Sequential Decision Making in Repeated Coalition Formation under Uncertainty

Georgios Chalkiadakis
School of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, United Kingdom
gc2@ecs.soton.ac.uk

Craig Boutilier
Department of Computer Science
University of Toronto
Toronto, M5S 3G4, Canada
cebly@cs.toronto.edu

ABSTRACT

The problem of coalition formation when agents are uncertain about the *types* or *capabilities* of their potential partners is a critical one. In [3] a Bayesian reinforcement learning framework is developed for this problem when coalitions are formed (and tasks undertaken) repeatedly: not only does the model allow agents to refine their beliefs about the types of others, but uses value of information to define optimal exploration policies. However, computational approximations in that work are purely myopic. We present novel, non-myopic learning algorithms to approximate the optimal Bayesian solution, providing tractable means to ensure good sequential performance. We evaluate our algorithms in a variety of settings, and show that one, in particular, exhibits consistently good sequential performance. Further, it enables the Bayesian agents to transfer acquired knowledge among different dynamic tasks.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Economics

Keywords

coalition formation, uncertainty, reinforcement learning

1. INTRODUCTION

Coalition formation has recently attracted much attention in AI, allowing the dynamic formation of teams of cooperating agents. Most existing models of coalition formation assume the values of coalitions to be known with certainty, assuming that agents have knowledge of their potential partners' capabilities, or at least that this knowledge can be reached via communication. In many natural settings, however, rational agents must agree to coalitions and the division of the value generated without *a priori* knowledge of this value. For instance, agents are often uncertain about the *types* (or *capabilities*) of potential partners, hence how well they are suited to a particular tasks, hence the value

of the coalitions in which they participate. The case of an enterprise trying to choose subcontractors while unsure of their capabilities is only one such example.

It is often the case that a set of rational agents have to engage in *repeated* coalition formation, consisting of a series of coalition formation episodes, each of which is followed by some collective action taken by the coalitions so formed. This suggests opportunities to *learn* about others' abilities (types) through repeated interaction, refining over time how coalitions are formed. It also poses the question of how to make decisions that are *sequentially rational*, given anticipated *future interactions* and the evolution of an agent's knowledge. Further, by progressively gaining knowledge about the types of others, agents can *reuse* this knowledge when meeting these same individuals in different settings (with different coalitional values, but still dependent on the types of the agents). Therefore, agents should make future use of information they gain in this process.

Intuitively, the effects of *collective (coalitional)* actions provide information about the capabilities of partners. Agents can use this to inform their decisions regarding future coalition participation. To account for such considerations, we adopt the framework we introduced in [3]: we describe there a *reinforcement learning* (RL) model that enables agents to improve the quality of the coalitions formed (and tasks undertaken) using experience gained by repeated interaction with others and the observed effects of coalitional actions. Specifically, a *Bayesian RL* model is proposed, in which agents maintain explicit beliefs about the types of others.

Our framework in [3] makes use of a *partially observable Markov decision process (POMDP)* formulation similar to that used for multiagent RL in stochastic games [2]. This enables the agents to assess the long-term value of coalition formation decisions, including both the immediate value of potential collective actions within a specific coalition and the *value of information*: how what is learned about one's partners can influence *future* coalitional decisions. Agents take action stochasticity and type uncertainty into account, translate it into coalitional value uncertainty, and choose actions and coalitions not only for their immediate value, but also for their informational value.

Our main contribution in the current paper is the development of new, *non-myopic* exploration methods for repeated coalition formation under uncertainty in this framework. Our approximation methods do not require the full solution of the exploration POMDP, hence are quite tractable in practice. One algorithm, in particular, consistently and handily outperforms the myopic approximation investigated in [3] in a variety of settings, and is shown to successfully

allow for the easy *transfer of knowledge* among tasks.

The rest of the paper is structured as follows: Section 2 provides background on coalition formation and Bayesian RL; Section 3 describes the Bayesian RL framework for optimal repeated coalition formation under uncertainty [3]. Section 4 presents our main contributions, Bayesian RL algorithms to approximate the solution of the exploration POMDP. Section 5 details our experimental results, and Section 6 provides a discussion of related work.

2. BACKGROUND

In this section we provide some essential background on coalition formation and Bayesian reinforcement learning.

Coalition formation is one of the fundamental areas of study within cooperative game theory, which deals with situations where players act together in a cooperative equilibrium selection process involving some form of bargaining, negotiation, or arbitration [11]. 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 in order to solve a problem by combining their efforts. The coalition formation process involves three main activities [12]: (a) searching for an optimal coalition structure; (b) solving a joint problem facing members of each coalition; and (c) dividing the value of the generated solution among the coalition members.

We assume that a *characteristic function* $v : 2^N \to \Re$ defines the *value* $v(S)$ of each coalition S [11]. This $v(S)$ represents the maximal payoff the members of S can jointly receive by cooperating effectively. An *allocation* is a vector of payoffs (or *demands*) $\boldsymbol{d} = (d_1, ..., d_n)$ assigning some payoff to each $i \in N$.[1]

Since we adopt a Bayesian RL approach to learn the abilities of other agents, we now review some relevant work. Consider an agent 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. Pr refers to a family of transition distributions $\Pr(s, a, \cdot)$, and $\Pr(s, a, s')$ is the probability of reaching state s' after taking action a at s. The probability with which reward r is obtained when state s is reached is denoted $R(s, r)$. The agent has to construct an optimal Markovian policy $\pi : \mathcal{S} \mapsto \mathcal{A}$ maximizing the expected sum of future discounted rewards over an infinite horizon. 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 [1]. An RL agent does not have direct access to R or Pr, so it must learn a policy based on its interactions with the environment.

When *model-based RL* is used, the agent 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 this MDP can be solved (or approximated). Single-agent Bayesian methods [6] assume some prior density P over possible dynamics D and reward distributions R, which is updated with each data point $\langle s, a, r, t \rangle$, and allow agents to explore optimally. In a similar fashion, multi-agent Bayesian RL agents [2] update prior distributions over the space of possible strategies of others, as well as the space of possible MDP models. There are two components of the value of performing an action at a belief state: an expected value given the current belief state, and a value of the action's impact on the current belief state. The second component, in particular, captures the *expected value of information* (EVOI) of an action. Each action gives rise to some immediate response by the environment changing the agent's beliefs, and subsequent action choice and expected reward is influenced by this change. EVOI computation can be combined with "object-level" expected value via Bellman equations that describe the solution to the POMDP representing the exploration-exploitation problem of the agents. Experiments [6, 2] demonstrate the practical value of the Bayesian approach, which allows for exploration costs to be effectively weighed against their expected benefits. This leads to informed, intelligent exploration, and better online performance while learning than other exploration models.

3. A BAYESIAN RL FRAMEWORK

The Bayesian RL models described have been applied to the problem of repeated coalition formation under *type uncertainty* by Chalkiadakis and Boutilier [3]. We describe that model and approach in some detail here, since our main contributions build on this model.[2]

In realistic settings, agents participating in coalition formation activities will have to face type uncertainty and uncertainty regarding coalitional actions and their results. The possibility of repeated interaction provides the agents with the ability to *learn*, progressively updating their beliefs about the types of their potential partners. However, agents should not seek to reduce uncertainty *for its own sake* by employing crude exploration policies: this often leads to poor online performance [2]. Indeed, we generally expect that in the limit type uncertainty will remain regarding the capabilities of unpromising partners.

To address this, we recap the model of [3] for *optimal repeated coalition formation*. Agents are interested in eventually forming efficient, profitable coalitions, but they also want to gather as much reward as possible while doing so. *Optimal repeated coalition formation*, or *optimal coalitional learning*, aims to maximize the long-term performance of an agent that repeatedly engages in coalition formation activities and receives its share of payoffs arising from agreed-upon coalitional actions—as specified in agreements reached during the aforementioned coalitional activities.

Agents repeatedly form coalitions and take coalitional actions. This gives agents the opportunity to update their beliefs about the types of their partners by observing the results of coalitional actions. Belief updates using our RL formulation will in turn influence future coalition formation decisions, which will be taken in a manner that is *sequentially rational*. When using a Bayesian approach to repeated coalition formation, agents are often satisfied *not* to learn about the abilities of potential partners, if they expect the costs of doing so to outweigh the anticipated benefits.

---

[1] *Coalitional stability* (e.g., notions like the *core*) is a critical aspect of any theory of coalitional formation. We do not address stability in this paper, but note that a stability concept under the form of type uncertainty that we adopt below, the *Bayesian core*, has been introduced in [3, 5].

[2] The model in [3] also deals extensively with issues of coalitional stability and various processes for coalition formation itself. Our present focus is entirely on the RL problem.

A Bayesian Coalition Formation Model.

Assume a set of agents $N = \{1, \ldots, n\}$. For each agent i, T_i is a finite set of possible *types*. Each agent has a specific type $t \in T_i$, which intuitively captures its "abilities." Let $T = \times_{i \in N} T_i$ denote the set of type profiles. For any coalition $C \subseteq N$, $T_C = \times_{i \in C} T_i$, and for any $i \in N$, $T_{-i} = \times_{j \neq i} T_j$. Each agent is aware of its own type t_i, but not those of other agents. The *beliefs* B_i of agent i comprise a joint distribution over T_{-i}, and $B_i(\boldsymbol{t}_{-i})$ is the probability i assigns to its peers having type profile $\boldsymbol{t}_{-i}$; further, $B_i(\boldsymbol{t}_C)$ denotes the marginal of B_i over any coalition of agents C.

A finite set of *coalitional actions* A_C is available to a coalition C. When a coalitional action is taken, it results in some outcome $o \in \mathcal{O}$. The odds with which an outcome is realized depends on the types of the coalition members with $\Pr(o|\alpha, \boldsymbol{t}_C)$ being the probability of outcome o given that C takes action $\alpha \in A_C$ and member types are given by $\boldsymbol{t}_C \in T_C$ (e.g., the outcome of building a house will depend on the abilities of the team members). We also assume that each outcome o of a coalitional action results in some *reward* $R(o)$ assigned to the coalition, and is assumed to be divisible among its members.

The *value* of coalition C with member types $\boldsymbol{t}_C$ is:

$$V(C|\boldsymbol{t}_C) = \max_{\alpha \in A_C} \sum_o \Pr(o|\alpha, \boldsymbol{t}_C) R(o) = \max_{\alpha \in A_C} Q(C, \alpha|\boldsymbol{t}_C)$$

Unfortunately, this value cannot be used in the coalition formation process when the agents are uncertain about their potential partners' types. Nevertheless, each i has beliefs about the value of any coalition based on its expectation of this value with respect to other agents's types, and thus can translate type uncertainty into coalitional value uncertainty:

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

Here $V_i(C)$ is not simply the expectation of $V(C)$ w.r.t. i's belief about types. The expectation Q_i of coalitional action values cannot be moved outside the max operator, since a single coalitional action must be chosen which is useful *given* i's uncertainty regarding its partners (i.e., i might prefer different coalitional actions to be performed by a specific coalition when he holds different beliefs). Of course, i's estimate of the value of any coalitional action, may not be shared by other agents. Nevertheless, any i is certain of its own *reservation value*, the amount it can attain by acting in a singleton coalition: $rv_i = V_i(\{i\}) = \max_{\alpha \in A_{\{i\}}} \sum_o \Pr(o|\alpha, t_i) R(o)$.

Optimal Repeated Coalition Formation.

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. Once coalitions are formed, each $C \in CS^t$ takes its agreed upon action α_C^t and observes the resulting outcome o of that action, as explained above. Each agent in C then updates its beliefs:

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

where z is a normalizing constant. We often denote the updated belief state as $B_i^{o,\alpha}$. In order to make our model applicable to realistic circumstances, and in order to be able to test the full potential of our RL algorithms, we assume only *limited* observability of the realized outcomes: the agents observe only the outcome of their own coalition's action.

The process then repeats. Thus, the RL process consists of coalition formation games being played (the *coalition formation stage*) and the execution of coalitional actions and subsequent belief updating (the *RL stage*).

The approach to optimal repeated coalition formation of [3] uses *Bayesian exploration* [6, 2]. Bayesian agents are able to balance exploration with exploitation, realizing *sequential* performance that is optimal w.r.t. their beliefs. Bayesian exploration outperforms in expectation any other method having the same prior knowledge. In [3], the problem of optimal coalitional learning is cast as a POMDP, or a belief-state MDP. Assuming an infinite horizon problem, with discount factor γ (with $0 \leq \gamma < 1$), it is possible to formulate the optimality equations for the POMDP; however, because an agent is unaware of others' beliefs, certain subtleties arise.

We let $Q_i(C, \alpha, \boldsymbol{d}_C, B_i)$ denote the *long-term value* i places on being a member of coalition C that has agreed action α and a vector of demands $\boldsymbol{d}_C$: that is, the agent realizes that after this action is taken, the coalition formation process will repeat. Thus, $Q_i(C, \alpha, \boldsymbol{d}_C, B_i)$ represents the quality value of a *coalitional agreement* $\langle C, \alpha, \boldsymbol{d}_C \rangle$ under belief state B_i. This is defined using Bellman-like equations:

$$Q_i(C, \alpha, \boldsymbol{d}_C, B_i) = \sum_o \Pr(o|C, \alpha, B_i)[r_i R(o) + \gamma V_i(B_i^{o,\alpha})] \tag{1}$$

$$= \sum_{\boldsymbol{t}_C} B_i(\boldsymbol{t}_C) \sum_o \Pr(o|\alpha, \boldsymbol{t}_C)[r_i R(o) + \gamma V_i(B_i^{o,\alpha})]$$

$$V_i(B_i) = \sum_{C|i \in C, \boldsymbol{d}_C} \Pr(C, \alpha, \boldsymbol{d}_C|B_i) Q_i(C, \alpha, \boldsymbol{d}_C, B_i) \tag{2}$$

Recall that $R(o)$ is the immediate reward realized by C for its action resulting to outcome o, and r_i is the relative demand $r_i = \frac{d_i}{\sum_{j \in C} d_j}$ of agent i given demand vector $\boldsymbol{d}_C$ (and thus $r_i R(o)$ describes i's reward share when coalitional action α results in o). $V_i(B_i)$ describes the value of belief state B_i to i, deriving from the fact that while in B_i, agent i may find itself participating in any of a number of possible agreements, each of which has some Q-value.

The agents' uncertainty is effectively encapsulated in the belief-state MDP described by Eqs. 1 and 2; also, the expected *value of information* of a coalitional agreement is naturally captured in these equations since the value equations include the value of possible future belief states $B_i^{o,\alpha}$ that result from current partnerships, thus reflecting new, more refined beliefs about certain potential partners and how it will impact future decisions. Agents enter in the negotiation (coalition formation) process using *Q-values* to value coalitional agreements rather than immediate expected reward estimates: that is, they incorporate the *long-term* value of their decisions in this repeated coalition formation environment. The optimal course of action for the agents, then, is to act greedily with respect to their Q-value function.[3]

The value function V_i above cannot be defined by maximizing Q-values, unlike typical Bellman-like equations. This is because agent i does not have complete control over the choice that dictates reward (i.e., the coalition that is formed). The agent must instead predict the probability $\Pr(C, \alpha, \boldsymbol{d}_C|B_i)$ with which a specific coalitional agreement $\langle C, \alpha, \boldsymbol{d}_C \rangle$ in

---

[3] We do not deal with negotiation processes in this paper, but refer to [3, 4] for such. The framework enables the agents to use the estimated Q-values of coalition formation decisions within *any* bargaining process employed during formation.

which it participates 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, they account for how i's beliefs will change in the future when deciding how useful a specific coalition is now. The sequential value of any coalitional agreement (and action), accounting for its value of information, is used in the formation process, as explained above.

The terms $\Pr(C, \alpha, \boldsymbol{d}_C | B_i)$ can be estimated in a variety of ways, depending on the coalition formation algorithm uses during the formation stage. However, it is particularly challenging in realistic environments where agents do not have full knowledge of all parameters of the negotiation process, and/or they have limited observability of the environment (as is our case). Since agents can observe the outcome of their *own coalition's action only*, it is not possible for them to monitor the way the beliefs of others are changing, and this affects their capability to estimate the $\Pr(C, \alpha, \boldsymbol{d}_C | B_i)$ probabilities. We return to this issue in the next section. We note however that one of our non-myopic algorithms allows the agents to sidestep this difficulty.

4. COMPUTATIONAL APPROXIMATIONS

The calculation of an exact solution to the repeated coalition formation problem using the Bayesian RL formulation of Eqs. 1 and 2 is generally infeasible. The solution of POMDPs is generally quite difficult: and our state space and action space grow exponentially with the number of agents. Thus computational approximations are needed. A simple myopic approach is developed in [3] (which we describe below). We describe several new algorithms that tackle this POMDP. These Bayesian RL algorithms can be combined with any underlying negotiation process: the agents evaluate any potential agreement $\langle C, \alpha, \boldsymbol{d}_C \rangle$ that may arise as a result of a negotiation, and use these valuations to enter negotiations that may be governed by any rules.[4] In contrast to the simple myopic approach, most of the algorithms we develop attempt to take the sequential value of decisions into account.

One-Step Lookahead Algorithm.

We first present a *one-step lookahead (OSLA)* algorithm, which deals only with immediate successor belief states following coalition action α and resulting outcome state s. The motivation for this is that computing a value for every possible belief state in order to solve the belief-state MDP is generally impractical (and it is further complicated by the fact that an agent is not in complete control of the choices that dictate reward). It is possible to *approximately* calculate the value of the belief states that can follow the execution of a coalitional action (and the subsequent observed outcome) under the current agreement. When employing the OSLA method, $V_i(B_i^{o,\alpha})$ in Eq. 1, the value of a successor belief state will be calculated *myopically*.

Specifically, we define the *1-step lookahead* Q-value of a

---

[4]Of course, the agents may need to take the specific set of rules into account when evaluating the various agreements.

$\langle C, \alpha, \boldsymbol{d}_C \rangle$ agreement for i, under belief state B_i, to be

$$Q_i^1(C, \alpha, \boldsymbol{d}_C, B_i) = \sum_o \Pr(o|C, \alpha, B_i)[r_i R(o) + \gamma V_i^0(B_i^{s,\alpha})]$$

(3)

$$= \sum_{\boldsymbol{t}_C} B_i(\boldsymbol{t}_C) \sum_o \Pr(o|\alpha, \boldsymbol{t}_C)[r_i R(o) + \gamma V_i^0(B_i')]$$

(where r_i is i's relative demand given $\boldsymbol{d}_C$). In this equation, $V_i^0(B_i')$ represents the myopic ("0-step" lookahead) value of successor belief state B_i', which can be calculated using the *0-step* (myopically calculated) Q-values under some B_i' as:

$$V_i^0(B_i') = \sum_{\beta \in A_{C'}, \boldsymbol{d}_{C'} | i \in C'} \Pr(C', \beta, \boldsymbol{d}_{C'} | B_i') Q_i^0(C', \beta, \boldsymbol{d}_{C'}, B_i')$$

(4)

$$Q_i^0(C', \beta, \boldsymbol{d}_{C'}, B_i') = r_i' \sum_{\boldsymbol{t}_{C'} \in T_{C'}} B_i'(\boldsymbol{t}_{C'}) \sum_{o'} \Pr(o'|\beta, \boldsymbol{t}_{C'}) R(o')$$

(5)

where r_i' is i's relative demand given $\boldsymbol{d}_{C'}$, and Q^0 values are calculated accounting only for the expected immediate reward of C' (with agreed $\boldsymbol{d}_{C'}$) for taking β under B_i'.

To approximate $\Pr(C', \beta, \boldsymbol{d}_{C'} | B_i')$, we view it as the probability of reaching a specific agreement $\langle C', \beta, \boldsymbol{d}_{C'} \rangle$ after one negotiation step rather than after a whole negotiation process. In other words, we apply a *lookahead bound $l = 1$* on the size of the underlying bargaining game tree. Specifically, an agent who wants to estimate the 1-step Q-value of (every potential) agreement $\langle C, \alpha, \boldsymbol{d}_C \rangle$ at some negotiation step must calculate (as in Eq. 4) $\Pr(C', \beta, \boldsymbol{d}_{C'} | B_i')$ for each B_i' reached after the execution of α and possible observation o. However, we assume that the agent cares only for agreements that are reachable within one negotiation step after "fixing" his beliefs to B_i'. We use a lookahead of 1 when solving the game tree in our experiments for efficiency. However, this tree-size lookahead bound could take any value of $l \geq 1$, depending on specific requirements. (An additional issue complicating the solution of the game tree is that the "common prior assumption" typically used when solving Bayesian games is not valid in our setting. We defer discussion of this issue to an extended version of this paper.) Two more computational difficulties arise when one tries to sum over all possible $\boldsymbol{t}_C$ in Eqs. 3 and 5, and over all possible formation moves (choice of coalition, action and demands) in Eq. 4 above. Nevertheless, sampling and appropriate discretization of demands can help alleviate these problems.

In summary, thus, the OSLA method proceeds as follows:

1. At the beginning of each RL stage, each agent i with belief state B_i calculates the 1-step Q-value Q_i^1 of any potential agreement $\langle C, \alpha, \boldsymbol{d}_C \rangle$, using Eqs. 3, 4 and 5. The $\Pr(C', \beta, \boldsymbol{d}_{C'} | B_i')$ in Eq. 4 are derived for each potential successor belief state B_i' by each agent solving a "fixed beliefs" game describing the anticipated negotiations, assuming a game tree size of l.

2. The calculated Q_i^1 values are then used by i in the subsequent coalition formation process.

VPI Exploration Method.

Here we recast the ideas of *VPI exploration* [7, 6, 2] to the repeated coalition formation setting, and propose a VPI exploration method that estimates the (myopic) value of obtaining perfect information about a coalitional agreement

given current beliefs. The sequential value of any coalitional action, accounting for its value of information, is then used in the formation process.

Though the basic idea of our algorithm is as in [7, 6, 2], it differs in that we do not sample over the space of MDP models, but rather over the space of types configurations. In addition, the actions for which VPI is calculated are the set of possible coalitional agreements instead of the actions of a single agent. Finally, we propose a way to combine our VPI technique with the OSLA technique introduced above.

To begin, consider what can be gained by learning the true value of *some* coalitional agreement $\sigma = \langle C, \alpha, \boldsymbol{d}_C \rangle$. Suppose σ is adopted and the corresponding action α is executed, and assume that it leads to specific *exact evidence* regarding the types of the agents in C (i.e., we assume the true type vector $\boldsymbol{t}_C^*$ is revealed following σ). In this way, the *true value* of σ is also revealed, and it can be defined as the share of the "true" coalitional agreement value that i gets; denote this by $q_\sigma^* = q_{\langle C, \alpha, \boldsymbol{d}_C \rangle}^* = Q_i(C, \alpha, \boldsymbol{d}_C | \boldsymbol{t}_C^*)$, with

$$Q_i(C, \alpha, \boldsymbol{d}_C | \boldsymbol{t}_C^*) = r_i \sum_o \Pr(o | \alpha, \boldsymbol{t}_C^*) R(o) \qquad (6)$$

where r_i is i's relative demand given $\boldsymbol{d}_C$. This is a "myopic" calculation of the specific (future) coalitional agreement value, assuming the definite adoption of this agreement, and the subsequent revelation of their true types.

This new knowledge is of interest only if it leads the agent to change its decision as to what strategy to follow. Specifically, there are two ways to take advantage of this new, "perfect" information.

First, suppose that under the current belief state B_i the value of i's current best action $\sigma_1 = \langle C_1, \alpha_1, \boldsymbol{d}_{C_1} \rangle$ is $q_1 = Q_i(C_1, \alpha_1, \boldsymbol{d}_{C1} | B_i)$, the expected value given belief state. If the new knowledge indicates that σ is a better action (i.e., $q_\sigma^* > q_1$), i should prefer σ to σ_1, gaining $q_\sigma^* - q_1$. Second, suppose that the value of the second best action $\sigma_2 = \langle C_2, \alpha_2, \boldsymbol{d}_{C2} \rangle$ is $q_2 = Q_i(C_2, \alpha_2, \boldsymbol{d}_{C2} | B_i)$. If action σ coincides with the action considered best, σ_1, and the new knowledge indicates that the real value $q_{\sigma_1}^* = q_\sigma^*$ is less than the value of the previously considered second-best action, then the agent should prefer σ_2 to σ_1, gaining $q_2 - q_{\sigma_1}^*$.

Thus, the *gain* from learning the true value q_σ^* of σ is:

$$gain_\sigma(q_\sigma^* | \boldsymbol{t}_C^*) = \begin{cases} q_\sigma^* - q_1, & \text{if } \sigma \neq \sigma_1 \text{ and } q_\sigma^* > q_1 \\ q_2 - q_\sigma^*, & \text{if } \sigma = \sigma_1 \text{ and } q_\sigma^* < q_2 \\ 0, & \text{otherwise} \end{cases} \qquad (7)$$

However, the agent does not know in advance what types (and, consequently, which Q-value) will be revealed for σ; therefore, it needs to take into account the expected gain given its prior beliefs. Hence, it computes the *expected value of perfect information* about σ as:

$$EVPI(\sigma | B_i) = \sum_{\boldsymbol{t}_C^*} gain_\sigma(q_\sigma^* | \boldsymbol{t}_C^*) B_i(\boldsymbol{t}_C^*) \qquad (8)$$

This expected value of perfect information (VPI) represents the expected gain deriving from learning the true value of coalitional agreement σ and it treated as a surrogate for the the value of "exploring" σ. Hence, the value of σ taking into account this expected gain is defined as:

$$QV_i(\sigma | B_i) = Q_i(\sigma | B_i) + EVPI(\sigma | B_i) \qquad (9)$$

and agents taking VPI into account should have a preference for agreements maximizing this equation. The agents should

then use these QV values instead of using the usual Q-value quantities in their decision making for forming coalitions. The calculation of expected values and VPI above can be done in a straightforward manner if the number of possible type configurations is small. If, however, this number is too large, sampling has to be employed.

In summary, the VPI algorithm proceeds as follows:

1. The "true" Q-values of any potential agreement σ are myopically calculated via Eq. 6.

2. The gain from reaching σ is calculated via Eq. 7.

3. The VPI for agreement σ is calculated via Eq. 8.

4. The Q-values QV_i for (any) σ are calculated through Eq. 9 (and are then used in the formation process).

VPI exploration is a non-myopic method, since it reasons about the value of future belief states, accounting for the expected VPI of coalitional agreements and its impact on the future decisions. Notice, however, that the VPI algorithm still uses myopic calculations when determining the value of agreements. Even though this is an approximation, it enables the method to focus on exploiting the value of (perfect) information regarding the types, however myopic the estimation of this value may be. This stands in contrast to lookahead methods which attempt to estimate the value of specific coalitional actions. Thus, unlike lookahead, the VPI algorithm does not have to explicitly incorporate the common prior hypothesis in the calculation of the Q-values to be used during formation—and does not need to account for the probability of agreement when transitioning to future belief states. The VPI exploration method is thus not tightly tied to the specific formation process used. This myopic VPI estimation proves to work well in a variety of experimental settings.

Nevertheless, for interest, we also developed and tested a method which combines VPI with OSLA. This *VPI-over-OSLA* method uses the application of VPI over Q-values estimated using the OSLA method. When this method is used, the values of currently expected best action, second best action and exploratory action σ are estimated using one-step lookahead (and, thus, there is a need to approximate the probabilities of future agreements in this case).

Maximum A Posteriori Type Assignment RL Algorithm.

A *maximum a posteriori (MAP) type assignment* algorithm can also be defined: Given a belief state B_i, agent i assumes that the type t_j^i of an opponent j is the one specified as the most probable by $B_i(t_j)$: that is, $t_j^i = argmax_{t_j} B_i(t_j)$. Thus, a vector of types $\boldsymbol{t}_C$ assumed by i to represent the true types of partners' (in any coalition C) can be defined, and agent i calculates Q-values as

$$Q_i(C, \alpha, \boldsymbol{d}_C | \boldsymbol{t}_C) = r_i \sum_o \Pr(o | \alpha, \boldsymbol{t}_C) R(o)$$

Notice that this calculation is myopic, not accounting for the sequential value of an agreement.

Myopic Bayesian RL Algorithm.

Finally, a myopic Bayesian RL algorithm was defined in the obvious way in [3]: the agents do not reason about future belief states, but rather just myopically assess the value of various coalitional moves, apply an inner coalition formation

process, and repeat. A myopic agent i calculates the value of agreement $\langle C, \alpha, \boldsymbol{d}_C \rangle$ under belief state B_i as:

$$Q_i(C, \alpha, \boldsymbol{d}_C, B_i) = r_i \sum_{\boldsymbol{t}_C \in T_C} B_i(\boldsymbol{t}_C) \sum_s \Pr(o|\alpha, \boldsymbol{t}_C) R(o)$$

5. EXPERIMENTAL EVALUATION

We conducted several types of experiments to evaluate our methods. For space reasons, however, we only present two in this paper. In the first set of experiments, we compare our methods to each other by requiring the agents to face the same coalition formation problem repeatedly. In the second, our agents act in a *dynamic* environment, in which potentially different tasks arise at each stage. This setting demonstrates that our approach allows for the *transfer of knowledge* between different tasks. Further, it helps demonstrate the benefits of using the VPI method in particular.

Our agents can observe the results of the action taken by the coalition to which they belong, but not those of any other coalition. Thus, they can only update their beliefs regarding their own partners at any stage. The coalition structure in place at the beginning of each RL step is the result of the preceding formation process. In all settings, the experiments are run in homogeneous environments (i.e., all agents employ the same algorithm). The main metric we use in all our experiments is discounted reward accumulated by the coalitions—this reflects the sequential rationality of agent decisions. The negotiation process used during the formation stages has 50 steps (per RL stage), and is the *best reply with experimentation (BRE)* dynamic process introduced in [3]. For sampling type vectors, we used the following approach: if $|T|^{|C|} \le 1000$, where $|T|$ is the number of types and $|C|$ is the size of coalition C, no sampling was used; otherwise, the sampling size was set to 100. Finally, each experiment consists of 30 runs, and each run employs 500 RL steps, and the discount factor used was 0.985.

Our first setting has 10 agents, with 10 possible types per agent, 3 actions per coalition, and 3 outcomes per action. The agents form companies to bid for software projects. There are 3 "major" types (project roles) for the agents, each with 3 or 4 "quality" types: *interface designer* $= \langle$ bad, average, expert $\rangle$, *programmer* $= \langle$ bad, average, good, expert $\rangle$ and *systems engineer* $= \langle$ bad, average, expert $\rangle$. The latter correspond to quality "points" (0 points for "bad" types and then increasing by 1), which, when summed, characterize the overall quality of a coalition. The agents know the major type of their opponents, but not their quality types. The companies bid for large, average-sized or small projects (actions), and expect to make large, average or small profit (outcomes). The outcome (and subsequent reward) of a coalitional action depends on the coalition's quality and the action performed. In general, bidding for large projects is unlikely to be rewarding: a coalition will be unable to receive large profits by doing so unless its overall quality is high *and* there is enough diversity (regarding major types) among its members. Also, it is to be expected that myopic agents will find it hard to form size 2 coalitions (starting from a configuration structure of singletons), even if these coalitions can serve as the "building blocks" for more promising ones. Due to space limitations, we omit further details.

We run our agents in two types of settings: one with a common prior that was uniform with respect to the quality types of opponents, and one with a misinformed common prior—in this case the agent has a belief of 0.7 that

each of its opponents is of a quality type other than its real type. The results, in terms of average discounted accumulated coalitional reward, are shown in Fig. 1. To have a comparison metric against some form of "optimal" behaviour of the agents, we also tested the behaviour of agents who were *fully informed* regarding each others' types (using a common prior that accurately depicted the assignment of types to agents); we do not show the plot in order not to congest the figure, but their discounted average (over 30 runs) accumulated payoff after 500 iterations was 258726.

The results clearly show that VPI is the most successful of the methods. It managed to accumulate 76.6% of the average discounted rewards accrued by fully informed agents in the misinformed priors case (and 67% in the uniform priors case), with other methods not exceeding 51.7%. One important observation is that the VPI method manages to achieve good performance without, in most cases, significantly reducing the agents' uncertainty regarding the true type of its partners. This observation reinforces the point that it is not always necessary for agents to seek to forcefully reduce uncertainty (e.g., via uninformed exploration) in order to achieve good performance.

The MAP method also does quite well in these experiments. It effectively employs "crude" exploration, with agents behaving "greedily" towards the value of information they receive (slight modification of beliefs may "point" to a different type for a partner to be taken for granted). This turned out to be helpful here, assuming major types which are known to agents, with only 3 or 4 unknown quality types each, and with a reward signal that can in fact be quite clear regarding the quality of coalitions. However, such conditions are unlikely to hold in realistic environments.

On the other hand, the performance of OSLA and VPI-over-OSLA in terms of discounted accumulated reward is, in general, poor. We attribute this to the fact that OSLA cannot successfully approximate $\Pr(C', \beta, \boldsymbol{d}_{C'}|B_i')$. Notably, however, VPI-over-OSLA achieves better performance than OSLA. (We also note that the reward-gathering performance of OSLA and VPI-over-OSLA in the final stages of the experiments, is in general comparable to that of methods that fare better in terms of discounted accumulated reward.)

The experiment above had agents facing the same coalition formation problem—with the same transition probabilities and outcomes—at each RL step. This is analogous to having the agents facing *static* tasks—the same set of tasks needs to be served at each time point—in distinction to facing *dynamic* tasks. One challenge in dynamic situations is for agents to discover the type of their opponents. The need to achieve this goal is more pressing in this case, since they will have to put their beliefs to test facing different situations each time.

The setting we now present tests the abilities of our agents to achieve *transfer of knowledge* between tasks; this is one of the benefits of using learning to address type uncertainty: once agents learn about the abilities of actual and potential partners, they can re-use this knowledge when encountering those partners again under different circumstances. Here we assume that the agents *do not* know in advance which tasks they are going to face in the next RL step. In other words, the agents do not know in advance which transition to outcome states model prevails in the next RL step (they only know the model in the current RL step). This makes the environment truly dynamic. The POMDP assumptions do not now hold, due the non-stationarity of the environ-
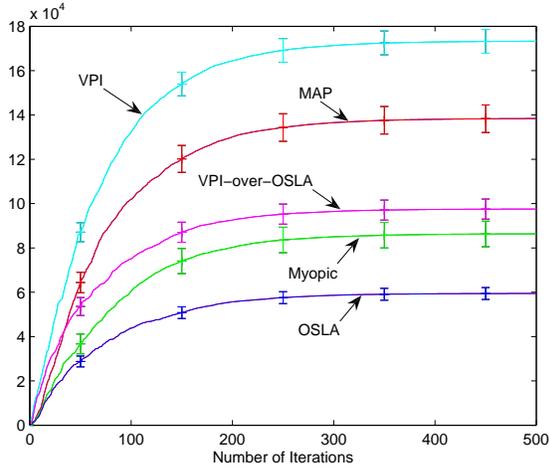
(a) Uniform priors
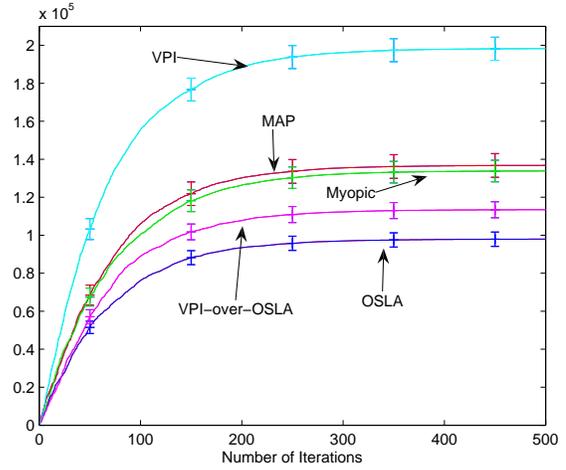
(b) Misinformed priors

Figure 1: First setting. Discounted average total payoff accumulated by coalitions (in 30 runs). Error bars represent 95% confidence intervals.

ment. Therefore, lookahead methods should do poorly here, since agents employing them would wrongfully anticipate to encounter specific tasks again in the near future.

In our setting, 5 agents co-exist in a homogeneous environment and form coalitions over a period of 500 RL steps. Each one of the agents is assigned with one of 5 different types (so that the agents are of different types). The agents share a uniform common prior regarding the types of opponents. At each RL step, the formed coalitions have 3 coalitional actions at their disposal, with 3 possible outcome states per action. We assume that five bandits in the Wild West face when trying to form a successful gang. Specifically, the "Good", the "Bad" and the "Ugly" (agents' types) have to discover each other and come together in order to "Rob the Train" (coalitional action), so as to get the "Big Money" (outcome state). In order to do so, they will go through an experience-gathering phase, during which it is possible to coalesce with other villains, performing "petit crime" actions of lesser significance given the coalition qualities and underlying stochasticity. So, during the first 400 RL steps, the agents are faced with two "petit-crime" problems, 1 and 2, alternatively (with a distinct outcome transition model for each, unknown to the agents beforehand) while they face problem 3 during the last 100 RL steps (the "Big Crime" phase). By the time RL step 401 is reached, they should have gained enough experience in order to tackle problem 3 (through identifying each other correctly) and fare well in their "Big Crime" activities, or else they are going to be making only "Some Change" during most of the last 100 RL steps. Specifically, if during problem 3, all three of them form a coalition *and* decide to rob the train, they have 85% probability of making Big Money. The setup of problems 1 and 2, in contrast, is such that it urges the agents to form two-agent coalitions, so that they get information regarding their partners' types.

It is obvious from the results of Figure 2 that VPI dominates the other methods in terms of discounted accumulated rewards (i.e., behaviour during the "experience-gathering" phase); it is also dominant in terms of accumulated rewards during the final stage of the experiment (154572 compared to 66051.3 for Myopic, 5651.13 for OSLA and 3087.73 for VPI-
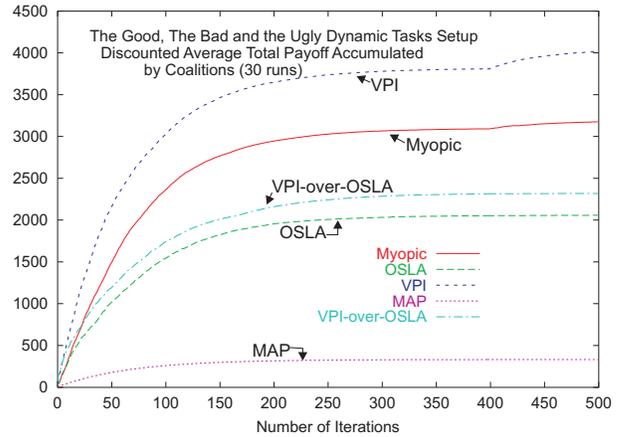
Figure 2: The Good, the Bad and the Ugly.

over-OSLA). This is because it uses myopic VPI regarding the types of partners, and does *not* deal intrinsically with the expected utility of future anticipated coalitional actions in subsequent belief states (the VPI method is not tightly linked to the "internal" coalition formation process used). In contrast, the lookahead methods' behaviour was much poorer, as expected. Nevertheless, OSLA and VPI-over-OSLA agents do manage to collect, during the last phase, approximately 10 and 6 times more reward, respectively, than the MAP agents, who have been utterly confused by the setup. We also note that in a similar setting—where agents knew beforehand the set of dynamic tasks to be faced in the next step—OSLA and VPI-over-OSLA did a lot better, with VPI-over-OSLA outperforming Myopic and OSLA.

To conclude, we note that apart from consistently outperforming the other methods, VPI is also a quite scalable method, whose worst case running time (for an *entire* run, including the negotiation phases, in the 10-agent setting) is on the order of 700s. This makes the method more attractive for realistic settings. By comparison, OSLA can exhibit running time of the order of >25 hours per run (in the 10-agent setting). No parallelism was used in those experiments; however, the autonomous agents can be assumed

to be performing their calculations in parallel, which could mean that the total time for the experiments would be reduced by a factor close to the number of agents.

6. RELATED WORK AND DISCUSSION

Much of the coalition formation related work done in AI is motivated by the need to serve tasks requiring the utilization of various resources found among a collection of agents. We note that resources can be readily incorporated in our model, since they can be thought of as determining the agent types.

Shehory and Kraus [13, 14] present coalition formation algorithms which take into account the capabilities of the various agents. However, the agents rely on information communicated to them by their potential partners in order to form their estimation of others' capabilities. This is the case also for the approach of [15], which utilizes coalition formation algorithms in order to achieve collaboration of agents so that tasks of common interest are served (but they do not deal with individual rationality).

Kraus, Shehory and Taase [8, 9] proposed a heuristic coalition formation method in the "Request for Proposal" domain, under a restricted form of coalitional value uncertainty. Again, the focus of their work is on social-welfare maximization rather than individual rationality. Furthermore, no learning is involved, since they do not tackle repeated coalition formation: a formed coalition "walks away" from negotiations and it cannot be decomposed. Campos and Willmott on the other hand, do bring repeated coalition formation into the picture in [10]. They tackle "iterative coalition games" that may involve up to 100 agents to fulfill a task which does not change over time. The agents use *pre-described* strategies (such as tit-for-tat) for joining or abandoning partners—thus, the approach is basically static.

In distinction to the aforementioned approaches, our framework enables agents to not only dynamically choose the tasks they wish to deal with, but also to choose the proper way (action) to deal with them. Task execution is readily incorporated in our model by simply viewing the tasks as entailing the use of specific action sets. Finally, our approach potentially enables the agents to form the most suitable coalitions for a new problem "online"—in the sense that knowledge acquired during executing one task is readily "transferable" to another through the estimation of the types (capabilities) of partners. Thus, the agents do not have to experience dealing with a new specific problem for some time period before deciding on ways to attack it.

7. CONCLUSIONS

In this paper we presented novel non-myopic algorithms to approximate the optimal Bayesian solution to the coalitional exploration-exploitation problem, weighing the need to explore the abilities of potential partners with the need to exploit acquired knowledge. We made use of the Bayesian multiagent RL framework for repeated coalition formation under type uncertainty and reward stochasticity originally proposed in [3]: this framework facilitates the sequential decision making of the agents through experience gathered by observation of the effects of coalitional actions. We further verified the flexibility and generality of this framework by exposing its ability to incorporate a variety of RL algorithms and prior beliefs, and demonstrating experimentally how it enables the agents to dynamically form coalitions, serve tasks and transfer knowledge among them.

Apart from verifying the overall effectiveness of our approach, our experiments also indicate that our Bayesian coalitional VPI method, in particular, is the most successful of our methods. We have recently applied these methods in larger, more realistic settings; specifically, we have seen very encouraging results when applying these ideas to the computational trust domain [16].

8. REFERENCES

[1] C. Boutilier, T. Dean, and S. Hanks. Decision theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.

[2] G. Chalkiadakis and C. Boutilier. Coordination in Multiagent Reinforcement Learning: A Bayesian Approach. In *AAMAS'03*, pages 709–716, 2003.

[3] G. Chalkiadakis and C. Boutilier. Bayesian Reinforcement Learning for Coalition Formation Under Uncertainty. In *AAMAS'04*, pages 1090–1097, 2004.

[4] G. Chalkiadakis and C. Boutilier. Coalitional Bargaining with Agent Type Uncertainty. In *IJCAI-07*, pages 1227–1232, 2007.

[5] G. Chalkiadakis, E. Markakis, and C. Boutilier. Coalition Formation under Uncertainty: Bargaining Equilibria and the Bayesian Core Stability Concept. In *AAMAS'07*, pages 400–407, 2007.

[6] R. Dearden, N. Friedman, and D. Andre. Model based Bayesian Exploration. In *UAI'99*, pages 150–159, 1999.

[7] R. Dearden, N. Friedman, and S. Russell. Bayesian Q-Learning. In *AAAI-98*, pages 761–768, 1998.

[8] S. Kraus, O. Shehory, and G. Taase. Coalition Formation with Uncertain Heterogeneous Information. In *AAMAS'03*, pages 1–8, 2003.

[9] S. Kraus, O. Shehory, and G. Taase. The Advantages of Compromising in Coalition Formation with Incomplete Information. In *AAMAS'04*, pages 588–595, 2004.

[10] C. Merida-Campos and S. Willmott. Modelling Coalition Formation Over Time for Iterative Coalition Games. In *AAMAS'04*, pages 572–579, 2004.

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

[12] T. Sandholm and V. Lesser. Coalitions Among Computationally Bounded Agents. *Artificial Intelligence*, 94(1):99 – 137, 1997.

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

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

[15] O. Shehory, K. Sycara, and S. Jha. Multiagent Coordination through Coalition Formation. In *Agent Theories, Architectures and Languages (ATAL'97)*, pages 143–154, 1997.

[16] W. T. L. Teacy, G. Chalkiadakis, A. Rogers, and N. R. Jennings. Sequential Decision Making with Untrustworthy Service Providers. In *AAMAS'08*, 2008.