv:1801.07140*.

Gintis, H. 2000. *Game theory evolving: A problem-centered introduction to modeling strategic behavior*. Princeton university press.

Greenwald, A.; Hall, K.; and Serrano, R. 2003. Correlated q-learning. In *ICML*, volume 3, 242–249.

Grenager, T.; Powers, R.; and Shoham, Y. 2002. Dispersion games: general definitions and some specific learning results. In *AAAI/IAAI*, 398–403.

Jain, R.; Chiu, D.; and Hawe, W. 1998. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. *CoRR* cs.NI/9809099.

Littman, M. L., and Stone, P. 2002. *Implicit Negotiation in Repeated Games*. Berlin, Heidelberg: Springer Berlin Heidelberg. 393–404.

Nisan, N.; Roughgarden, T.; Tardos, E.; and Vazirani, V. V. 2007. *Algorithmic game theory*, volume 1. Cambridge University Press Cambridge.

Papadimitriou, C. H., and Roughgarden, T. 2008. Computing correlated equilibria in multi-player games. *J. ACM* 55(3):14:1–14:29.

Stone, P.; Kaminka, G. A.; Kraus, S.; and Rosenschein, J. S. 2010. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence*.

Tuyls, K., and Parsons, S. 2007. What evolutionary game theory tells us about multiagent learning. *Artificial Intelligence* 171(7):406 – 416. Foundations of Multi-Agent Learning.

Watkins, C. J. C. H., and Dayan, P. 1992. Q-learning. *Machine Learning* 8(3):279–292.

Zhou, L. 2015. A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326*.