Bayesian Reinforcement Learning for Coalition Formation under Uncertainty

Georgios Chalkiadakis
Dept. of Computer Science, Univ. of Toronto
Toronto, ON, M5S 3G4, Canada
gehalk@cs.toronto.edu

Craig Boutilier
Dept. of Computer Science, Univ. of Toronto
Toronto, ON, M5S 3G4, Canada
cebly@cs.toronto.edu

Abstract

Research on coalition formation usually assumes the values of potential coalitions to be known with certainty. Furthermore, settings in which agents lack sufficient knowledge of the capabilities of potential partners is rarely, if ever, touched upon. We remove these often unrealistic assumptions and propose a model that utilizes Bayesian (multiagent) reinforcement learning in a way that enables coalition participants to reduce their uncertainty regarding coalitional values and the capabilities of others. In addition, we introduce the Bayesian Core, *a new stability concept for coalition formation under uncertainty. Preliminary experimental evidence demonstrates the effectiveness of our approach.*

1. Introduction

Coalition formation, widely studied in game theory and economics [7], has attracted much attention in AI as means of dynamically forming partnerships or teams of cooperating agents. Most models of coalition formation assume that the values of potential coalitions are known with certainty, implying that agents possess knowledge of the capabilities of their potential partners, or at least that this knowledge can be reached via communication (e.g., see [11, 12]). However, in many natural settings, rational agents must form coalitions and divide the generated value without knowing *a priori* what this value may be or how suitable their potential partners are for the task at hand. The case of an enterprise trying to choose subcontractors while unsure of their capabilities is only one such example. The creation and interaction of *virtual organizations* has been anticipated as a long term impact of agent coalition technologies on e-commerce; this cannot possibly be achieved without dealing with the problem of uncertainty.

The presence of uncertainty poses interesting theoretical questions, such as the discovery of analogs of the traditional concepts of stability. Furthermore, it suggests opportunities for agents to learn about each others' abilities through repeated interaction, refining how coalitions are formed over time. As a consequence, realistic models of coalition formation must be able to deal with situations in which the presence of action uncertainty and the types of the potential partners is translated into uncertainty about the values of various coalitions.

To this end, we propose a new model of coalition formation in which agents must derive coalitional values by reasoning about the *types* of other agents and the uncertainty inherent in the actions a coalition may take. We propose a new stability concept, the *Bayesian core (BC)*, suitable for this setting, and describe a dynamic coalition formation process that will converge to the BC if it exists. Furthermore, since one agent will generally learn something about the abilities of other agents via interaction with them in a coalition, we propose a *reinforcement learning* (RL) model in which agents refine their beliefs about others through repeated interaction. We propose a specific Bayesian RL model in which agents maintain explicit beliefs about the types of others, and choose actions and coalitions not only for their immediate value, but also for their *value of information* (i.e., what can be learned about other agents). We show that the agents in this framework can reach coalition and payoff configurations that are stable given the agents' beliefs, while learning the types of their partners and the values of coalitions. We believe that these ideas could be of value for, say, e-commerce applications where *trust* among potential partners (e.g., where each is uncertain about the other's capabilities) is an issue.

We begin in Section 2 with a discussion of relevant work on coalition formation, including recent work that deals with dynamic coalition formation and some forms of uncertainty. In Section 3 we propose a model for coalition formation in which agent abilities are not known with certainty, and actions have stochastic effects. We introduce the Bayesian core concept and a suitable dynamic formation process. We then describe a Bayesian RL model in Section 4 that allows agents to learn about their partners through their interactions in coalitions.

2. Background

Cooperative game theory deals with situations where players act together in a cooperative equilibrium selection process involving some form of bargaining, negotiation, or arbitration [7]. Coalition formation is one of the fundamental areas of study within cooperative game theory. We briefly review relevant work in this section.

2.1. Coalition Formation

Let $N = \{1, \ldots, n\}$, $n > 2$, be a set of players. A subset $S \subseteq N$ is called a *coalition*, and we assume that agents participating in a coalition may coordinate their activities for mutual benefit. A *coalition structure* is a partition of the set of agents containing exhaustive and disjoint coalitions. *Coalition formation* is the process by which individual agents form such coalitions, generally to solve a problem by coordinating their efforts. The coalition formation process can be seen as being composed of the following activities [1, 10]: (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.[1]

While seemingly complex, coalition formation can be abstracted into a fairly simple model [7]. A *characteristic function* $v : 2^N \rightarrow \Re$ defines the *value* $v(S)$ of each coalition S. Intuitively, $v(S)$ represents the maximal payoff the members of S can jointly receive by cooperating effectively. An *allocation* is a vector of payoffs $\vec{x} = (x_1, ..., x_n)$ assigning some payoff to each $i \in N$. An allocation is *feasible* w.r.t. coalition structure CS if $\sum_{i \in S} x_i \leq v(S)$ for each $S \in CS$, and is *efficient* if this holds with equality. When rational agents seek to maximize their individual payoffs, *stability* becomes critical. Research in coalition formation has developed several notions of stability, among the strongest being the *core*.

Defn 1 Let CS be a coalition structure, and let $\vec{x} \in \Re^n$ be some allocation of payoffs to the agents. The *core* is the set of payoff configurations

$$C = \{(\vec{x}, CS) | \forall S \subseteq N, \sum_{i \in S} x_i \geq v(S) \text{and} \sum_{i \in N} x_i = \sum_{S \in CS} v(S)\}$$

In a core allocation, no subgroup of players can guarantee all of its members a higher payoff. As such, no coalition would ever "block" the proposal for a core allocation. Unfortunately, the core might be empty, and, furthermore, it is exponentially hard to compute. [2] Apart from the core, there exist other solution concepts such as the *Shapley value* and the *kernel* [7].

In recent years, there has been extensive research covering many aspects of the coalition formation problem. None has yet dealt with dynamic coalition formation under the "extreme" uncertainty we tackle in this paper. However, various coalition formation processes and some types of uncertainty have been studied. We briefly review some of the work upon which we draw.

Dieckmann and Schwalbe [5] describe a *dynamic* process of coalition formation (under the usual deterministic coalition model). This process allows for exploration of suboptimal "coalition formation actions." At each stage, a randomly chosen player decides which of the existing coalitions to join, and demands a payoff. A player will join a coalition if and only if he believes it is in his best interest to do so. These decisions are determined by a "non-cooperative best-reply rule", given the coalition structure and allocation in the previous period: a player switches coalitions if his expected payoff in the new coalition exceeds his current payoff; and he demands the most he can get subject to feasibility. The players observe the present coalitional structure and the demands of the other agents, and expect the current coalition structure and demand to prevail in the next period. The induced Markov process (when all players adopt the best-reply rule) converges to an absorbing state; and if players can *explore* with myopically suboptimal actions all absorbing states are core allocations. Konishi and Ray [6] study a somewhat related coalition formation process.

Sjuis *et al.* [14, 13] introduce *stochastic cooperative games (SCGs)*, comprising a set of agents, a set of coalitional actions, and a function assigning to each action a random variable with finite expectation, representing the payoff to the coalition when this action is taken. These papers provide strong theoretical foundations for games with this restricted form of action uncertainty, and describe classes of games for which the core of a SCG is nonempty (though no coalition formation process is explicitly modeled).

Finally, Shehory and Kraus have proposed coalition formation mechanisms that take into account the capabilities of the various agents [11] and deal with expected payoff allocation [12]. However, information about the capabilities or resources of others is obtained via communication.

2.2. Bayesian Reinforcement Learning

Since we will adopt a Bayesian approach to learning about the abilities of other agents, we briefly review relevant prior work on Bayesian RL. Assume an agent is learning to control a stochastic environment modeled as a Markov decision process (MDP) $\langle \mathcal{S}, \mathcal{A}, R, \Pr \rangle$, with finite state and action sets $\mathcal{S}, \mathcal{A}$, reward function R, and dynamics $\Pr$. The

1 Throughout we assume transferable utility.

2 We do not deal with complexity issues related with coalition formation in this work; we note, however, that such issues have become the focus of recent research (e.g., [9]).

dynamics $\Pr$ refers to a family of transition distributions $\Pr(s, a, \cdot)$, where $\Pr(s, a, s')$ is the probability with which state s' is reached when action a is taken at s. $R(s, r)$ denotes the probability with which reward r is obtained when state s is reached. The agent is charged with constructing an optimal Markovian policy $\pi : \mathcal{S} \mapsto \mathcal{A}$ that maximizes the expected sum of future discounted rewards over an infinite horizon: $E_\pi[\sum_{t=0}^{\infty} \gamma^t R^t | S^0 = s]$. This policy, and its value, $V^*(s)$ at each $s \in \mathcal{S}$, can be computed using standard algorithms such as policy or value iteration.

In the RL setting, the agent does not have direct access to R and $\Pr$, so it must learn a policy based on its interactions with the environment. Any of a number of RL techniques can be used to learn an optimal policy.

In *model-based RL* methods, the learner maintains an estimated MDP $\langle \mathcal{S}, \mathcal{A}, \widehat{R}, \widehat{\Pr} \rangle$, based on the set of experiences $\langle s, a, r, t \rangle$ obtained so far. At each stage (or at suitable intervals) this MDP can be solved (exactly or approximately). Single-agent Bayesian methods [4] allow agents to incorporate priors and explore optimally, assuming some prior density P over possible dynamics D and reward distributions R, which is updated with each data point $\langle s, a, r, t \rangle$.

Similarly, multi-agent Bayesian RL agents [3] update prior distributions over the space of possible models as well as the space of possible strategies being employed by other agent. The value of performing an action at a belief state involves two main components: an expected value with respect to the current belief state; and its impact on the current belief state. The first component is typical in RL, while the second captures the *expected value of information* (EVOI) of an action. Each action gives rise to some "response" by the environment that changes the agent's beliefs, and these changes can influence subsequent action choice and expected reward. EVOI need not be computed directly, but can be combined with "object-level" expected value through Bellman equations describing the solution to the POMDP that represents the exploration-exploitation problem by conversion to a belief state MDP.

3. A Bayesian Coalition Formation Model

In this section we introduce the problem of Bayesian coalition formation, define the Bayesian core, and describe a dynamic coalition formation process for this setting.

3.1. The Model

A *Bayesian coalition formation problem* is characterized by a set of agents, a set of types, a set of coalitional actions, a set of outcomes or states, a reward function, and agent beliefs over types. We describe each of these components in turn.

We assume a set of agents $N = \{1, \ldots, n\}$, and for each agent i a finite set of possible *types* T_i. Each agent i has a specific type $t \in T_i$, which intuitively captures i's "abilities" (in a way that will become apparent when we describe actions). We let $T = \times_{i \in N} T_i$ denote the set of type profiles. For any coalition $C \subseteq N$, $T_C = \times_{i \in N} T_i$, and for any $i \in N$, $T_{-i} = \times_{j \neq i} T_j$. Each i knows its own type t_i, but not those of other agents. Agent i's *beliefs* B_i comprise a joint distribution over T_{-i}, where $B_i(\vec{t}_{-i})$ is the probability i assigns to other agents having type profile $\vec{t}_{-i}$. We use $B_i(\vec{t}_C)$ to denote the marginal of B_i over any subset C of agents, and for ease of notation, we let $B_i(t_i)$ refer to i "beliefs" about its own type (assigning probability 1 to its actual type and 0 to all others).

A coalition C has available to it a finite set of *coalitional actions* A_C. When an action is taken, it results in some outcome or *state* $s \in \mathcal{S}$. The odds with which an outcome is realized depends on the types of the coalition members (e.g., the outcome of building a house will depend on the abilities of the team members). We let $\Pr(s|\alpha, \vec{t}_C)$ denote the probability of outcome s given that coalition C takes action $\alpha \in A_C$ and member types are given by $\vec{t}_C \in T_C$. Finally, we assume that each state s results in some *reward* $R(s)$. If s results from a coalitional action, the members are assigned $R(s)$, which is assumed to be divisible/transferable among the members.

The *value* of coalition C with members of type $\vec{t}_C$ is:

$$V(C|\vec{t}_C) = \max_{\alpha \in A_C} \sum_s \Pr(s|\alpha, \vec{t}_C) R(s) = \max_{\alpha \in A_C} Q(C, \alpha|\vec{t}_C)$$

Unfortunately, this coalition value cannot be used in the coalition formation process if the agents are uncertain about the types of their potential partners. However, each i has beliefs about the value of any coalition based on its expectation of this value w.r.t. other agents's types:

$$V_i(C) = \max_{\alpha \in A_C} \sum_{\vec{t}_C \in T_C} B_i(\vec{t}_C) Q_i(C, \alpha|\vec{t}_C) = \max_{\alpha \in A_C} Q_i(C, \alpha)$$

Note that $V_i(C)$ is not simply the expectation of $V(C)$ w.r.t. i's belief about types. The expectation Q_i of action values (i.e., Q-values) cannot be moved outside the max operator: a single action must be chosen which is useful *given* i's uncertainty. Of course, i's estimate of the value of a coalition, or any coalitional action, may not conform with those of other agents. This leads to additional complexity when defining suitable stability concepts. We turn to this issue in the next section. However, i is certain of its *reservation value*, the amount it can attain by acting alone:

$$rv_i = V_i(\{i\}) = \max_{\alpha \in A_{\{i\}}} \sum_s \Pr(s|\alpha, t_i) R(s)$$

3.2. The Bayesian Core

We define an analog of the traditional core concept for the Bayesian coalition formation scenario. The notion of stability is made somewhat more difficult by the uncertainty associated with actions: since the payoffs associated with coalitional actions are stochastic, allocations must reflect this [14, 13]. Stability is rendered much more complex still by the fact that different agents have potentially different beliefs about the types of other agents.

Because of the stochastic nature of payoffs, we assume that players join a coalition with certain *relative payoff demands* [14]. Let $\vec{d}$ represent the payoff demand vector $\langle d_1, \ldots, d_n \rangle$, and $\vec{d}_C$ the demands of those players in coalition C. For any agent $i \in C$ we define the relative demand of agent to be $r_i = \frac{d_i}{\sum_{j \in C} d_j}$. If reward R is received by coalition C as a result of its choice of action, each i receives payoff $r_i R$. This means that the excesses or losses deriving from the fact that the reward function is stochastic are expected to be allocated to the agents in proportion to their agreed upon demands. As such, each agent has beliefs about any other agent's expected payoff given a coalition structure and demand vector. Specifically, agent i's beliefs about the *expected stochastic payoff* of some agent $j \in C$ is denoted $\bar{p}^i_j = r_j V_i(C)$. If $i \in C$, i believes its *own* expected payoff to be $\bar{p}^i_i = r_i V_i(C)$.

A difficulty with using $V_i(C)$ in the above definition of expected stochastic payoff is that it assumes that all coalition members agree with i's assessment of the best (expected reward-maximizing) action for C. Instead, we suppose that coalitions are formed using a process by which some coalitional action α is agreed upon, much like demands. In this case, i's beliefs about j's expected payoff is $\bar{p}^i_j(\alpha, C) = r_j Q_i(C, \alpha)$. Finally, we let $\bar{p}^i_{j,C}(\alpha, \vec{d}_C)$ denote i's beliefs about j's expected payoff if it were a member of any $C \subseteq N$ with demand $\vec{d}_C$ taking action α:

$$\bar{p}^i_{j,C}(\alpha, \vec{d}_C) = \frac{d_j Q_i(C, \alpha)}{\sum_{k \in C} d_k}$$

Intuitively, if a coalition structure and payoff allocation are stable, we would expect: (a) no agent believes it will receive a payoff (in expectation) that is less than its reservation value; and (b) based on its beliefs, no agent will have an incentive to suggest that the coalition structure (or its allocation or action choice) is changed—specifically, there is no alternative coalition it could reasonably expect to join that offers it a better payoff than it expects to receive given the action choice and allocation agreed upon by the coalition to which it belongs.

Thus we define the *Bayesian core (BC)* as follows:

Defn 2 Let $\langle CS, \vec{d} \rangle$ be a coalition-structure, demand vector pair, with C_i denoting the $C \in CS$ of which i is a member. Then $\langle CS, \vec{d} \rangle$ is in the *Bayesian core* of a Bayesian coalition problem iff, for all $C \in CS$, there exists an $\alpha \in A_C$ such that, for no $S \subseteq N$ is there an action $\beta \in A_S$ and demand vector $\vec{d}_S$ s.t. $\bar{p}^i_{i,S}(\beta, \vec{d}_S) > \bar{p}^i_i(\alpha, C_i), \forall i \in S$.

In words, all agents in every $C \in CS$ believe that the coalition structure and payoff allocation currently in place ensure them expected payoffs that are as good as any they might realize in any other coalition.[3] Furthermore, their beliefs "coincide" in the weak sense that there is some coalitional action α that they commonly believe to ensure this better payoff. This doesn't mean that α is what each believes to be best. But an agreement to do α is enough to keep *each* member of C from defecting.

The *core* is a special case of the BC when all agents know the types of other agents (which is the only way their beliefs can coincide, since each agent knows its own type). In this case, all beliefs about coalitional values coincide, and the BC coincides with the core of the induced characteristic function game. Since the core is not always non-empty, it follows that the BC is not always non-empty.

3.3. Dynamic Coalition Formation

We now propose a protocol for dynamic coalition formation. The protocol is derived from the process defined in [5], with two main differences: it deals with expected, rather than certain, coalitional values; and it allows for the proposal of a coalitional action during formation.

The process proceeds in stages. At any point in time, we suppose there exists a structure CS, demand vector $\vec{d}$, and a set of agreed upon coalition actions $\vec{\alpha}_{CS}$ (with one $\alpha \in A_C$ for each $C \in CS$).[4] With some probability γ_i, agent i, the *proposer*, is given the opportunity to propose a change to the current structure. We assume $\gamma_i > 0$ for each $i \in N$, and permit i the following options: it can propose to stay in its current coalition, but propose a new demand d_i and/or a new coalitional action; or it can propose to join any other existing coalition with a new demand d_i and a suggested coalitional action. The second option includes the possibility that i "breaks away" into a singleton. If i proposes a change to the current structure/demand/action, then the new arrangement will occur only if all "affected" coalition member agree to the change. Otherwise, the current structure and agreements remain in force.

To reflect the rationality of the players, we impose restrictions on the possible proposal and acceptance decisions. Specifically, we require the proposer to suggest a new

3 Note that if some S, β, and $\vec{d}_S$ exist that makes some $i \in S$ strictly better off while keeping all other $j \in S$ equally well off, then there must exist a $\vec{d}_S$ that makes *all* $j \in S$ strictly better off.

4 We might initially start with singleton coalitions with the obvious choices of actions.

demand that maximizes its payoff, while taking into consideration its beliefs about whether affected agents will accept this demand. Thus for any coalition it proposes to join (or new demand it makes of its own coalition), it will ask for the maximum demand that it believes affected members will find acceptable.

Let $\bar{p}^i_{i,C}(\alpha, d_i) = \bar{p}^i_{i,C}(\alpha, \vec{d}_C \circ d_i)$ denote i's beliefs about its expected payoff should it join coalition $C \in CS$ with demand d_i (or make a new demand of its own coalition), with $C \cup \{i\}$ taking action α. When proposing to join C, i should make the maximum demand (d_i and α) that is *feasible* according to its beliefs, in other words, that it believes the other agents will accept. More precisely, we say $\langle C, d_i, \alpha \rangle$ is feasible for i if:

$$\forall j \in C, \frac{d_j Q_i(C \cup \{i\}, \alpha)}{\sum_{k \in C \cup \{i\}, \text{s.t.} k \neq i} d_k + d_i} \geq \bar{p}^i_j$$

If $\langle C, d_i, \alpha \rangle$ is feasible for i, then i expects the members of C to accept this demand. Of course, i does not know this for sure, since it does not know what the members of C believe, but has its own estimates of their current values $\bar{p}^i_j$.[5] Agent i can directly calculate its maximum rational demand w.r.t. C and action α:

$$d^{max}_i(C, \alpha) = \min_j \frac{d_j Q_i(C \cup \{i\}, \alpha) - \bar{p}^i_j \sum_{k \in C \cup \{i\}, k \neq i} d_k}{\bar{p}^i_j}$$

Assumption 1 *Let $0 < \delta < 1$ be a sufficiently small* smallest accounting unit. *When any i makes a demand $\langle C, d_i, \alpha \rangle$ to coalition C, its payoff demand d_i is restricted to the finite set $D_i(C, \alpha)$ of all integral multiples of δ in the closed interval $[rv_i, d^{max}_i(C, \alpha)]$.*

With this model in place, we can define two related coalition formation processes. In the *best reply (BR) process*, proposers are chosen randomly as described above, and any proposer i is required to make its maximal feasible demand:

$$\max_C \max_{\alpha \in A_C} \max_{d_i \in D_i(C,\alpha)} \bar{p}^i_{i,C}(\alpha, d_i).$$

If there are several maximal feasible demands, i chooses among them with equal probability. As above, such a proposal is accepted only in all members of affected coalition are no worse off in expectation (w.r.t. their own beliefs).

This best reply process induces a discrete-time, finite-state Markov chain with states of the form $\langle CS^t, \vec{\alpha}^t, \vec{d}^t \rangle$. This state at time t is sufficient to determine the probability of transitioning to any new state at time $t + 1$.

We also consider a slight modification of the best reply process, the *best reply with experimentation (BRE) process*. It proceeds similarly to BR with the following exception: if proposer i believes there is a coalition S which will be beneficial to it, but which cannot be derived starting from the existing CS, it can propose an arbitrary feasible demand in order to destabilize the current state in hopes of reaching a better structure. More precisely, the best reply is chosen with probability $1 - \varepsilon$, and some other feasible demand with arbitrarily small $d_i \geq 0$ is chosen with probability ε (each with equal probability). This can be viewed as a "trembling" mechanism or as explicit experimentation. Furthermore, any agent j that is part of an affected coalition will choose to accept a demand from i that lowers its payoff with probability ε iff j believes there exists some coalition S, with $i, j \in S$, such that all members of S are better off than currently (i.e., $V_j(S) > \sum_{k \in S} \bar{p}^j_k$).

The BRE process has some reasonable properties. First we note that absorbing states of the process coincide with Bayesian core allocations.

Theorem 1 *The set of demand vectors associated with an absorbing state of the BRE process coincides with the set of BC allocations. Specifically, $\omega = \langle CS, \vec{d}, \vec{\alpha}_{CS} \rangle$ is an absorbing state of the BRE process iff $\langle CS, \vec{d} \rangle \in \text{BC}$ and each $\alpha_C, C \in CS$ satisfies the stability requirement.*

Proof sketch: Part (i): If a state ω is in the BC, no agent believes that he can gain either by switching coalitions or by changing his demand. Moreover, as no agent believes that a "blocking" coalition exists, no agent experiments.

Part (ii): Suppose that $\omega = \langle CS, \vec{d}, \vec{\alpha}_{CS} \rangle$ is a non-BC absorbing state of the BRE process. Since it is not in the BC, then there exists some i that believes there exists an S and α' s.t. $\bar{p}^i_{i,S}(\alpha') > \bar{p}^i_{i,C}(\alpha^i_{CS})$. Consequently, with probability ϵ, at least i will experiment, potentially asking for zero payoff. Thus, there exists a positive probability that ω will be left, which contradicts the statement that ω is absorbing.

Theorem 1 does not guarantee that a BC allocation will actually be reached by the BRE process. However, we can prove the following theorem:

Theorem 2 *If the BC is non-empty, the BRE process will converge to an absorbing state with probability one.*

Proof sketch: The proof is completely analogous to the proof for the deterministic coalition formation model [5]. The basic idea is that when the BC is not empty, all ergodic sets reached by the BRE process are singletons, therefore the BRE process will converge to an absorbing state.

Theorems 1 and 2 together ensure that if the BC is not empty then the BRE process will eventually reach a BC allocation, no matter what the initial coalition structure.

To test the validity of this approach empirically, we examined the BRE process on several simple Bayesian coali-

5 This poses the interesting question of how best to model one agent's beliefs about another's beliefs in this setting.

tion problems. The first test was very simple: a game with three agents, each having common beliefs about coalition values. We start the BRE process in an absorbing state of the best reply (*without* experimentation) process, but which is not in the BC. Unsurprisingly, the BR process never left the absorbing state, while the BRE process always (30/30 runs) converges to an absorbing configuration in the BC. In 19 runs, the process converged in fewer than fifty proposals (typically less than 25) but in one case took 207 rounds.

For interest, we tested BRE on the 3-player majority game [5, 7], a well-known example of a game having an empty core. The agents shared common beliefs about the coalitional values to mimic a standard characteristic function game. As expected, neither the BR nor the BRE process converged to a stable configuration.

We also tested a game having a non-empty BC in which the initial beliefs of the three agents differed and each coalition had three actions available to it. The BRE process managed to reach a BC configuration in all 30 runs tested. The greatest number of negotiation rounds to convergence was 292, however, a BC configuration was typically reached in less than 100 rounds. In five of the 30 runs, the agents reached a BC configuration almost instantaneously (in less than five rounds). The BR process, on the other hand, converged to a BC configuration in only 19/30 runs. Interestingly, in all runs it did reach a coalition *structure* in the BC, but not always with an appropriate payoff allocation.

4. A Bayesian RL Framework

While the core and related stability concepts provide firm foundations for cooperative games when all coalitions have known value, their applicability in realistic settings must be called into question. Generally, agents will face the types of uncertainty described above. One may ask, of course, if agents are faced with the possibility of repeated interaction, would most uncertainty about agent types eventually vanish? We argue that in fact, not only is it generally infeasible for "type uncertainty" to vanish, but furthermore that agents often have no incentive to engage in actions (or interactions) that would reduce this uncertainty.

In this section, we describe an RL model in which agents repeatedly form coalitions and take coalitional actions. This gives agents the opportunity, through observation of the outcome of coalitional actions, to update their beliefs about the types of their partners. This will, using notions like the BC, influence future coalition formation decisions. With a Bayesian approach to *repeated coalition formation*, agents are often satisfied not to learn about the abilities of potential partners, if the costs of doing so outweigh the anticipated benefits (or *value of information*).

We suppose a standard Bayesian coalition problem as before, with N players, each having *initial beliefs* B_i. The RL process proceeds in stages: at each stage t, the agents engage in some coalition formation process, based on their current beliefs B_i^t.[6] Once coalitions are formed, each $C \in CS^t$ takes its agreed upon action α_C^t and observes the resulting state s. Each member of the coalition then updates its beliefs about its partners' types:

$$B_i^{t+1}(\vec{t}_C) = z \Pr(s|\alpha, \vec{t}_C) B_i^t(\vec{t}_C)$$

where z is a normalizing constant (we sometimes denote the updated belief state as $B_i^{s,\alpha}$). The process then repeats.

We adopt an approach to optimal repeated coalition formation that uses *Bayesian exploration*. As demonstrated in our approach to multiagent RL [3], Bayesian agents in multiagent interaction can balance exploration with exploitation, effectively realizing *sequential* performance that is optimal with respect to their beliefs about other agents.[7] We cast the problem of optimal learning as a partially observable MDP (POMDP), or a *belief-state MDP*.

If we assume an infinite horizon problem, with discount factor $0 \leq \gamma < 1$, it is reasonably straightforward to formulate the optimality equations for the POMDP; however, certain subtleties will arise because of an agent's lack of knowledge of other agent beliefs. Let agent i have beliefs B_i about the types of other agents. Let $Q_i(C, \alpha, \vec{d}_C, B_i)$ denote the (long-term) value i places on being a member of coalition C that has agreed action α and demands $\vec{d}_C$, realizing that after this action is taken, the coalition formation process will repeat. This is defined as:

$$\begin{aligned} Q_i(C, \alpha, \vec{d}_C, B_i) &= \sum_s \Pr(s|C, \alpha, B_i)[r_i R(s) + \gamma V_i(B_i^{s,\alpha})] \quad (1) \\ &= \sum_{\vec{t}_C} B_i(\vec{t}_C) \sum_s \Pr(s|C, \alpha, \vec{t}_C)[r_i R(s) + \gamma V_i(B_i^{s,\alpha})] \\ V_i(B_i) &= \sum_{C|i \in C, \vec{d}_C} \Pr(C, \alpha, \vec{d}_C|B_i) Q_i(C, \alpha, \vec{d}_C, B_i) \quad (2) \end{aligned}$$

Unlike the typical Bellman equations, the value function V_i cannot be defined by maximizing Q-values. This is because the choices that dictate reward, namely, the coalition that is formed, are not in complete control of agent i. Instead, i must predict, based on it beliefs, the probability $\Pr(C, \alpha, \vec{d}_C|B_i)$ with which a specific coalition C (to which it belongs) and agreement $\langle \alpha, \vec{d}_C \rangle$ will arise as a result of negotiation. However, with this in hand, the value equations provide the means to determine the *long-term value* of any coalitional agreement. Specifically, it accounts for how i's beliefs will change in the future when deciding how useful a specific coalition is now.

6 The model can be extended by allowing state transitions—we simply let the value of any coalitional action depend on the current state. This would allow for a sequential environment model (an underlying MDP). We don't consider this here, instead focusing on the sequential nature of repeated coalition formation itself.

7 Of course, this draws heavily on methods for optimal Bayesian exploration in bandit problems and single-agent RL [2, 4].

We now consider four types of reinforcement learners. The first are *Non-myopic/full negotiation (NM-FN)*. Agents in this class employ *full negotiation* when forming coalitions, attempting to find a BC structure and allocation before engaging in their actions. For instance, they might use the dynamic process described above to determine suitable coalitions given their *current* beliefs. Furthermore, they employ lookahead, or sequential reasoning, in their attempt to solve (possibly approximately) the POMDP described by Eqs. 1 and 2. Several difficulties face non-myopic RL agents. One bottleneck is calculating $\Pr(C, \alpha, \vec{d}|B_i)$ in Eq. 2, the probability of negotiation ending with the agent in a specific coalition in a specific state of the coalition formation process Markov chain (i.e., in a specific coalition structure under a specific agreement). While the Markov chain can be analyzed readily (thus determining the steady state distribution) if the parameters are known, agent i does not have full knowledge of it, since it is unaware of other agents' beliefs. However, these beliefs can be approximated in a variety of ways, and the approximate Markov chain solved.

A second difficulty facing NM-FN agents is the difficulty of solving the optimal exploration POMDP. Several computational approximations can be used in order to make this tractable. Instead of dealing with every possible future belief state, we may instead use "one-step lookahead", dealing only with immediate successor states. Alternatively, we can employ *VPI sampling*, a method developed in [4], and adapted to the multiagent RL context in [3]. This technique estimates the (myopic) value of obtaining perfect information about a coalitional action given current beliefs. In either case, the sequential value of any coalitional action, accounting for its value of information, is then used in the formation process.

Myopic/full negotiation (M-FN) agents use full negotiation to determine coalitions at each stage. However, they do not reason about future (belief) states when assessing the value of coalitional moves. Essentially, M-FN agents engage in repeated application of a myopic formation process (e.g., the straightforward proposal process described above), choose actions, and repeat.

Myopic/one-step proposers (M-OSP) are agents that are myopic regarding the use of their beliefs when estimating coalition values (like M-FN), but do not employ full negotiation to form coalitions. Rather, at each stage of the RL process, one random proposer is chosen, and once a proposal has been made and accepted or rejected, no further negotiations take place: the coalitional action is executed after a *single* proposal. *Non-myopic/one-step proposers (NM-OSP)* are, naturally, the obvious combination of NM-FN and M-OSP agents.

When comparing these approaches, we see that FN approaches have the advantage that at the end of each RL stage, before actions are executed, the coalition structure is in a stable state, provided that a coalition formation process which ensures this is employed (e.g., if the BC in nonempty and BRE is used). Another advantage of FN is that agents have the opportunity to update their beliefs regarding other agents' types during the negotiation itself (though we do not explore the possibility here). In contrast, OSP approaches, have the advantage of giving more flexibility to the agents to investigate the space of structures, trying out different coalitions without being bound to reach a stable structure at each stage before acting (and gaining information). Finally, OSP methods have the obvious advantage that they apply best in situations where "real-time" performance is an issue (since no lengthy negotiations are used after each RL stage). One can show, in fact, OSP methods will converge to the BC of a game (if it is nonempty): it is sufficient to ensure that agents' beliefs regarding coalitional values stabilize over time.

We have to date only experimented with myopic approaches, but experiments with non-myopic approaches are under way. We should also note that we allowed for observability of occurring outcome states by both the members and non-members of an acting coalition, even though this is not required by our model.

To test M-FN and M-OSP, we first ran an experiment with three agents, two types per agent, three actions per coalition and three outcomes per action. When an M-FN approach was used, 1000 steps were used for the coalition formation process at each RL step. In many cases, these steps are sufficient for the agents to converge in the BC even without their beliefs having converged to true values regarding partners' types and coalitional values. After an average of 31 RL steps (in 30 runs), the agents' beliefs converge to such values that they always reach BC configurations in all subsequent coalition formation attempts. As expected with full negotiation, agents do not get to know with certainty the true types of all players. However, the agents' beliefs (after 100 RL steps) do converge on the *true types* of their partners in the BC coalitions to which they have converged, and the true values of coalitions with these partners. For "non-partners," At convergence, each agent has on average a degree of belief 0.7 regarding the true types of its "non-partners." On the other hand, agents using an M-OSP approach have the opportunity to explore the space of coalition structures more broadly. In this small problem, it takes on average 392 RL steps for their beliefs to stabilize, after which they converge to BC configurations. Furthermore, they eventually learn the true types of *all* other agents.

We also tested our approach in a setting with 5 agents, 10 types/agent, 3 actions/coalition and 3 outcome states/action. The agents form companies to bid for software development projects. There exist 3 "major" types having 3 or 4 "quality" types each: *interface designer* = $\langle bad, average, expert \rangle$,

programmer = $\langle bad, average, good, expert\rangle$ and *systems engineer* = $\langle bad, average, expert\rangle$. The companies can bid for a large, average-sized or small project, and they expect to make large, average or small profit, given their choices and their members' types. In general bidding for large projects is less likely to be rewarding, but the more members a coalition has, the more likely it is that it will be successful in getting higher profits if it tries to bid for large projects. Coalitions with competent members of different major types have more potential for high reward, in contrast to coalitions with incompetent members. The actual types of the agents were set to *a1 = bad programmer*, *a2 = good programmer*, *a3 = expert programmer*, *a4 = bad interface designer* and *a5 = bad systems engineer*. The agents know the major type of their opponents, but not their quality types. The most profitable coalition structure in our scenario was $\{\langle a1\rangle, \langle a2, a3\rangle, \langle a4\rangle, \langle a5\rangle\}$, where the two "competent" programmers form a coalition and bid for average-sized projects.

After 30 runs of 200 RL steps, FN-agents learn the actual types of the opponents with certainty 20/30 times. In these cases, they manage to converge to the most profitable coalition structure. In addition, they manage to converge to BC configurations in 20/30 runs, on average within 118 RL steps, behaving optimally to the best of their knowledge. The 10 runs that do not converge to BC configurations, have nevertheless converged to stable coalition structures, but slight changes in their beliefs make some agents occasionally unsure about their payoff allocation, and this results to the agents alternating between BC configurations and non-BC configurations.

We also chose to employ the OSP approach with only 200 RL steps for this time-consuming problem. The OSP agents manage to discover and converge to the most profitable CS in 15/30 runs—but they are not (yet) really "convinced": they converged to BC configurations in only 6/30 runs.

5. Conclusions

We proposed a new model for coalition formation with type uncertainty reflecting uncertain knowledge about the abilities of potential partners. The Bayesian core concept seems fairly natural, and we also described a dynamic coalition formation process and an RL model for learning about potential partners through repeated interaction, which is of great applicability in real-world situations. Note that the proposed Bayesian RL framework is independent of the underlying negotiation process, or the requirement to convergence to a specific stability concept. It enables agents to weigh their need to explore the abilities of their potential partners with their need to exploit knowledge acquired so far.

There are several interesting directions we intend to pursue. Among these are extending the BC concept so that it provides for "meta-reasoning" regarding one's beliefs about the beliefs of others. We are also interested in deriving conditions under which the BC is non-empty, as well as recasting the ideas presented here to coalitional bargaining with discounted payoffs in the presence of uncertainty, deriving Bayes-Nash equilibria of the negotiation process [8].

References

[1] B. Banerjee and S. Sen. Selecting Partners. In C. Sierra, M. Gini, and J. Rosenschein, editors, *Proceedings of the Fourth International Conference on Autonomous Agents*, pages 261–262, Barcelona, Catalonia, Spain, 2000. ACM Press.

[2] D. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London, 1985.

[3] G. Chalkiadakis and C. Boutilier. Coordination in Multiagent Reinforcement Learning: A Bayesian Approach. In *Proceedings of AAMAS'03*, 2003.

[4] R. Dearden, N. Friedman, and D. Andre. Model based Bayesian Exploration. In *Proceedings of Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 150–159, 1999.

[5] T. Dieckmann and U. Schwalbe. Dynamic Coalition Formation and the Core, 1998. Economics Department Working Paper Series, Department of Economics, National University of Ireland - Maynooth.

[6] H. Konishi and D. Ray. Coalition Formation as a Dynamic Process, 2002. Boston College Working Papers in Economics 478.

[7] R. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1991.

[8] A. Okada. A Noncooperative Coalitional Bargaining Game With Random Proposers. *Games and Economic Behavior*, 16:97–108, 1996.

[9] T. Sandholm, K. Larson, M. Andersson, O. Shehory, and F. Tohme. Coalition Structure Generation with Worst Case Guarantees. *Artificial Intelligence*, 111(1–2):209–238, 1999.

[10] T. Sandholm and V. Lesser. Coalitions Among Computationally Bounded Agents. *Artificial Intelligence*, 94(1-2), 1997.

[11] O. Shehory and S. Kraus. Methods for Task Allocation via Agent Coalition Formation. *Artificial Intelligence*, 101(1–2):165–200, 1998.

[12] O. Shehory and S. Kraus. Feasible Formation of Coalitions among Autonomous Agents in Nonsuperadditve Environments. *Computational Intelligence*, 15:218–251, 1999.

[13] J. Suijs and P. Borm. Stochastic cooperative games: superadditivity, convexity and certainty equivalents. *Journal of Games and Economic Behavior*, 27:331–345, 1999.

[14] J. Suijs, P. Borm, A. D. Wagenaere, and S. Tijs. Cooperative games with stochastic payoffs. *European Journal of Operational Research*, 113:193–205, 1999.