A Learning Approach to Dynamic Coalition Formation

Enrique Munoz de Cote
Politecnico di Milano
Department of Electronics and Information
piazza Leonardo da Vinci 32, I-20133 Milan, Italy
munoz@elet.polimi.it

Alessandro Lazaric
Politecnico di Milano
Department of Electronics and Information
piazza Leonardo da Vinci 32, I-20133 Milan, Italy
lazaric@elet.polimi.it

Marcello Restelli
Politecnico di Milano
Department of Electronics and Information
piazza Leonardo da Vinci 32, I-20133 Milan, Italy
restelli@elet.polimi.it

ABSTRACT

We consider the problem where self-interested learning agents need to cooperate with others in order to achieve better outcomes. Cooperative game theory studies the problem of coalition formation and provides a reliable way to sustain equilibrium points inside the Pareto region for "large" groups of individuals. However, research in this field provides one-shot solutions that in real scenarios gives us a partial and static view of the problem. In this paper we present a transformation from the static to the dynamical view using learning techniques. We study the conditions in which the learning dynamics of a "large" population of Q-learning agents can evolve and learn policies that will settle into stable coalition structures. We then present how this learning dynamics relate to game theoretical stability concepts.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Concept Learning; I.2.11 [**Distributed Artificial Intelligence**]: Multiagent Systems

General Terms

Algorithms, Experimentation

Keywords

Multi Agent Learning, Cooperative Game Theory, Dynamic Coalition Formation

1. INTRODUCTION

Automated negotiations among *self-interested* agents are becoming increasingly important due to the rapid changes in technology. Nowadays, separately designed agents belonging to different organizations can interact in an open environment and safely carry out transactions. This kind of scenarios could benefit if autonomous self-interested agents could form "temporal agreements" to profit from other agents' resources, capabilities and experience. One reliable way of sustaining this kind of cooperation among self-interested agents in a *multi-agent system* (MAS) can emerge if agents can fraction/ensemble themselves into coalitions. Coalition formation (CF) rely on *cooperative game theory* (CGT) [1] as the mathematical framework for finding resulting steady states of games in coalitional form. However, the concepts it addresses are static by definition and therefore, the theory fails at explaining how agents arrive at certain coalitions and at equilibrium in general.

From the system designers viewpoint, one may concern for the stable partitions of the MAS (coalition configurations), and the final outcomes resulting from such configurations. *Dynamic coalition formation* [12, 13, 5] was born as an extension of CGT to fill the gap between the static and the dynamic worlds. It bases its assumptions on agents that follow simple adaptation rules, which are based on myopic optimization. This approach lacks on methods and algorithms to compute solutions, however, it gives insights to questions like: how do coalitions form, and how do agents decide on the division of the coalitional gains.

In recent years, research in multi-agent systems has concentrated on the problem of coalition structure (CS) generation, an activity previously brushed aside in CGT but of great importance in real scenarios. The biggest limitation in this field is the complexity of searching through an exponential search space, which is NP-complete [8]. There are several solutions known from prior art, and, given that the number of possible coalitions is exponential in the number of agents (2^n), the mainstreams for reducing this number are heuristic search algorithms like greedy search [11] or best-first search [4, 15, 9]. All of these approaches compromise optimality in order to relax complexity, they have been tested and applied in interesting scenarios like task allocation problems [11] and combinatorial optimization problems [9]. Sen and Dutta interested in the problem of identifying the optimal coalition structure, in [10], they propose an order-based genetic algorithm as a stochastic search process to identify the optimal coalition structure. They compare their results to a deterministic CS search algorithm [8] and obtain better results when the optimal CS is complex enough.

Our perspective differs from previous ones in that this work is intended to provide a dynamic and adaptive solution that satisfies both, a stable coalition structure generation process that is completely distributed and a MAS that is reactive to changes in the environment (e.g. an agent signing-in). There has been little work in the art covering this problems. An approach related to ours is presented in

AAMAS'07 May 14–18 2007, Honolulu, Hawai'i, USA.

[2], where a Bayesian RL approach is used to create a model of the other agents in the system. Creating a model of the other agents is a strong demand so it is unclear how scalable this approach can be. We instead do not need a model of the other agents and rather just let the agents experience different coalitions to rate their value, making this approach much more scalable.

A game in coalitional form is usually dealt with as a one-shot game, rather than a sequential process. In a one-shot game all agents are required to simultaneously choose actions, and payoffs are given to coalitions according to the specified characteristic function v. In a real situation, coordination problems may arise due to incompatibility of players' plans (e.g when one agent pretends to join another agent and the former joins some other coalition). To alleviate this inconvenience we study coalition formation as an ongoing, dynamic process, with payoffs generated as coalitions form, disintegrate, or regroup. In our model, only one agent is able to revise its strategy at any time[1]. The motivation for introducing turns is that now the environment has a one-step deterministic dynamics, where an agent i chooses an action and the coalition structure at the next step is completely determined, showing with this a Markov-like one-step dynamics (important property for any TD algorithm). We let an external process (nature's hands) govern the agent selection (for instance, the agent turn may be drawn from a uniform distribution over the population).

The contribution of this paper is twofold, first we study how, using simple reinforcement learning techniques, a system composed by self-interested learning agents can develop social "bonds" when shown beneficial, and second, we analyse whether this approach shows characteristics that fall into some stability definition from game theory. This is a preliminary work in this line; here, we put out the conditions in which a MAS of standard Q-learners are able to find stable coalition configurations that fall into a stability concept known as the *core* [1].

2. COOPERATIVE GAME THEORY

Coalition formation can be addressed under two completely different scenarios. When under a purely cooperative setting, agents can be modeled as *group rational* individuals, where they care to maximize the coalition's utility. In contrast, we study the problem in a competitive setting (e.g. an open environment). In this case the problem becomes non trivial given that agents are modeled as *individually rational*, caring to maximize their individual utility. Here, an agent will join (or quit) a coalition only if it believes it is in its own best interest to do so.

The formal mathematical representation used to study CF is a *characteristic function game* (CFG) [1]. A CFG is usually dealt with as a one-shot game, rather than a sequential process (just as in non-cooperative games).

2.1 Assumptions and Definitions

Formally, let $N = \{1, \ldots, n\}$ be a set of players, a *coalition* (S) is a subset, $S \subseteq N$ such that $S \in 2^n$, and the coalition composed by the entire set N is called *grand coalition*. The partition of N in the set of exhaustive and disjoint coalitions is called a *coalition structure* (CS),

[1]These new dynamics poses no real drawbacks given that in unsynchronized scenarios, agents do not execute actions in a contemporarily manner (unless there is some coordinator).

Definition 1. A coalition structure of a CFG, $CS = \{S_1, \ldots, S_K\}$, is a partition of N. That is, $S_k \neq \emptyset$ for any $k \in \{1, \ldots, K\}, \bigcup_{k=1}^{K} S_k = N$, and $S_k \cap S_l = \emptyset$ for any $k, l \in \{1, \ldots, K\}$ with $k \neq l$.

The value of a coalition is computed by a commonly known characteristic function, $v(S)$, which determines the utility a coalition is able to obtain. These kind of games are called *transferable utility* (TU for short). Since utility is assigned to coalitions in TU games, agents inside the coalitions need to divide the surplus from collaborating. The outcome from dividing the coalitions' surplus is a vector of payoffs $\mathbf{x} = (x_1, \ldots, x_n)$, that should be *feasible*. Let $\mathbb{R}^S$ denote the $|S|$-dimensional Euclidean space with coordinates indexed by elements of S. Given this notation, we may restrict the vector of payoffs $\mathbf{x} = (x_1, \ldots, x_n)$, called an *imputation*[2] to the subset $\mathbf{x}_S$, which will denote its restriction to $\mathbb{R}^S$. Feasibility implies that for any S there exists a real number $v(S) \geq 0$ which creates a set of vectors $\{\mathbf{x}_S : \sum_{i \in S} x_i \leq v(S)\}$ that is feasible.

Most works in literature have been focused on games characterized by a function $v(S)$ that is increasing in the size of S. This class of games are named *superadditive*, where for any two distinct coalitions S and T it holds that $v(S) + v(T) \leq v(S \cup T)$. A non-increasing to scale characteristic function is a function v where the previous assumption does not hold. This type of function v defines a *non-superadditive CFG*. In what follows we will refer to a non-superadditive CFG just as CFG.

Let $\mathcal{P}(S)$ denote the set of all possible coalitions. Any population defined by N generates a characteristic function defined over the set $\mathcal{P}(S)$; such characteristic function and the set N define a *coalition game* [1].

Definition 2. A coalition game is a pair (N, v) such that,
(i) v is a mapping $\mathcal{P}(S) \backslash \{\emptyset\} \mapsto \mathbb{R}$, called characteristic function;
(ii) $S \subseteq N$ such that $S \in 2^n$.

The meaning of stability in a CFG basically depends on two things: (a) the way the surplus $v(S)$ is divided among the coalition participants and (b) the type of rationality the participants hold. When rational agents seek to maximize their individual payoffs, stability becomes critical. The equilibrium concept chosen for this model entirely relies in a game theoretic concept for payoff division within coalitions, known as *the core*.

Definition 3. Let $\mathcal{P}(CS)$ denote the set of all possible coalition structures.

The core of a non-superadditive CFG (v, N), is the set of vectors $\mathbf{x} = (x_1, \ldots, x_n)$ such that,
(i) $\sum_{i \in N} x_i = \max_{CS \in \mathcal{P}(CS)} \sum_{S \in \mathcal{P}(S)} v(S)$
(ii) $\sum_{i \in S} x_i \geq v(S),\ \forall S \in \mathcal{P}(S)$

The issue of assigning an instantaneous reward to each individual agent is a hard problem, that is because CFG assigns a value to each coalition and not to each individual agent. As stated earlier, letting agents divide the surplus from the coalition's value is not an easy task and we are not interested in building a complex solution concept based on payoff distribution methods. Instead, we will take a solution based on the *marginal contribution*,

[2]Also referred as *allocation* in literature.

Definition 4. For an agent i, the marginal contribution is defined to be, $\vartheta_i = v(S) - v(S\backslash\{i\})$ where $v(S\backslash\{i\})$ is just the restriction of S to $S\backslash\{i\}$.

Therefore, the marginal contribution measures the level of participation of an agent to the coalitional gains (note that in a non-superadditive game, a coalition may also perceive a loss with the inclusion of an agent).

Seminal works on dynamic coalition formation literature follow some standard assumptions: payoffs can be transfered between agents belonging to the same coalition, $v(S)+v(T) \leq v(S \cup T)$ holds (superadditivity condition) and agents follow simple adaptation rules which are based on myopic optimization. This assumptions are used to study the properties of the dynamical process of forming stable coalition structures. Our motivations, on the other hand, are different in the sense that this approach is agent design centered, meaning that agents are bounded rational players[3], are self-interested utility maximizers and act based upon an expectation about future (discounted) rewards. Therefore, our assumptions overlap with the standard ones but are not equal:

- Communication is not allowed.
- Information about each agent's payoff function and preferences is not available.
- Payoffs are transferable between agents belonging to the same coalition.
- Superadditivity does not hold.

2.2 Complexity in finding an optimal solution

The problem of finding a coalition structure that maximizes the population gains is computationally complex. The input to any coalition structure search algorithm involves knowing the value $v(S)$ of each coalition $\forall S \in 2^n$ (exponential in the number of agents), and finding the coalition structure that maximizes the social welfare,

$$CS^* = \arg \max_{CS \in \mathcal{P}(S)} v(CS) \tag{1}$$

where $v(CS) = \sum_{S \in CS} v(S)$. The exact number of coalition structures is a function $O(n^n)$, and therefore exhaustive enumeration is not a viable method for searching the optimal coalition structure [8].

In the next section we redefine the concepts previously presented in order to formalize the model of coalition formation with the inclusion of types.

2.3 Cooperative game theory with types

This section contains a characterization of an agent economy with a finite number of types of agents with effective small groups.

In this new perspective, there is a finite number of agent types $\{1, \ldots, n\}$, such that any player in this economy is completely specified by its type, which is fixed for all times. A coalition can be identified with a point $\mathbf{S} \in \mathbb{R}^n$ such that $0 \leq \mathbf{S} \leq \mathbf{Q}$, where $\mathbf{Q} \in \mathbb{R}^n_+$ specifies the total number of players of each type (n is the number of types).

Let $\mathcal{P}(\mathbf{S})$ denote the set of feasible coalitions. Any population defined by the vector $\mathbf{Q} \in \mathbb{R}^n_+$ generates a characteristic function defined over the set $\mathcal{P}(\mathbf{S}) = \{\mathbf{S} \in \mathbb{R}^n; 0 \leq \mathbf{S} \leq \mathbf{Q}\}$ A CFG with types is defined by a pair $(\mathbf{Q}, v)$ such that, v is a mapping $v : \mathcal{P}(\mathbf{S})\backslash\{\emptyset\} \mapsto \mathbb{R}$, with $\mathbf{Q}$ interpreted as the grand coalition.

[3]Agents have limited computational resources.

Definition 5. A coalition structure of a CFG with types, $CS = \{\mathbf{S}_1, \ldots, \mathbf{S}_K\}$, is a partition of $\mathbf{Q}$. That is, $\mathbf{S}_k \neq \emptyset$ for any $k \in \{1, \ldots, K\}, \bigcup_{k=1}^{K} \mathbf{S}_k = \mathbf{Q}$, and $\mathbf{S}_k \cap \mathbf{S}_l = \emptyset$ for any $k, l \in \{1, \ldots, K\}$ with $k \neq l$.

Definition 6. Let $\mathcal{P}(CS)$ denote the set of all possible coalition structures. The *core* of a CFG with types $(v, \mathbf{Q})$, is the set of vectors $\mathbf{x} = (x_1, \ldots, x_n)$ such that,
(i) $\mathbf{xQ} = \max_{CS \in \mathcal{P}(CS)} \sum_{\mathbf{S} \in \mathcal{P}(\mathbf{S})} v(\mathbf{S})$
(ii) $\mathbf{xS} \geq v(\mathbf{S}),\ \forall \mathbf{S} \in \mathcal{P}(\mathbf{S})$

Intuitively, the core of a game w.r.t. a given CS is the set of all feasible payoffs that cannot be blocked by any coalition, such that, when under a core allocation, no subgroup of agents is motivated to depart from the given structure. Assumption *(ii)* from Definition 6 state the condition that if $\mathbf{x}$ is a core allocation then no coalition value will surpass the sum of core allocations for that coalition, this (no blocking coalition) condition is necessary for a stable coalition to exist. Also note that $\max_{CS \in \mathcal{P}(CS)} \sum_{\mathbf{S} \in \mathcal{P}(\mathbf{S})} v(\mathbf{S})$ is equal to $v(\mathbf{Q})$ if v is superadditive, returning to the definition of the superadditive core.

3. REINFORCEMENT LEARNING

In RL, the goal of an agent is formalized in terms of the reward signal r, where the agents' objective is to maximize expected return. The expected return is formally represented in terms of an *objective function*, which will become the function for the agent to maximize. Formally, an agent's objective function is,

$$E_\pi \left\{ \sum_{j=0}^{\infty} \gamma^j r_{t+j} \mid s \in \mathcal{S} \right\} \tag{2}$$

where r_{t+j} is the reward received j steps into the future given that the agent follows policy π and $\mathcal{S}$ is the agent state space. A policy, π, is a mapping that defines the probability of selecting a coalition from a particular state.

Farsightedness is governed by a discount factor, $0 \leq \gamma < 1$ that controls how much effect future rewards have on the optimal decisions, with small values of γ emphasizing near-term gain and larger values giving significant weight to later rewards. At any time, an agent is able to estimate the value of a state, by using (2),

$$V^\pi(s) = E_\pi \left\{ \sum_{j=0}^{\infty} \gamma^j r_{t+j} \mid s \in \mathcal{S} \right\} \tag{3}$$

which estimates the farsighted value of the present s^t when using policy π. Similar to (3) we can define an state-action value as,

$$Q^\pi(s, a) = E_\pi \left\{ \sum_{j=0}^{\infty} \gamma^j r_{t+j} \mid s \in \mathcal{S}, a \in \mathcal{A}) \right\} \tag{4}$$

which estimates the farsighted value of taking action $a^t \in s^t$ when using policy π.

What is left is to express the relationship between the Q-value of the present state s^t and its possible successor states.

This is capture by the *Bellman equation* for Q^π,

$$Q^\pi(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s,a,s') V^\pi(s') \qquad (5)$$

Q-Learning is one of the most used learning algorithms both in single and multiagent problems. In its basic form it updates it table of Q-values according to the following rule,

$$Q(s,a) \leftarrow (1-\alpha) \cdot Q(s,a) + \alpha \cdot \left(r + \gamma \max_{a' \in \mathcal{A}} Q(s',a') \right) \quad (6)$$

where α is the learning rate. In order to explore all the states and action spaces, at each time step agent i plays an ϵ-greedy strategy that takes her best action $a = \arg\max_{a'} Q_i(s,a)$ with probability $1-\epsilon$ and one random action with probability ϵ.

4. LEARNING TO FORM COALITIONS

Given that the focus of this approach is on the dynamics of coalition formation and on its convergence to stable coalition structures, we will present a smooth transition form the original one-shot model to the dynamical perspective of coalition formation. We study coalition formation as an ongoing, dynamic process, with payoffs generated as coalitions form, disintegrate, or regroup. The consequences of small group effectiveness and the proportion of agents of each type are examined within this learning framework.

A CFG is usually dealt with as a one-shot game, rather than a sequential process. In a one-shot game all agents are required to simultaneously choose actions, and payoffs are given to coalitions according to the specified characteristic function v. Under this perspective, GT asks if under a given CS there exists an allocation vector $\mathbf{x}$ such that no agent has an incentive to deviate from its present coalition. In a real situation, coordination problems may arise due to incompatibility of players' plans[4].

A dynamical coalition formation process can be seen as being composed of the following activities: (a) the search for an optimal coalition structure; (b) the solution of a joint problem facing members of each coalition; and (c) division of the value of the generated solution among the coalition members. Point (a) is the only interest in this work, and to avoid point (b), we suppose the coalition knows how to best solve its joint problem. Point (c) is an intrinsic part of the coalition formation stability and thus cannot be obviated, it is this distribution of the coalition's profit to its members that ensures individual rational payoffs. This is what provides a minimum of incentive to the agents to collaborate. We will come to this issue in the proposed solution (section 4).

The first intuitive direction to port this one-shot CFG into a dynamic perspective suitable for learning is to allow all agents choose contemporarily their actions.

4.1 Simultaneous action selection model

We assume that each agent can always leave its current coalition and join the coalition of its choice, but no agent can be forced to stay in any coalition (it would be against the rationality principle to force any agent's action).

[4]Imagine that at time t player i is in the singleton coalition $\{i\}$ but plans to join the coalition $\mathbf{S}$ at time $t+1$, while another player j plans to join $\{i\}$ in $t+1$. If both agents get to perform their actions simultaneously, agent j will find itself with an unavailable move.

The formal description of the dynamical model goes as follows. At time t, the state of the environment is the coalition structure CS^t, which is fully observable for all agents in the system. At any time, any agent has the possibility to revise her strategy, by changing her current coalition to any $\mathbf{S}^t \in CS^t \cup \{\emptyset\}$, including forming the singleton coalition $\{i\}$. Clearly the agent's choice of coalition membership is restricted by the current coalition structure.

Given a coalition structure CS, the set of actions available to any agent i is,

$$A_i(CS) := \{\mathbf{S}_i | \mathbf{S}_i = \mathbf{S} \cup \{i\},\ \forall\ \mathbf{S} \in CS \cup \{\emptyset\}\} \qquad (7)$$

and a joint action is defined as $A = \times_{i \in \mathbf{Q}} A_i(CS)$.

Under this approach, at any time t action selection is performed in parallel by all the agents. At this point a CS^{t+1} is achieved. According to the principle presented earlier, if an agent i chooses an action $\mathbf{S}_i^t$ at time t and $\mathbf{S}_i^t \neq \mathbf{S}_i^{t+1}$ (meaning that the coalition $\mathbf{S}_i^t$ not longer exists), it will end up in the singleton coalition $\{i\}$ (recall footnote 4).

One may reason that even if the original coalition $\mathbf{S}_i^t$ has changed at time $t+1$, agent i can still join the remaining agents. Even though it is completely true, if we allow this to happen, the agent i would not be longer choosing the action that she wanted, making this action unpredictable, thus preventing the agent from correctly evaluate the value of the state-action pair $Q_i(CS^t, \mathbf{S^t})$ during the learning process.

Once the convention for the dynamics of the model is given, the state transitions and the reward assignment can be computed.

- $T(CS^t, \mathbf{S}^t) \rightarrow CS^{t+1}$. The state transition is a deterministic function of the present state and joint action.
- $r_i = \vartheta_i$. The reward r_i can be determined after making a move $A_i(CS)$ (this in short means: changing coalition, keeping the same coalition or forming the singleton coalition) by following the *marginal contribution* principle.

At each time point, agent i chooses an action a at state s[5]. The action value function is updated with the update formula (6), this happens contemporarily for all agents, so the ending state s', will depend on the joint action.

We now propose a *turn dynamics* where only one agent is able to revise its strategy at any time. This new dynamics poses no real drawbacks given that, in unsynchronized scenarios, agents do not execute actions in a contemporarily manner (unless there is some coordinator). We let an external process govern the agent selection. We will first use a round-robin agent selection process to extract agents from the population set in a fixed manner and present the dynamics of the coalition formation process. We will then expand this notions using a uniform distribution to extract agents at random.

The motivation for introducing turns is that now the environment has a one-step deterministic dynamics, where an agent i chooses an action $\mathbf{S}^t \in CS^t$ and the coalition structure CS^{t+1} at the next step is completely determined, showing with this a Markov-like (i.e., memoryless) property.

4.2 Fixed turn action selection model

[5]For ease of notation we write $s = CS$, $a = \mathbf{S}$

By introducing the concept of a turn, the type of game changes and the correct representation for this type of dynamics are *extensive form games*. The model of an extensive form game, by contrast to the strategic one, describes the sequential structure of decision making explicitly, allowing the study of situations in which each agent is free to change her mind as events unfold [3]. In order to apply learning algorithms to extensive form we adopt the more general framework of stochastic games. In a stochastic game, each agent's reward depends on the current state, while state transitions obey the Markov property.

The agent that acts at time t is given by the turn function $\iota : \mathbb{N} \rightarrow \mathcal{N}$ where ι is such that $\iota(t) \neq \iota(t+1)$.

This translation is admissible every time the dynamics of the turn-based game shows a Markov-like (memoryless) property[6].

Note that in this case T is the fixed period of inactivity $T =\mid N-1 \mid$. The process starts at time $t = 0$, where an agent a_0 is drawn from a uniform distribution over the population, at this time, an order is fixed for all times, starting from agent a_0 and ending each *epoch* after $\mid N-1 \mid$ turns.

Let $T =\mid N-1 \mid$ and $\iota(t+T) = i$, then, at time $t+T$ agent i chooses an action a^{t+T} at state s^{t+T}. The action value function is updated with the following update formula [14]:

$$Q_i(s^t, a^t) = (1-\alpha)Q_i(s^t, a^t) + \alpha(\mathcal{R}_i + \gamma \max_{a^{t+T}} Q_i(s^{t+T}, a^{t+T}))$$

where α is the learning rate, $0 \leq \gamma < 1$ is a discount factor.

Due by the turn dynamics, agents do not get the opportunity for action selection at all time steps, therefore the appropriate notion of reward function is the accumulative sum of discounted rewards,

$$R_i = \sum_{j=0}^{T} \gamma^j r_{t+j+1} \tag{8}$$

where r_{t+j} is the reward received j steps into the future.

In the next subsection we present the randomized model dynamics of coalition formation, which only defers from this by the function ι.

4.3 Randomized turn action selection model

This is a generalization of the model presented in section 4.2. The agent that acts at time t is given by the agent function $\iota : \mathbb{N} \rightarrow \mathcal{N}$ where ι is a random function in the entire population.

The reward function $\mathcal{R}_i$ is now the reward accumulated over a variable number T of $\{pass\}$ actions. The expected value of T is still $|N-1|$ since agents are selected according to a uniform distribution probability over the whole population.The updating rule is exactly as presented in 4.2, so the agent in turn update its past state using the accumulated $\mathcal{R}_i$.

From a learning viewpoint, using a randomized turn is translated in a more complete exploration of the state space. Therefore, the learning curve is expected to be smoother due by the fact that the agents are having a more complete experience. Furthermore, given that random turn selection is a generalization of the previous cases, this randomness can be thought of as a more general learning framework, therefore if agents are able to learn in this setting it is most probable that it will in other setting with simpler dynamics.

The two basic properties present in every agent in our model, individual rationality and farsightedness guarantee that the agents will care about the "ultimate" payoff from a move, and not its immediate consequences. Therefore, a player will join (or quit) a coalition if and only if he believes it is in his own best interest to do so. This creates an individual rational distribution that assigns to each agent at least the gain it may get without collaborating within any coalition,

Definition 7. Individual (farsighted) rationality. An agent moves away from a coalition only if in the future it expects to benefit for moving from its present coalition given the present CS^t. Formally, an agent will move from coalition $\mathbf{S}_i$ if $\exists S_j \in CS^t \cup \{\emptyset\}$ such that,

$$Q(CS^t, \mathbf{S}_j \cup \{i\}) > Q(CS^t, \mathbf{S}_i), \qquad \mathbf{S}_i, \mathbf{S}_j \in CS^t \tag{9}$$

A *best-response* agent's objective is to find a policy π mapping its interaction history to a current choice of action so as to maximize its objective function. It is this best-response property that defines the dynamics of the coalition formation process.

5. CASE STUDY

This section demonstrates the performance of the model using a non-superadditive CFG with types. The goal of these experiments is to show that the learning process of the MAS converges to the theoretical equilibrium points inside the core. The case study is such that the allocation vector $\mathbf{x}$ that belongs to the core is not available if the grand coalition is formed. When the population splits into small groups, the net worth of the population is increased and the allocation $\mathbf{x}$ becomes available.

Production with Two Types of Players.
There are 2 types of players $\{cooker, helper\}$, 2 types of products $\{cake = 10, cookie = 1\}$ and 4 sorts of cooking teams:
(i) 1 cooker and 2 helpers can make a cake;
(ii) 4 cookers alone can make a cake (too many cookers have difficulty reaching an agreement);
(iii) a helper alone can make a cookie and
(iv) a cooker alone can do nothing.

A group (x, y) consisting of x cookers and y helpers can realize the maximal total payoff possible from splitting into teams of the sorts described above.

5.1 Equilibria (core) analysis for different populations $(\mathbf{x}, \mathbf{y})$

Case 1 (only cookers) $\mathbf{x > 0, \ y = 0}$: If $x \leq 4$ the core is nonempty. If $x = 4k$ for some integer $k \geq 2$, the core is nonempty and consists of the payoff imputing $5/2$ to each cooker. If $x > 4$ and $x \neq 4k$ for some integer k, then in any partition of the population into teams there will be some leftover cookers. These cookers create instability (for any division of the payoff among the employed cookers, unemployed cookers can profit by offering to work for a lower payment).

Case 2 (only helpers) $\mathbf{x = 0, \ y > 0}$: For any $y > 0$ the core is nonempty and assigns 1 to each helper.

[6]the information about the time point t and the joint action $\mathbf{a}(t)$ are enough to define the state s at time $t+1$. In that case, the extensive form game can be compressed into a stochastic game that completely ignores the history of the system.

Case 3 (many helpers): For population $\mathbf{Q} = (x, y)$, where $\mathbf{Q} = r_1(1, 2) + r_2(0, 1)$ for some positive integers r_1 and r_2, there are leftover helpers from coalitions $(1, 2)$. The core is nonempty, with imputation $(8, 1)$. 1 to each helper and 8 to each cooker. Intuitively, "competition" between helpers keeps the price of a helper down to his opportunity price in a helper-only group.

Case 4 (many cookers): For $\mathbf{Q} = r_1(1, 2) + r_2(4, 0)$, the core is nonempty, with imputation $(5/2, 15/4)$. Competition between cookers for helpers keeps the price (core payoff) for helpers up to $15/4$, while cookers get no surplus from being in "mixed" groups.

Case 5 (exact proportions): For $\mathbf{Q} = r(1, 2)$ the core contains a continuum of points and its extreme points are described by the cores in Cases 3 and 4 above.

What is interesting about this case is that it can easily be verified that if $\mathbf{Q}$ is a total population with $x \geq 4$ and $y \geq 2$ the core is nonempty only if the population is described by one of the cases 2 to 5 or by Case 1 with $x \leq 4$ or $x = 4k$ for some integer k. In any other case, there will be "leftover" cookers or helpers who cannot realize the payoffs received by other players of the same type. These players create the instability associated with an empty core.

6. RESULTS

The results presented in the previous section show the learning curve of the MAS. The x-axis represents the learning trials, where each trial is composed of a variable number of steps and at the end of each trial, the environment is reset. The y-axis represents the sum of rewards by all coalitions using the characteristic function presented in the case study, e.g. for $\mathbf{Q} = \{2, 6\}$ (Case 3), the core solution is to have two coalitions $\{1, 2\}$ and two $\{0, 2\}$, achieving in total: $v(\{1, 2\}) \times 2 + v(\{0, 2\}) \times 2 = 20 + 2$. All unsuccessful coalitions (all not mentioned coalitions) are assigned a reward of zero. The results presented here use the random action selection model presented in Section 4.3. All experiments were performed in self play with the following parameterization: $\alpha = 0.3$ with a decreasing rate of 0.001, $\epsilon = \max(0.3 - 0.001t, 0)$ and Q-tables with low initialization. We let the learning process execute $n \times 3$ steps for each trial using random turns, therefore, in average, each agent will have a change to change action 3 times in each trial, and after $n \times 3$ steps, the environment is reset.

In Figure 1 we present experimental results of the MAS learning process using most interesting cases from the case study. The plots shows how a MAS integrated by all learners can effectively learn equilibrium points that coincide with the stability concept of the core. The chosen case study, as shown in literature [7], has a solution to all the previously presented cases. Beside those cases with a stable solution inside the core, we also present a random unstable case, Fig. 1(d), that shows that even if the MAS could potentially form a much better CS, this is no stable and will not be formed. As can be seen in the plots, the MAS reaches the CS inside the core as the farsightedness γ of the agents diminishes until some lower threshold ($\gamma = 0.2$), this is because each agent gets to update a state 3 times in average each trial, given that γ can be interpreted as the expected probability of the game being finished the next time step, leaving a low γ induces agents to form coalitions sooner than later.

6.1 Discussion

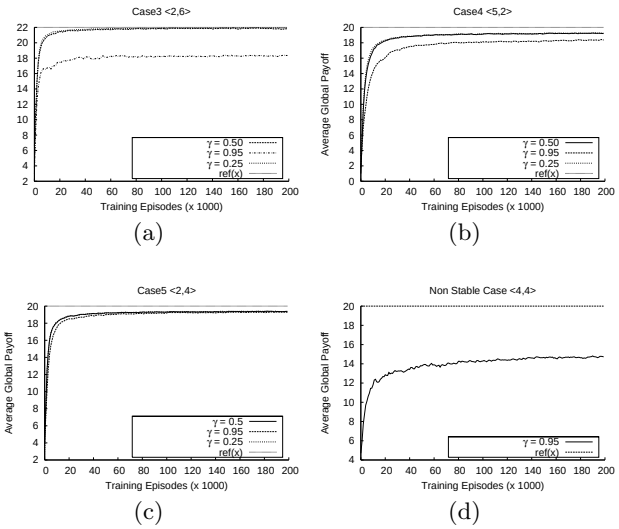

Figure 1: Plots showing convergence performance of the learning system to different populations. 1(a): Case 3 (many helpers); 1(b): Case 4 (many cookers); 1(c): Case 5 (perfect proportion); 1(d): non stable case. Each graph presents the global reward obtained by the MAS and is referenced with the core solution.

The results from the previous section show how a MAS integrated by all learners can effectively learn equilibrium points that coincide with the stability concept of the core. Even though these results are encouraging, there is much to be done. This paper uses the concept of *marginal contribution*, which sounds appealingly straightforward to use. We obtained good results with it, which suggests that the limitations and consequences of using it as an instantaneous reward for a learning agent should be taken seriously.

We presented results from the random action selection model, which has showed the best performance in the learning task. From a learning viewpoint, using a randomized turn is translated in a more complete exploration of the state space. This can be better viewed by analyzing the fixed turn model. In this model the state an agent perceives is completely determined by the state-action pair of the previous agent, therefore, the $\max_{a' \in \mathcal{A}} Q(s', a')$ in Equation (6) will make the next-turn agent experience only a subgroup of the state spate.

7. CONCLUSIONS

The results obtained are a first step towards this new learning framework for coalition formation where adaptability is the main concern. The work is relevant in that it is a distributed solution that can react to unknown changes to the environment (e.g. changes in the characteristic function or agents signing in/out), which is a key aspect of nowadays MAS domains. The work by Chalkiadakis and Boutilier [2] has been, to our knowledge, the only one work that fits into this description. They showed that belief based learning techniques can successfully be used for coalition formation. This approach is specially useful when faced against uncertain opponents, given that they build a model of each other agent in the environment, however the computational expense of such effort is high, and therefore, is hardly scalable.

We in the other hand are concerned with the problem of studying a model-free ML technique in CFGs, and provide conditions for which this techniques can be successful in achieving optimality. We then experimentally showed that the learning process coincide with the theoretical equilibria inside the core, and in cases where the core is empty, the agents can still learn to form stable coalitions; the process converges and stabilizes to some solution that is at least better than no coalitions at all.

The ideas stated here where thought of to provide a dynamic and adaptive solution that satisfies both, an stable coalition structure generation process that is completely distributed and a MAS that is reactive to changes in the environment (e.g. an agent signing-in). Although, his work is just the start, we provide enough evidence to sustain that a process of coalition formation can be reliably achieved by self-interested autonomous agents and this process can attain better outcomes at the single agent and MAS perspectives. Furthermore, this outcomes can be achieved under an heterogeneous environment. If agents sign-in or out at any time in the process, previous experiences will continue to be as valid, therefore the process needs not to start all over.

8. FUTURE WORKS

There are many directions that this paper opens for future research. A further and deeper understanding of credit assignment for a reinforcement learning problem is needed. Studying works in bargaining and other ways of payoff division could turn useful in understanding this issue. It would be interesting to study how the line of research followed by N. Vlassis [6], as a way of payoff redistribution could help in these settings.

As explained in Section 2.2, the search for the best coalition structure is of size $O(n^n)$ and the state of our proposal is the current CS^t, therefore this approach by no means simplifies the problem. The main interest of this work was to study the feasibility of multiagent RL techniques in the problem of coalition formation, excluding the problem of complexity, we proved this to be a feasible solution and to be useful in changing environments. We are now working in the issue of reducing the complexity of the problem in order to make this approach further scalable.

This learning framework suggests its application to task allocation problems where hard tasks can only be executed by a known configuration of agent aptitudes and in problems where the characteristic function cannot be known *a priori* (given that $v(S)$ could be generated on-line when the coalition is formed and its performance can be valued numerically). Furthermore, there are many problems in operations research literature that are suitable for testing under this framework, e.g. set partition/covering problems, it would be interesting to confront these results in terms of computational complexity and optimality.

9. ADDITIONAL AUTHORS

Additional authors: Andrea Bonarini,
email: `bonarini@elet.polimi.it`.

10. REFERENCES

[1] R. J. Aumann and S. Hart, editors. *Handbook of Game Theory, with Economic Applications, Vol. 2.* North-Holland, Amsterdam, 1994.

[2] G. Chalkiadakis and C. Boutilier. Bayesian reinforcement learning for coalition formation under uncertainty. In *AAMAS 2004 proceedings*. ACM Press, 2004.

[3] D. Fudenberg and J. Tirole. *Game Theory*. The MIT Press, Cambridge, MA, USA, 1991.

[4] S. P. Ketchpel. Forming coalitions in the face of uncertain rewards. In *AAAI*, pages 414–419, 1994.

[5] M. Klusch and A. Gerber. Dynamic coalition formation among rational agents. *IEEE Intelligent Systems*, 17(3):42–47, 2002.

[6] J. R. Kok and N. Vlassis. Collaborative multiagent reinforcement learning by payoff propagation. *Journal of Machine Learning Research*, 2006. to appear.

[7] A. Kovalenkov and M. Wooders. Approximate cores of games and economies with clubs. *Journal of Economic Theory*, (110):87–120, 2003.

[8] T. Sandholm, K. Larson, M. Andersson, O. Shehory, and F. Tohmé. Coalition structure generation with worst case guarantees. *Artificial Intelligence*, 111:209–238, 1999.

[9] T. Sandholm and V. R. Lesser. Coalitions among computationally bounded agents. *Artificial Intelligence*, 94(1-2):99–137, 1997.

[10] S. Sen and P. S. Dutta. Searching for optimal coalition structures. In *ICMAS '00: Proceedings of the Fourth International Conference on MultiAgent Systems (ICMAS-2000)*, page 287, Washington, DC, USA, 2000. IEEE Computer Society.

[11] O. Shehory and S. Kraus. Methods for task allocation via agent coalition formation. *Artificial Intelligence*, 101(1–2):165–200, 1998.

[12] P. P. Shenoy. On coalition formation: a game-theoretical approach. *International Journal of Game Theory*, 8(3):133–164, 1979.

[13] P. P. Shenoy. A dynamic solution concept for abstract games. *Journal of Optimization Theory and Applications*, 32(2):151–169, 1980.

[14] C. J. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

[15] G. Zlotkin and J. S. Rosenschein. Coalition, cryptography, and stability: Mechanisms for coalition formation in task oriented domains. In *AAAI*, pages 432–437, 1994.