A Bayesian Approach to Multiagent Reinforcement Learning and Coalition Formation under Uncertainty

by

Georgios Chalkiadakis

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

A Bayesian Approach to Multiagent Reinforcement Learning and Coalition Formation under Uncertainty

Georgios Chalkiadakis
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2007

Sequential decision making under uncertainty is always a challenge for autonomous agents populating a multiagent environment, since their behaviour is inevitably influenced by the behaviour of others. Further, agents have to constantly struggle to find the right balance between *exploiting* current information regarding the environment and the rest of its inhabitants, and *exploring* so that they acquire additional information. Moreover, they need to profitably trade off short-term rewards with anticipated long-term ones, while learning through interaction about the environment and others—employing techniques from *reinforcement learning (RL)*, a fundamental area of study within artificial intelligence (AI).

Coalition formation is a problem of great interest within game theory and AI, allowing autonomous individually rational agents to form stable or transient teams (or *coalitions*) to tackle an underlying task. Agents participating in realistic scenarios of *repeated coalition formation under uncertainty* face the issues identified above, and need to bargain to succesfully negotiate the terms of their participation in coalitions—often having to compromise individual with team welfare effectively.

In this thesis, we provide theoretical and algorithmic tools to accommodate *sequential decision making under uncertainty in multiagent settings*, dealing with the issues above. Specifically, we combine *multiagent Bayesian RL* with game theoretic ideas to facilitate the agents' sequential decision making. We deal with popular multiagent problems which were to date not tackled under uncertainty, or more specifically under *type uncertainty*. In our work, we assume

that the environment dynamics or the types (capabilities) of other agents are not known, and thus the agents have to account for this uncertainty, in a Bayesian way, when making decisions. Handling type uncertainty allows information about others acquired within one setting to be exploited in possibly different settings in the future.

The core of our contributions lies in the area of coalition formation under uncertainty. We studied several aspects of both the cooperative and non-cooperative facets of this problem, coining new theoretical concepts, proving theoretical results, presenting and evaluating algorithms for use in this context, and proposing a Bayesian RL framework for optimal repeated coalition formation under uncertainty.

Acknowledgements

I would like to acknowledge the contributions of many people who helped me directly and indirectly to complete this work. First and foremost, I am hugely indebted to a great scientist, my supervisor and mentor Craig Boutilier, who guided and supported me throughout the PhD process, always providing me with insightful advice. He has undoubtedly shaped me as a researcher. Many thanks also go to my external thesis appraiser and examiner, Katia Sycara, and to the members of my thesis committee, Sam Roweis, Grigoris Karakoulas and Peter Marbach for their helpful comments.

I would also like to thank Bob Price, for his invaluable help during the first years of my thesis; also, Pascal Poupart and Relu Patrascu, for many helpful discussions. Further, I would like to thank Vangelis Markakis, for all his help and discussions, and for being a good friend. Moreover, I would like to thank all the good friends and colleagues that made my years in Toronto truly unforgetable. Among them, I would like to give special thanks to Theophanis Tsandilas, for being a great friend throughout all these years.

I also feel obliged to commemorate here the name of my MSc supervisor in the University of Crete, the late Stelios Orphanoudakis, a visionary scientist and an inspiring man, without whom I most probably wouldn't have turned to research. Finally, I would like to thank my parents, Charalampos and Maria, for their love and support. This thesis is dedicated to them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Learning through interaction with the environment is a fundamental idea underlying many theories of learning and intelligence. *Reinforcement Learning (RL)* [KLM96, SB98], from a computer science point of view, is the problem that an agent faces when trying to learn to act by trial and error through interactions with a dynamic environment. This is different from *supervised learning*, which is learning from examples provided by an external supervisor; it also contrasts with classical planning, in that agents do not know *a priori* how their actions will affect the world. An RL agent is a goal-seeking agent that can sense aspects of its environment and can choose actions to influence this environment, so as to maximize a numerical reward signal. Reinforcement learning is a highly active research area, and is strongly connected not only with artificial intelligence (AI) research—such as machine learning, planning and neural networks—but neuroscience, psychology, and statistics, as well.

The reinforcement learning problem becomes even more interesting when studied in the context of multiple reinforcement learning agents co-existing in the same environment. This situation constitutes the *multiagent reinforcement learning (MARL)* problem. The presence of multiple RL agents creates new opportunities for them to seize or obstacles to overcome so that they enhance their learning capabilities. It also raises important questions[1] regarding the value of employing "social behaviours" like cooperation and coordination. Agents also face questions such as whether it is worth sacrificing short-term rewards in anticipation of long-term ones, whether it is valuable to them—under certain circumstances or conditions—to form teams, and whether and in which ways their own behaviour will influence the behaviour of others (and in which way these changes will affect them in the short or long term). Learning in

[1] Sometimes, even questions of philosophical nature (see, e.g., [Axe84]).

a multiagent world of this kind can be viewed as learning in a game with multiple players.

This chapter underscores the motivations and questions which lead us to explore research at the borderline of game theory and reinforcement learning It also briefly outlines the solutions and answers we provide to the questions explored in this thesis, and highlights the main research contributions of our work.

1.1 Motivation: Sequential Decision Making in Uncertain Multiagent Environments

In realistic settings, agents have to repeatedly make decisions over time, having only incomplete information about the environment and the other agents present. In a multiagent world, others' actions influence one's own decisions and actions; furthermore, when the agents are faced with the question of forming teams or coalitions, the issue of finding the right balance in sacrificing individual welfare for team welfare or vice-versa is important (*individual* vs. *team* rationality). Taking all those issues into consideration, other interesting questions then arise: *What does it mean to "act optimally" in such an environment? How do we make (a series of) decisions that are beneficial both in a short and in a long term?* In other words, *how can an agent in a multiagent world make sequentially optimal decisions under incomplete information?*

Game Theory and Reinforcement Learning

Game theory is the study of strategic interactions between multiple intelligent *rational* decision-makers—rationality meaning that the agents (or, in a game setting, "players")[2] consistently pursue their own objectives, trying to maximize the expected value of their own payoffs, which is measured in some *utility* scale [Mye91]. Being the theory of strategic interactions between multiple goal-driven agents, game theory provides many of the tools for dealing with situations and questions such as those mentioned above. The importance of game theory is evident in the fact that it is now widely applied in various fields, such as economics, biology, political science, social psychology, sociology and anthropology [Gin00]. In addition, a part of game theory deals with *learning in games*. Learning in games involves modeling the processes by which players change the strategies they are using to play a game over time. The very notion

[2] We will be using the terms "agents" and "players" interchangeably in this thesis.

of *equilibrium* can be considered to be the long-run result of a process where rational players try to learn and play optimally over time [FL98].

Of course, as the critics of game theory argue, the assumption of rationality does not always hold in realistic environments. Especially in AI, "...we do not generally have the luxury of assuming rationality—it is our burden to explain how to realize approximately rational behaviors in operational computational terms"[BSW97]. Moreover, the players' decisions are arguably stochastic or "noisy", perhaps due to errors in perception, calculation, or the recording of actions. However, the relaxation of the classical assumptions of game theory and the incorporation of stochasticity (or "noise") into the agents' introspection process can some times provide the leverage to overcome those problems [GH99].

Further, the repeated encounter of the same or similar game scenarios provides the agents with the opportunity to learn, revising their beliefs, predictions and decisions over time. Reinforcement learning techniques, in particular, can be successfully employed in multiagent game settings in tasks as diverse as robots competing in soccer [BV01b, Lit94] and agents coordinating their actions in order to rescue civilians trapped in buildings collapsed after an earthquake [KT01]. The assumption underlying most of these techniques is that the MARL problem can profitably be defined *within a stochastic games framework*; this game-theoretic framework enables the participating agents to better align their action choices with those of others—since in a multiagent world the effects of an agent's actions are directly influenced by the actions of others, and also, crucially, have the potential to influence the future deliberations and actions of others. Therefore, one is forced to cast aside the direct adoption of single-agent RL techniques as simplistic and, largely, inappropriate.

On the whole, several learning scenarios in repeated games pose interesting questions regarding the rationality of agents and the effectiveness of predictions: Can rational players really learn to play effectively in repeated games? Can they achieve optimality in their play, and end up in equilibrium? In addition, can convergence to equilibrium *at any cost* be considered as the optimal behaviour?

The *Bayesian approach* is an attractive model for defining *optimal* sequential behaviour, when agents act under uncertainty: since it is impossible to remove all underlying uncertainty in a realistic environment, the reasonable behaviour for a rational agent would be to act according to its own evolving beliefs, trying—by being Bayesian—to implicitly take into account all eventualities concerning possible models of the world, and other agents inhabiting this world.

Combining (Bayesian) RL with Coalition Formation

While in *non-cooperative* ("strategic") game theoretic settings the players just pursue their own interests, in *cooperative* game theoretic settings players are allowed to form *coalitions* and combine their decision-making problems in order to achieve reward through cooperation. Cooperative game theory concentrates on the question of *what* a coalition can get, without saying *how*: instead of concentrating on the strategic choices of the individual, it deals with the options available to the group [AH92].

This does not mean that the players cease to be rational. If, in particular, *transferable utility* [Mye91] is assumed, the participants in a cooperative game can reach *enforceable* agreements that will exploit side payments among the agents. These side payments can induce players to use specific mutually beneficial strategies. The question of the *stability* of coalitions is central to cooperative coalition formation research, and several *cooperative solution concepts* to the problem of stability have been proposed.[3] In any case, rational players should seek to join the coalition that guarantees them the highest return. In fact, John Nash claimed, in his position known as the "Nash Program", that all cooperative games can be reduced into some non-cooperative form [Nas51, Nas53]. Moreover, there exists a substantial body of game theory research which examines (mainly from a non-cooperative standpoint) the processes by which coalitions emerge.

Game theory provides nowadays the main methods of examining economics. Non-cooperative game theory can be used, among its other applications, to analyze models for oligopolies or auctions, while cooperative game theory can find many applications in real-life problems, such as bilateral or multilateral bargaining [AH92, Fri91]; furthermore, e-commerce is another field for the potential application of both non-cooperative and cooperative game theory and coalition formation in particular. However, in many cases the assumptions made in existing research are unrealistic with respect to the requirements of real-life environments: in particular, the inevitable burden of uncertainty. Moreover, when scenarios of *repeated* coalition formation activities come into consideration, the possibility of employing learning mechanisms in order to enhance the decision making of agents presents itself. This is the case, for example, when the opportunity exists to bargain *repeatedly* with partners encountered in the past, or to revise the structure or the strategies adopted by the formed coalitions. It is quite natural that agents should then be capable of exploiting the experience they gathered in the past in order to make more informed decisions. As we will argue throughout our thesis, these are all areas where em-

[3] We expand on all these issues in section 2.3.

ploying Bayesian reasoning and Bayesian RL techniques can prove to be of value to rational agents that have to operate and interact under uncertainty.

To provide some intuitions into the dynamics of the problem of repeated coalition formation under uncertainty, and its implications for strategic reasoning and learning, let us consider a simple example. Suppose there exists a set of agents that have to form teams to participate in a construction project (such as one that would require them to build a new residential area). The agents may have different professional training, for example they can be carpenters, plumbers or electricians, each with different degrees of skills or expertise—an agent's training and expertise define its capabilities, or its *type*. Each agent may be uncertain regarding the others' types, for example, their degree of expertise. When a project is to be undertaken, the agents have the choice to form coalitions whose members will collectively decide to act and complete one of possibly several project components—such as choosing to build an apartment building, a single house, or a skyscraper. The choice and execution of a project or a project component is a *coalitional action*. The degree of the successful completion of the project component (i.e., the outcome of the coalitional action) would depend on the capabilities of the agents, and would result in an eventual payoff to the coalition. Naturally, there is inherent uncertainty in the environment regarding the outcomes of the coalitional actions (i.e., the outcomes are stochastic). Each member will get a share of the eventual payoff, and this share should be decided in advance. The choice of teammates, the choice of the coalitional action to perform, and the agreement over the shares to be allocated, must all be simultaneously decided via bargaining among all the agents, each of whom seek to maximize their expected payoff shares. Thus, the coalitions form *endogenously*, rather than being determined by some external force. Once the coalitions have formed and acted, a new, possibly different, project may be announced—thus, the agents would have the opportunity to engage in similar coalition formation activities over and over again.

This setting of repeated coalition formation under uncertainty presents an agent with many challenges and questions. The agent will have to decide whether he should stay in a coalition or abandon it. He might be content working with a specific electrician, yet he may be tempted to team-up with someone else and learn about this new partner's capabilities. While bargaining, the question of how big a share to ask for is always pressing, especially since one is unsure of the potential that a coalition has given uncertainty regarding his partners. Thus, it is difficult to decide how much payoff he could reasonably expect to claim successfully. Further questions might arise, such as what are the bargaining processes, if any, that can potentially guarantee the stability of formed coalitions, or how can one learn and benefit by observing the behaviour of

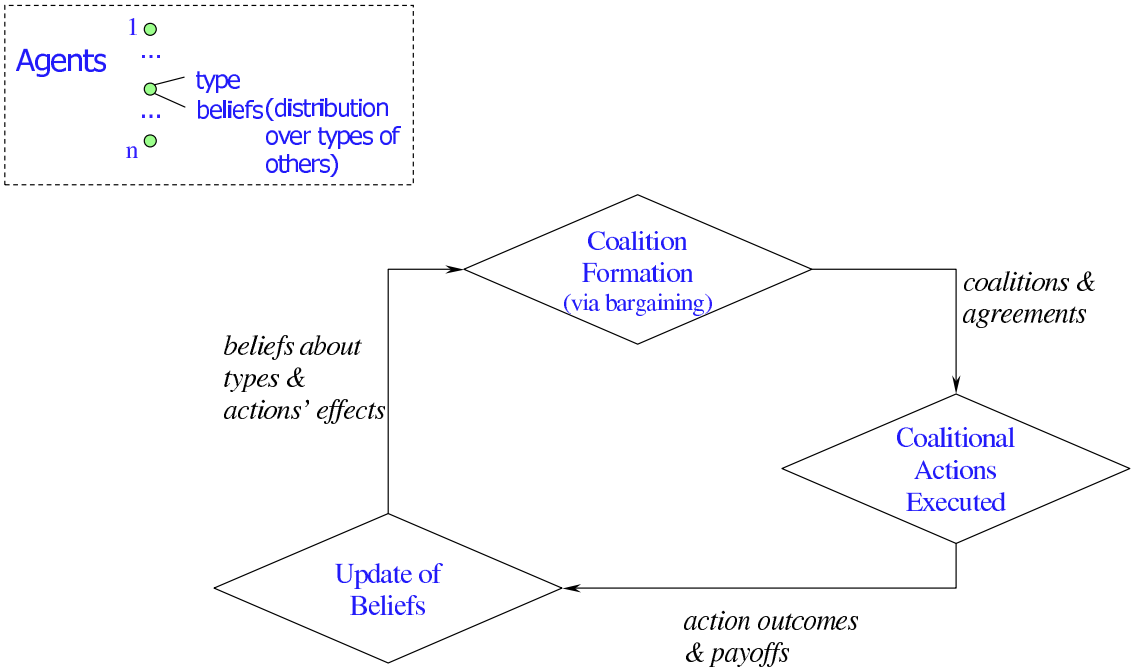


Figure 1.1: A Bayesian RL framework for repeated coalition formation under uncertainty.

others during bargaining. Thus, as demonstrated by even this simple example, the introduction of type-related uncertainty in the coalition formation problem can provides a rich agenda of research questions.

In this dissertation, we formally state and attempt to resolve several of those questions. In order to do so, we develop a Bayesian RL framework for repeated coalition formation under uncertainty, the outline of which is shown in Figure 1.1. The agents in this framework engage repeatedly in coalition formation activities resulting in a set of coalitional agreements, comprising a set of formed coalitions and a set of payoff allocations and coalitional actions (one per coalition) to perform (thus, we say that the agents participate in the "coalition formation stage" of the repeated process). The formed coalitions take the actions agreed by their members, and the observation of their outcomes enables the agents to update beliefs regarding the types of partners. This, in turn, revises their expectations regarding the *sequential* value of the potential coalitional agreements that may emerge as the result of the coalition formation activities ("RL

stage"). The evaluations of these agreements are then used in subsequent coalition formation activities.

We consider the development of this framework to be an important contribution of this thesis: in fact, the work presented in the various chapters of this dissertation builds towards this framework (or informs aspects of it). We now present a more comprehensive list of our major contributions.

1.2 Contributions and Thesis Outline

In this thesis we provide theoretical and algorithmic tools to accommodate sequential decision making under uncertainty in multiagent settings, dealing with issues and questions such as the ones identified above. Starting with the observation that the MARL problem can be best viewed through a game-theoretic perspective, we combine it with various interesting cooperative and non-cooperative game theory problems, and by doing so we facilitate both online learning and decision making under incomplete information.

Combining *multiagent* Bayesian RL with game theoretic ideas to facilitate agents' sequential decision making is the major contribution of this work. We deal with popular multiagent problems which research has so far hesitated to tackle *under uncertainty*, or more specifically under *type uncertainty*. Our work explicitly models this uncertainty: our research assumes that the environment dynamics or the capabilities (types) of the environment's co-inhabitants are not known, and thus the agents have to account for this uncertainty, in a Bayesian way, when making decisions.

Figure 1.2 provides a map of the research areas wherein our contributions lie. Other researchers have already defined the MARL problem within a game-theoretic, stochastic games framework. We have contributed a *Bayesian* approach to the MARL problem, introducing *a model-based Bayesian MARL framework for sequential decision-making in stochastic games*. Stochastic games lie within the non-cooperative branch of game theory, a branch that focuses on studying the strategic interactions of rational players and on describing their equilibrium behaviour. Cooperative game theory, on the other hand, deals basically with the coalition formation problem, focusing on the question of stability of formed coalitions, irrespectively of the process through which the coalitions have emerged. The branch of coalition formation research that does examine the dynamic processes by which coalitions emerge is usually referred to as

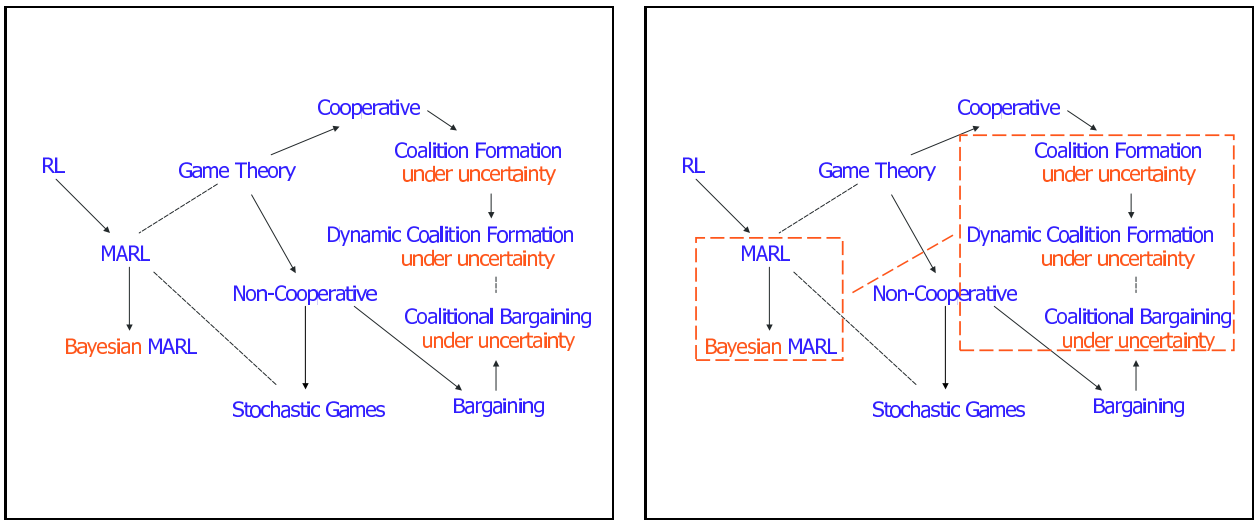


Figure 1.2: Areas of contributions (with our contributions marked in red).

dynamic coalition formation, and has strong links with the purely non-cooperative problem of *coalitional bargaining*, that examines the equilibrium behaviour of agents during multilateral bargaining to form coalitions. Dynamic coalition formation examines the endogenous formation of coalitions by some bargaining process,[4] usually modeled as a Markov process, mainly with respect to their potential of giving rise to stable coalitions (stability being described by some cooperative coalition formation solution concept). Coalitional bargaining research, on the other hand, focuses on examining the (purely non-cooperative) equilibrium solutions of the coalition formation problem. As we have already mentioned, however, and as we shall see in more detail, the treatment of uncertainty in coalition formation, dynamic coalition formation, and coalitional bargaining research is very limited. We, on the other hand, make contributions in all those fields, providing models, solution concepts and algorithms that take uncertainty (more specifically, type uncertainty and reward stochasticity) into account. As mentioned above, and as shown in the right part of Figure 1.2, an important contribution of our work is providing a link between Bayesian MARL and coalition formation research, by introducing a *Bayesian RL framework for repeated coalition formation under uncertainty*, which we detail in Chapter 6 of this thesis.

One main idea that links various pieces of our work together is the formulation of the unknown environment and agent interactions as a Partially Observable Markov Decision Process (POMDP). This provides firm foundations upon which agents can base their deliberations in

[4]The bargaining process in question may be cooperative or non-cooperative—focusing more on the joint or individual value to be gained, respectively.

order to determine rewarding behaviour. Thus, our work provides a *belief-state MDP* formulation for *Bayesian sequential decision making under uncertainty in multiagent environments*, which allows for the *expected value of information* of agents' actions to be taken into account during deliberations (the "value of information" represents the ability this information has to change future decisions). Such an approach intertwines online learning with optimal decision making under uncertainty. Our work shows that this approach can be successfully employed both in stochastic games, and in the coalition formation problem. In particular, our introduction of a model of Bayesian coalition formation under type uncertainty, and its integration within a Bayesian RL framework for coalition formation under uncertainty are two aspects of our work that make it quite original.

Modeling coalition formation under *type uncertainty* (i.e., uncertainty regarding the capabilities of others) is another important contribution of our work. Dealing with coalitional value uncertainty alone[5] does not allow an agent to learn in a way general enough to enable him to tackle similar or different problems in the future. On the contrary, if an agent is able to progressively enrich its knowledge regarding the capabilities of its peers, this will allow it to tackle other problems involving those same peers in the future, as these capabilities influence the values of the potential coalitions under any possible scenario. Uncertainty regarding types can then be readily translated into uncertainty regarding coalitional values (given any other necessary specifics of the particular problem), and thus the acquired knowledge (or uncertainty) about the types of others can be applied in different coalition formation scenarios.[6]

More specifically, our work has resulted in the following list of main research contributions, which we present and discuss in detail in subsequent chapters of this thesis:

1. To the best of our knowledge, we are the first to introduce *Bayesian Multiagent Reinforcement Learning*, proposing a *multiagent*, *model-based*, *Bayesian* RL framework for optimal sequential decision making *in stochastic games* (*Chapter 3*).

[5] Coalitional value uncertainty may depend on the particular characteristics of a specific problem—such as, perhaps, a particular model of stochasticity regarding the rewards paid to coalitions, or the particular uncertainty regarding resources to be allocated.

[6] For instance, consider a construction coalition in the toy example described earlier: the knowledge (or uncertainty) regarding the capabilities of team members (e.g., knowledge of whether agents A or B are good carpenters or not and of whether agents C or D are good plumbers or not) will guide an agent (perhaps agent E, an electrician) in his choice of partners, even if the particulars of the construction project change (e.g., even if project $\mathcal{A}$ requires him to team up with a plumber and project $\mathcal{B}$ with a carpenter, or if project $\mathcal{C}$ pays more for building a mansion in Toronto while project $\mathcal{D}$ suggests it is better to build an apartment in San Francisco). Not taking into account type uncertainty would have required agent E to try and learn the values of different coalitions for each different project from scratch.

- We focus on the problem of multiagent coordination in stochastic games and show that a Bayesian approach enables the agents to deal with the multiagent exploration-exploitation problem.
- The solution to the multiagent exploration-exploitation problem can be cast as the solution of a belief-state MDP, capturing the *expected value of information* of the agents' actions (and its impact to subsequent actions).

2. We provide a formal model for the problem of *coalition formation under type uncertainty* (*Chapter 4*).

   - We formulate a *Bayesian* coalition formation model that enables the agents to have beliefs about their potential partners' types, and expectations about the values of coalitions and their own payoff shares (in environments with stochastic rewards and unknown partner types).
   - We identify a *stability concept for coalition formation under uncertainty*, the *Bayesian core*; we note that ours is one of a few existing coalitional stability concepts under uncertainty—and the first proposed for coalition formation under type uncertainty.
   - We also discuss the problem of the existence of the Bayesian core (i.e., the problem of verifying whether stable coalitional configurations exist while bargaining under uncertainty). For small games, we provide an algorithmic method to decide whether the BC is empty or not.
   - Finally, we propose two algorithms for *dynamic coalition formation under uncertainty*, and prove that one of them converges to the Bayesian core.

3. We examine the problem of *discounted coalitional bargaining under type uncertainty* (*Chapter 5*).

   - Our discounted coalitional bargaining framework focuses on the strategic considerations required during coalition formation. We formulate the (Perfect Bayesian) equilibrium solution to the problem, and devise a coalitional bargaining algorithm to approximate it. The algorithm allows the agents to update beliefs regarding the types of others while observing their bargaining behaviour. In addition, we examine the effects of learning by observation of coalitional actions to the quality of decisions taken in such repeated coalitional bargaining settings.

- We are the first to examine the correspondence between cooperative (Bayesian core) and non-cooperative (Bayesian bargaining equilibria) solution concepts for coalition formation under uncertainty, proving propositions that relate them to each other. (This is similar to results that related research has obtained for deterministic, "certain" environments.) We thus provide a *non-cooperative justification* of the Bayesian core as a coalitional stability concept under uncertainty. As a corollary to our results, we also establish a sufficient condition for the existence of the Bayesian core (but only under specific assumptions regarding the bargaining strategies of the agents).

4. We bring together the coalition formation problem with Bayesian MARL, providing a *Bayesian RL framework for coalition formation under uncertainty* (*Chapter 6*).

   - Critically, our Bayesian approach to repeated coalition formation enables rational agents to make sequential decisions on taking formation (bargaining) actions and coalitional actions, in an informed manner:
   - "Coalition formation stages" alternate with "RL stages", in accordance to Figure 1.1, and the agents use one of several algorithms we propose in order to evaluate the quality of the various potential agreements.
   - The evaluation of the quality of the coalitional agreements is done by using the (approximate) *sequential* value of these agreements—we have proposed a belief-state MDP formulation that describes the long-term value that the agents place on the agreements (incorporating their expected value of information), and contributed RL algorithms that solve it approximately. Thus, our approach provides a solution for what we can call the problem of *optimal repeated coalition formation under uncertainty* (in which the participating agents are interested in eventually forming efficient, profitable coalitions, but they are also interested in gathering as much reward as possible while doing so).
   - Furthermore, we demonstrate experimentally that our approach enables the agents to effectively deal with situations where the coalition formation scenarios change over time (i.e., situations where the tasks the coalitions have to face are different after each coalition formation stage).
   - Notice that our framework can readily incorporate the dynamic coalition formation algorithms provided in Chapter 4 and the discounted coalitional bargaining algo-

rithm provided in Chapter 5 as its coalition formation components.

- We argue that this approach provides a novel perspective to the (repeated) coalition formation problem, interweaving as it does Bayesian deliberations based on observations with formation decisions and collective coalitional action-taking.

At the beginning of each of the chapters, we provide further motivation for the work described there, and elaborate on our relevant contributions. Furthermore, in each chapter we discuss related work as appropriate, and provide experimental results that support our approaches. Finally, Chapter 2 provides the background necessary for our thesis, while Chapter 7 draws conclusions and outlines possible future work. There, we also expand a bit on some of the potential application domains for our research, which include multi-agent planning (e.g., logistics, supply chain planning, robotic teams' formation and decision-making), e-commerce (e.g., many forms of multilateral bargaining, supply chain formation), computational trust, grid computing, ubiquitous computing, and so on.

Chapter 2

Background

Here we provide an overview of the basic concepts related to the problems we tackle throughout this thesis, specifically background on single-agent and multiagent reinforcement learning, and on non-cooperative and cooperative game theory. In the process, we draw the connections among those fields, focusing on the concepts, problems and approaches that are of the most relevance to our work.

2.1 Single-Agent Reinforcement Learning

Reinforcement learning is learning how to map situations to actions, so as to maximize a sequence of rewards. There exist two important features of reinforcement learning, distinguishing it from other kinds of machine learning. The first one is that the learner has to discover which actions yield the most reward through *trial-and-error search*. The second is *delayed reward*: actions may affect not only the immediate reward, but also the next situation and, through that, all subsequent rewards [SB98].

More specifically, an *agent* in a reinforcement learning system has in its disposal a finite[1] set of actions, A, through which it interacts with the *environment*, that has a finite set of states S. At each time step, the agent observes the environment's state, and on that basis selects an action; one time step later, in part as a consequence of its action, the agent receives a *reward* and finds itself in a new state. The reward is provided by a (real-valued) *reward function*. The reward function defines the goals of the reinforcement learning problem, by providing the agent with reinforcement signals that depend on the current state or state-action pairs.

[1] In many cases in this article—for the sake of simplicity— we will be restricting attention to discrete time, and finite number of actions, states or rewards; however, more general formulations are possible.

The rewards determine the immediate desirability of states. However, the long-term goal of an agent acting within a reinforcement learning system is to maximize the total amount of accumulated reward, therefore delayed reward has to be considered along with instantaneous reward [KLM96, SB98].

In order to achieve its goal, the agent has to follow a *policy*. The policy π of the agent is a mapping from perceived states of the environment to actions to be taken when in those states. More accurately, π is a mapping from each state $s \in S$ and action $a \in A$ to the probability $\pi(s, a)$ of taking action a when in state s. In order for the agent to determine its policy, it needs to consider the *value* of each state. The value of a state is the total amount of reward an agent can expect to accumulate over the future starting from that state, and is specified by a *value function*. Values indicate the long-term desirability of states.

Some reinforcement learning systems have as an element a *model* of the environment, that can be learned and used for predicting the next state and next reward, given the current state and action. A model of the environment is not an essential part of a reinforcement learning system; as a matter of fact, it is debated whether a model should or should not be learned and used by RL algorithms. (We discuss this issue in 2.1.2.)

The current section will allow for a more detailed description of the reinforcement learning concepts mentioned above.

2.1.1 Markov Decision Processes

A key assumption underlying much reinforcement learning research is that the interaction between an agent and the environment can be modeled as a *Markov Decision Process (MDP)* [Bel57, How60, Put94, BDH99].

An MDP is a 4-tuple $\langle S, A, p_T, p_R \rangle$. S is a set of states, A is a set of actions, p_T is a transition model that captures the probability $p_T(s, a, t)$ of reaching state t after executing action a at state s, and p_R is a reward model that captures the probability $p_R(s, a, r)$ that we receive reward r after executing a at s. From an RL point of view, an MDP can be viewed as a complete specification of the reinforcement learning environment that satisfies the Markov property [SB98].[2]

[2]The environment within which an RL agent operates is said to have the Markov property if its response at $t + 1$ depends only on the state and agent's action at t, in which case the environment's dynamics can be defined by specifying only the probability distribution $Pr\{s_{t+1} = s', r_{t+1} = r | s_t, a_t\}$ for all s', r, s_t and a_t. If an environment has the Markov property, then its one-step dynamics enable us to predict the next state and expected next reward given the current state and action [KLM96, SB98].

A history h is the sequence of states and actions generated from the beginning of system's evolution to some point of interest [BDH99]. The solution of an *infinite-horizon* MDP requires that the agent's performance be evaluated over an infinite history. Therefore, the discounted *infinite-horizon model of optimal behaviour* [Bel57, How60] assumes that rewards received in the future are geometrically discounted according to a discount factor γ ($0 \leq \gamma \leq 1$). The expected reward that has to be optimized is then given by:

$$E(\sum_{t=0}^{\infty} \gamma^t r_t)$$

where r_t is the reward at time step t. The agent's aim is to maximize the expected discounted total reward it receives. This requires computing an optimal value function V^* and a Q-function Q^*. These functions satisfy the Bellman optimality equations (for all a and s)[Bel57, SB98]:

$$V^*(s) = \max_{a \in A} Q^*(s,a) \tag{2.1}$$

where

$$Q^*(s,a) = E_{p_R(s,a,r)}[r|s,a] + \gamma \sum_{s' \in S} p_T(s,a,s') V^*(s') \tag{2.2}$$

These equations say that the *quality Q of a state-action pair* is the immediate reward plus the expected discounted value of all succeeding states weighted by their likelihood, and that the *value V of a state* is the quality of the best action for that state.

The *state-value function* $V^*(s)$ is the unique solution of (2.1) and (2.2). Once one has V^* it is easy to determine an optimal policy: any policy that is *greedy*[3] with respect to V^* is an optimal policy. In other words, once the optimal value function is known, the actions that appear to be best (i.e., the actions at which the maximum of (2.1) is attained) after a one-step search will be optimal actions. In the case the transition and reward models are given, *dynamic programming* [Bel57] techniques can be used to determine the optimal policy:

The method of *value iteration* (Figure 2.1) starts with estimates Q and V of Q^* and V^*, respectively, and updates them repeatedly by applying the previous equations to get new values for Q and V [KLM96, SB98]. It has been shown that the estimated values for Q and V converge to their true values [Ber87, Put94]. This in theory requires an infinite number of iterations, but

[3]A policy π is greedy with respect to a value function f if, for all states s, $\pi(s)$ is an action satisfying $Q^f(s,\pi(s)) = \max_{a \in A} Q^f(s,a)$. A greedy policy selects an alternative based only on local or immediate considerations, without considering the possibility that such a selection may prevent future access to even better alternatives.

initialize V(s) arbitrarily, e.g. $V(s) = 0$, for all $s \in S$
repeat //(until policy is good enough)
$\Delta := 0$
For $s \in S$
{
$u := V(s)$
For $a \in A$
{
$Q(s,a) = E_{p_R(s,a,r)}[r|s,a] + \gamma \sum_{s' \in S} p_T(s,a,s')V(s')$
}
$V(s) := max_a Q(s,a)$
$\Delta := max(\Delta, |\upsilon - V(s)|)$
}
until $\Delta < \epsilon$ (a small positive number)

output a policy π such that for each s
$\pi(s) = argmax_a Q(s,a)$

Figure 2.1: The value iteration algorithm

choose an arbitrary policy π'
repeat
$\pi := \pi'$
compute the value function of policy π:
solve the linear equations
$V_\pi(s) = E_{p_R(s,\pi(s),r)}[r|s,\pi(s)] + \gamma \sum_{s' \in S} p_T(s,\pi(s),s')V_\pi(s')$

improve the policy at each state:
$\pi'(s) := argmax_a E_{p_R(s,a,r)}[r|s,a] + \gamma \sum_{s' \in S} p_T(s,a,s')V_\pi(s')$
until $\pi = \pi'$

Figure 2.2: The policy iteration algorithm

in practice the execution of the algorithm is stopped once an appropriate *stopping criterion* is satisfied. Such a criterion could be that the value function changes by a small amount ϵ in a sweep, such that $\epsilon = \epsilon'(1-\gamma)/2\gamma$. It is proved that the resulting policy is ϵ'-optimal [WB93]. The value iteration algorithm guarantees that—in many cases—the greedy policy is optimal long before the value function has converged [Ber87]. Moreover, the assignments to V can occur asynchronously, provided that the value of each state gets updated infinitely often on an infinite run, so that faster convergence is achieved.

Policy iteration (Figure 2.2), on the other hand, manipulates the policy directly rather than finding it indirectly via the state-value function. The *value function of a policy*, V_π, is the

expected infinite discounted reward that will be gained, at each state, by the execution of that policy. It can be computed by solving a set of linear equations. This constitutes the *policy evaluation* step of the algorithm. Once the value of each state under the current policy is known, a policy improvement step is tried by changing the first action taken when in a state. If the value of the state can be improved, the new action is adopted by the policy; thus, the policy is strictly improved. The algorithm iterates policy evaluation and improvement steps, until no further improvements are possible. The policy is then guaranteed to be optimal [How60].

The computational complexity of the value iteration algorithm, per iteration, is linear in the number of actions and quadratic in the number of states ($O(|A||S|^2)$). However, the number of iterations required can grow exponentially in the discount factor, because, as the discount factor approaches 1, the decisions must be based on results that happen further and further into the future [LDK95]. In practice, policy iteration tends to converge in far fewer iterations than does value iteration: its rate of convergence is quadratic, rather than linear as is the case for value iteration [Put94]. However, its per iteration costs of $O(|A||S|^2 + |S|^3)$ can be prohibitive.[4]

Dynamic programming techniques may not be practical for large problems because of the *curse of dimensionality* [Bel57], the fact that the number of states often grows exponentially with the number of state variables. However, there exist methods to avoid some of the computational difficulties that arise in large state spaces. *Real-time dynamic programming* [BBS95], for example, a form of online asynchronous value iteration, has been proposed as a way of approximately solving large MDPs.[5] Another way to alleviate the computational burden of a large state space is to use *sampling methods* [DeG70], which sample from the space of possible histories and use this information to provide estimates of the values of specific policies.

More generally, *learning in large state spaces* can be addressed through the adoption of *generalization techniques*, which allow compact storage of learned information and transfer of knowledge between "similar" states and actions [KLM96]. Popular techniques include various function approximation methods, such as neural network methods and generalizations of nearest neighbours method. For instance, a function approximator such as a backpropagation network can be used to take examples from a value function and attempt to generalize from

[4]Policy improvement can be performed in $O(|A||S|^2)$ steps and value determination by solving the system of linear equations requires $O(|S|^3)$ steps at most; in practice, however, the second cost is usually much less than its forementioned theoretical worst case value [KLM96].

[5]An *online* RL algorithm is one that not only learns a value function but also simultaneously controls a real environment. *Asynchronous* algorithms, unlike synchronous algorithms, do not place any constraints on the order in which the state-update equation is applied to update the values of the states; however, the values of all statesshould in the limit be updated infinitely often.

them to construct an approximation of the entire function.

Finally, representational economy can be achieved by adopting *factored representations* of states, actions, rewards and other components of an MDP [BDH99]. A state space is factored if it is determined by a set of state variables, representing the *features* that are required in order to describe the state. Instead of viewing the state as a single variable taking on a huge number of values, it can be viewed as a cross product of its features, each one of them taking on substantially fewer values. A factored action representation describes the effect of an action on specific state features rather than on entire states.

We end this subsection by briefly introducing an extension to the MDP model: An MDP assumes full observability of the current state; a *Partially Observable Markov Decision Process (POMDP)* [Mon82, KLC98], on the other hand, is an MDP in which the agent is unable to observe the current state, but is able to make (probabilistic) observations about it. (The probability of making an observation depends on the action taken and the resulting state.) Any given (discrete) POMDP induces an MDP whose states are belief states—or a *belief-state MDP* [Ast65, SS73, Son78]. Belief states are probability distributions over the states of the world, and they comprise a *sufficient statistic* for the past history (the process over belief states is Markov) [Ast65, SS73, Son78]. The optimal solution of the belief-state MDP gives rise to optimal behaviour for the original POMDP [Ast65, SS73, Son78]. There are many techniques for the efficient solution of POMDPs, even in the case of large state spaces [KLC98, Pou05].

2.1.2 Reinforcement Learning

As mentioned above, it is a common assumption in RL research that the agent-environment interaction can be modeled as an MDP. Learning by reinforcement is well-suited to situations where there is significant uncertainty about some parameters of the MDP model: RL algorithms try to maximize the agent's expected reward when the transition model p_T and sometimes the reward model p_R are initially unknown. So, in a nutshell, the reinforcement learning problem can be defined as an agent's attempt to maximize its long-term reward through trial-and-error, while operating in an environment that is modeled as an MDP with unknown p_T and, possibly, unknown p_R.

Maximization of long-term reward, either in a finite or infinite horizon, is the criterion to use in order to assess the policies learned by a given algorithm. The *quality of learning*, however, can be assessed by using measures such as whether the algorithm provably *converges to optimal behaviour or not*, whether the algorithm converges *fast* to optimality or near-optimality, and

whether the *regret* of the algorithm is large or not (i.e., what is the expected decrease in reward gained due to executing the learning algorithm instead of behaving optimally from the very beginning). Unfortunately, these learning performance metrics may be incompatible with each other: for example, an algorithm may be provably always convergent but with a very slow rate, or it can take long to achieve optimality but at the same time be very effective in terms of accumulating reward along the way [KLM96].

A central debate in RL research is over whether learning and using the unknown transition and reward models is important in order to control the actions of an agent [KLM96, SB98]. *Model-based* algorithms use a model of the environment to update the value function. In case the model is not given a priori, then—as is common in the RL case—it has to be estimated: real experience is used for improving the model (*model learning*). The current learned model is then used to update the corresponding value (or Q-value) function and plan a policy that optimizes a given criterion. *Model-free* (or *direct reinforcement learning*) algorithms do not have access to the transition or the reward model; they just apply the state-update equation to the state of the real environment, trying to directly learn the Q-function and policy.

Both model-based and model-free methods have advantages and disadvantages. A common argument for the use of model-based algorithms is that by learning a model the agent can avoid costly repetition of steps in the environment. The agent is able to use the model to reason about the effects of its actions. Thus, the number of steps actually executed by the agent is reduced, since simulated steps in the model can be used for learning or computing a value function. On the other hand, model-free algorithms have the advantage of being simpler than their model-based counterparts, and, in addition to that, they are not affected by biases in the design of the model. In the following paragraphs we will present some well-known model-based and model-free algorithms.

Model-Based Algorithms

Model-based algorithms need to make use of the transition probabilities and/or a reward model.

A conceptually straightforward model-based algorithm is the *certainty equivalence* method [KLM96, SB98]. It is given its name because it assumes that the model is known with certainty, even though in reality it is being approximated. The method tries to continually learn the model by keeping statistics about the results of each action. The estimated transition probability from state s to state s' under action a might be, for example, the fraction of observed transitions that from s to state s', and the associated expected reward is the average of the rewards observed

on those transitions. *At each step*, the current model is used to compute an optimal policy and value function, employing, perhaps, some dynamic programming technique. Thus, even though the use of data available is effective, the method is computationally demanding (since at each step a complete computation of the optimal policy is required). In addition, the algorithm does not *explore* the environment at all.[6]

Another well known model-based RL algorithm is *prioritized sweeping* [MA93]. Prioritized sweeping, being a model-based method, works by maintaining an estimated transition model $\widehat{p_T}$ and a reward model $\widehat{p_R}$. Whenever an experience tuple $\langle s, a, t, r \rangle$ is sampled, the estimated model at state s can change; a Bellman backup is done at s to incorporate the revised model and some number of additional backups are performed at selected states. States are selected according to a *priority* that estimates the potential change in their values, based on the size of the changes occured by earlier backups (the backups are focused on states that can benefit from them the most). Prioritized sweeping has found several applications, such as in *learning by imitation* [PB99].

Last but not least, Dearden *et al.*[DFA99] have presented a way to be Bayesian when employing model-based RL. We will describe this approach later in this dissertation.

Model-Free Algorithms

Unlike their model-based counterparts, model-free algorithms try to learn and produce an optimal policy without having to learn and use a model of the environment.

TD(0) [Sut88] is a model-free method that learns the value of a policy π by using the update rule

$$V_\pi(s) := V_\pi(s) + \alpha(r + \gamma V_\pi(t) - V_\pi(s))$$

Whenever a state s is visited, its estimated value is updated[7] so that it approaches $r + \gamma V_\pi(t)$, where r is the instantaneous reward received, $0 < \alpha < 1$ is the *learning rate* governing to what extent the new sample is replacing the current estimate, and $V(t)$ is the estimated value of the observed next state t, after taking action a prescribed by π for s. The quantity $r + \gamma V_\pi(t)$ is a sample of the real $V_\pi(s)$; it depends in part on the error-prone $V(t)$ estimate, but it also incorporates the real, error-free r. In this way, convergence can be achieved because of per-update reduction of the largest error between true and estimated values of $V_\pi(s)$. If α is slowly decreased and the policy π held fixed, TD(0) is guaranteed to converge to the value

[6] We will refer in detail to the question of *exploration* shortly.

[7] TD(0) updates are *sample backups*, because they are based on a single sample successor state.

function V_π [JJS94]. Then, in order to estimate the optimal policy, one can combine TD with an "actor-critic" method [KLM96, SB98]. These essentially alternate policy evaluation stages (which "criticize" a policy π using TD) with policy improvement steps (that try to improve π by taking into account the "critique" provided to each state by the estimated value function $V_\pi(s)$ and its corresponding error $r + \gamma V_\pi(t) - V_\pi(s)$).

TD(0) looks only one step ahead when adjusting the value estimates. Its *multi-step* version, TD(λ) [Sut88], uses an update rule that is similar to the one used by TD(0):

$$V_\pi(u) := V_\pi(u) + \alpha(r + \gamma V_\pi(t) - V_\pi(s))e(u)$$

The rule is not just applied to the immediately previous state, but to *every* state u, according to its eligibility $e(u)$, which shows the degree it has been visited in the past and is a function of γ and λ, $0 \leq \lambda \leq 1$. TD(λ) is a way of averaging the n-step backups. The one-step return has weight $1 - \lambda$, the two-step return has weight $(1 - \lambda)\lambda$, the three-step return has weight $(1 - \lambda)\lambda^2$ and so on. Thus, for $\lambda = 0$, TD(λ) becomes TD(0), while when $\lambda = 1$ it is roughly equivalent to updating all the states according to the number of times they were visited by the end of a run.

Q-learning [WD92] is an *off-policy*[8] TD control method; it estimates the Q-values online[9] using essentially TD(0), but at the same time it uses them to define a policy, because an action can be chosen just by acting greedily with respect to the Q-values. An update is performed by an agent whenever it receives a reward of r when making a transition from s to s' after taking action a. The update rule is

$$Q(s, a) := Q(s, a) + \alpha(r + \gamma max_{a'} Q(s', a') - Q(s, a))$$

The probability with which this happens is precisely $p_T(s, a, s')$, which is why it is possible for an agent to carry out the appropriate update without using a transition model. The learned Q-value function directly approximates Q^*, independently of the policy being followed, but, of course, the policy followed has an effect on which state-action pairs are visited and updated. In any case, if each action is tried infinitely often and α is decayed appropriately, the Q-values

[8] Off-policy algorithms may update estimated value functions on the basis of actions other than those actually executed. In this way, off-policy algorithms can conceptually separate exploration from control; so, convergence proofs are enabled independently of the exploration technique. *On-policy* methods, on the other hand, update value functions just on the basis of experience gained from executing a policy [KLM96, SB98].

[9] Online algorithms face the *exploration vs. exploitation tradeoff* to be discussed shortly. Therefore, Q-learning *does* face the exploration vs. exploitation tradeoff.

will converge to Q^* [WD92, JJS94].

At this point, it may be worth pointing out some analogies between TD(0) and TD(λ) to policy iteration and Q-learning to value iteration. Interestingly, the update rules used by TD(0) and TD(λ) methods refer to the value of states given a policy, as does the update rule used by policy iteration; on the other hand, the update rule used by Q-learning updates a state-action pair, while the backup is also maximizing over all those actions possible in the next state, exactly as backups used by value iteration do [KLM96, SB98].

Finally, in departure from the methods mentioned above, Arai and Sycara propose in [AS00, AS01] an RL method that does not attempt to satisfy the Bellman equations. Their approach is closely related to the Profit-Sharing method [Gre88], that does not utilize successive approximations to compute state values. Rather, their method piles reward on successful (i.e., leading to goals) state-action pairs encountered within episodes of task execution in stochastic domains. An internal episodic memory is used to successfully identify perceptual aliasing states, discard looping behaviour, and thus form effective stochastic policies—and even resolve potential conflicts in multiagent domains.

The Exploration vs. Exploitation Problem

One major characteristic of reinforcement learning is that the agent should explicitly *explore* its environment in order to acquire knowledge about it. This is where the *exploration vs. exploitation problem* [SB98, Thr92] arises: the agent has to choose between executing actions that allow it to improve it estimate of the value function (*exploration*) and actions that return high (anticipated) payoffs (*exploitation*). This is often a hard dilemma, since rewards cannot really be maximized over time without exploring the environment, while exploration can waste much time in exploring costly parts of the environment or losing the opportunity to act in more rewarding regions. So, optimal combination of exploration and exploitation strongly depends on the particular learning problem, as well as the particular learning technique employed.

Two categories of exploration techniques can be identified. *Undirected* exploration techniques explore an environment based on randomness, while *directed* exploration utilizes some exploration-specific knowledge for guiding exploration: actions are selected so that the expected knowledge gain is optimized.

The most basic undirected exploration technique, *random exploration*, always selects actions randomly, with uniform probability. On the other hand, the purest way to minimize exploration is to adopt a *greedy* action-selection policy. Greedy agents always perform the

action with the highest estimated value. *ϵ-greedy* methods adopt a greedy policy most of the time, but once in a while, with a small probability ϵ select an action at random [SB98, Thr92].

The exploration policies that become more and more greedy over time are called *decaying exploration policies*. The class of decaying exploration policies that demand that each action is tried infinitely often in every state that is visited infinitely often, and that, in the limit, the learned policy is greedy with probability 1, is labeled *GLIE*, i.e., "greedy in the limit with infinite exploration". An example of a GLIE technique is *Boltzmann exploration*: action a is chosen at state s with probability

$$\frac{e^{Q(s,a)/T}}{\sum_{a' \in A} e^{Q(s,a')/T}}$$

where $Q(s, a)$ is the agent's estimate of the Q-value of performing a at s. The *temperature parameter* T can be decreased over time so that the exploitation probability increases; this can be done in such a way that convergence is assured [SJLS00].

Boltzmann exploration is an undirected exploration technique. *Counter-based exploration*, on the contrary, is a directed exploration technique that tries to drive the agent to less explored states of the environment. This, in the simplest way, is achieved by counting the occurrences of each state and then using a rule that forces the agent to visit the least visited neighbouring states of the current state [Thr92].

The *interval estimation method* [Kae93], proposed originally for bandit problems [BF85] and generalized by [Wie99] to general MDPs, is a popular statistical algorithm that tries to control the exploration-exploitation tradeoff. *Confidence-interval estimates* of each action's value are constructed. They serve to capture the uncertainty for each action and some information about "how good" the action is—in other words, what the potential of it being the best action in terms of expected reward is—based on statistics kept regarding the success of the action. An action a is chosen according to $a = argmax_a\{Q(s, a) + U(s, a)\}$, where $U(s, a)$ is a $(1 - \delta)$ upper confidence interval on the point estimate $Q(s, a)$. Intuitively, the more uncertain a's value is, the greater the probability that a will prove to be optimal. Naturally, the method can be misled by particular configurations of the environment so that to suggest suboptimal actions; this is a "curse" associated with algorithms that simply define local measures of uncertainty based on the theory of bandit problems.

Meauleau and Bourgine[MB99], on the other hand, propose the *IEQL+* algorithm, which is closely related to interval estimation but is less likely to fall prey to the aforementioned "curse". IEQL+ backs up Q-values and uses them to compute a local exploration bonus. However,

unlike interval estimation, IEQL+ backs up an exploration bonus and combines the two to compute the new exploration value of the action. Nevertheless, this is still in essence a heuristic approach (even if a principled one) to resolving the exploration-exploitation tradeoff.

Bayesian Reinforcement Learning

In contrast, the *Bayesian approach* [SL73, DFR98, DFA99, Duf02, Pri03] is a principled, non-problem-specific approach that provides an *optimal solution* to the action choice problem in RL. The optimal solution to the RL action selection problem—or *optimal learning*—is the pattern of behaviour that, *in expectation*, maximizes performance over the entire history of interactions of an agent with the world. As mentioned earlier, a performance metric commonly used in RL—and, actually, the main metric we will be using when evaluating the performance of RL algorithms throughout this thesis—is discounted accumulated reward.

The Bayesian approach formulates the action selection problem as a *belief-state MDP*, representing the agent's beliefs as a prior density over (all) the possible dynamics' and rewards' models of the environment, allowing thus for the explicit representation of his uncertainty. The solution to the belief-state MDP provides the optimal solution to the action selection problem (by choosing the policy that obtains maximum expected reward in the belief-state MDP), without requiring the definition of an explicit exploration strategy (therefore, the Bayesian action selection is also referred to as *Bayesian exploration*): the agents simply act greedily with respect to the Bayesian Q-values. No other method outperforms the Bayesian method in expectation, when using the same prior information [Bel61, Mar67]. However, there is no guaranteed efficient way to compute the aforementioned solution [MGLA00, LGM01], and thus approximation algorithms have to be devised.

Dearden *et al.*[DFR98, DFA99] were the first[10] to explicitly define and utilize *Bayesian RL*. [DFR98] provides a Bayesian approach to Q-learning: probability distributions over the Q-values are maintained and propagated, in order to represent the uncertainty the agent has about its estimate of the Q-value of each state. Actions to be performed are selected according to local Q-value information, but the distributions are used so as to compute a myopic approximation to the value of information for each action, and hence to select the action that balances exploitation and exploration in the best possible way. To put it differently, this work uses the concepts of the *value of perfect information (VPI)* and of the *value of exploration estimates* in

[10]However, there do exist earlier attempts for the Bayesian control of Markov chains and MDPs, such as [CGZM65, SL73].

order to boost the desirability of different actions. [DFA99] proposes a *model-based* Bayesian RL approach. Under some reasonable assumptions, a posterior distribution over possible dynamics and reward models given past experience can be represented and updated in a tractable fashion. The Bayesian model-based RL approach starts with some *conjugate prior*[11] over all possible MDPs, and progressively the mass of the *posterior* distribution is focused on those MDPs in which the observed experience tuples are most probable. If *parameter independence* is assumed, then the learning problem can be reformulated as a collection of unrelated local learning problems, in each of which the estimation of a probability distribution over all states and all rewards is required. Then, at each stage, the models are updated, and k MDPs are sampled from the distributions to be solved for sample Q-values for each state-action pair. VPI is again used to estimate the benefit of exploration. The main advantage of the Bayesian approach over existing model-based approaches which use simple estimation methods to learn the environment and keep a point estimate of its dynamics, is that it does not ignore the agent's uncertainty about those dynamics.[12] By representing a distribution over possible models, the agent's uncertainty can be quantified, which can in turn be used to inform it as to what are the best actions to perform. We will provide more details on [DFA99] in Section 3.1, as it provides the basis of our *multiagent* Bayesian RL approach (presented in Chapter 3). Bayesian agents [DFR98, DFA99] have been experimentally shown to effectively balance exploration with exploitation of actions.

More recently, [Duf02] has proposed several other computational procedures to sidestep the intractability of solving the belief-state MDP representing the action selection problem. Finally, the Bayesian approach has also been used in *learning by imitation*[PB03], and it has been combined with *sparse sampling* to improve action selection by intelligently growing sparse lookahead trees [WLBS05].

2.2 Non-Cooperative Game Theory

In this section, we provide the reader with a small review of basic non-cooperative game theory concepts. Non-cooperative game theory deals with the strategic interactions of players pursuing their own individual interests in games. We also illustrate how specific *learning models*

[11] A conjugate prior is a family of prior probability distributions which has the property that the posterior probability distribution also belongs to that family [DeG70].

[12] It is worth noting here that Wyatt [Wya01] has combined the Bayesian view of exploration with model-based interval estimation, via the use of an "optimistic model selection" algorithm.

	s_2^1	s_2^2
s_1^1	$u^1(s_1^1, s_2^1), u^2(s_1^1, s_2^1)$	$u^1(s_1^1, s_2^2), u^2(s_1^1, s_2^2)$
s_1^2	$u^1(s_1^2, s_2^1), u^2(s_1^2, s_2^1)$	$u^1(s_1^2, s_2^2), u^2(s_1^2, s_2^2)$

Figure 2.3: The matrix representation of a two-player, two-action strategic form game.

can be utilized in the process of game-playing. (A learning model [FL98] is any model that specifies the learning rules used by individual players, and examines their interaction when a game is played repeatedly.[13]) We then proceed to define the multiagent reinforcement learning (MARL) problem and present attempts to solve it within a principled game-theoretic framework.

2.2.1 Strategic Games

One very simple way to represent a game is to use the *strategic form*. A *strategic (or normal) form game* [FL98, Mye91] is a tuple $\langle N, (A_i)_{i \in N}, (u^i)_{i \in N} \rangle$. N is a finite set of n players, A_i is a finite set of actions available to player i, and u^i is the real-valued utility (payoff) function for player i.

More specifically, u^i is defined as $u^i : \Sigma \rightarrow \Re$, $\Sigma = \Sigma_1 \times ... \times \Sigma_n$, where Σ is the set of the possible *strategy profiles*. The strategy profile $\boldsymbol{\sigma}$ is the combination of *strategies* $\sigma_i \in \Sigma_i$ that the players might choose. A player follows a *pure strategy* if he selects and executes a single action from the set of actions available to him. A randomization by a player over his pure strategies, according to a fixed probability distribution, is called a *mixed strategy*; a strategy profile composed of mixed strategies is a *mixed-strategy profile*. The vector containing the strategies of i's opponents is denoted by $\boldsymbol{\sigma}_{-i}$.

Strategic form games with two or three players can usually be easily represented by one or more matrices depicting the players, their possible actions, and their payoffs. This is why they are also called "matrix" games. The matrix representation of a two-player, two-action normal form game is shown in Figure 2.3.

Finally, strategic games with *incomplete information*, meaning that players enter the game with private information, are called *Bayesian games*. The private information of the players is captured by their *type*: a Bayesian game is a tuple $\langle N, (A_i)_{i \in N}, (\Theta_i)_{i \in N}, (p_i)_{i \in N}, (u^i)_{i \in N} \rangle$,

[13] In other words, a learning model attempts to model the processes by which players change their behaviour while playing a game.

where N, A_i are as before (and let $A = \times_{i \in N} A_i$); Θ_i is a set of possible types for player i (and let $\Theta = \times_{i \in N} \Theta_i$; p_i summarizes what i believes about the types of others: $p_i(\boldsymbol{\theta}_{-i}|\theta_i)$ is the probability that i of type θ_i assigns to opponents' profile $\boldsymbol{\theta}_{-i} \in \Theta_{-i}$; and $u^i : A \times \Theta \rightarrow \Re$, ($u^i(a|\theta)$ is the payoff to i when the action profile is a and the type profile is θ—however, frequently u^i depends only on θ_i rather than θ as a whole).

2.2.2 Equilibria and Equilibrium Selection

An important concept in a game is the concept of *Pareto optimality* [FL98, Mye91]. A strategy profile of a game is *Pareto optimal* (or *Pareto efficient*) if and only if there exists no other strategy profile in the game that would have a higher payoff for some players without resulting to a lesser payoff for any other players. Obviously, a Pareto optimal outcome is highly desired. A different game solution concept, but indeed really important and much celebrated, is the concept of *Nash equilibrium*.

A mixed-strategy profile $\boldsymbol{\sigma}$ of a game is a Nash equilibrium if and only if, for all $i \in N$ and for all $\tilde{\sigma}_i$ in the set of all possible randomized strategies for i,

$$u^i(\boldsymbol{\sigma}) \geq u^i(\tilde{\sigma}_i, \boldsymbol{\sigma}_{-i})$$

Equivalently, a mixed strategy profile of a game is a Nash equilibrium if and only if each player plays a *best response*[14] strategy to the other players' choices of strategies [FL98, Mye91], which means that *no player can benefit by unilaterally deviating from the equilibrium*. The best response $BR^i(\boldsymbol{\sigma}_{-i})$ of player i to his opponents' strategies is the strategy that would maximize his expected utility payoff, given the strategy profile vector of all players' strategies: $BR^i(\boldsymbol{\sigma}_{-i}) = argmax_{\tilde{\sigma}_i} u^i(\tilde{\sigma}_i, \boldsymbol{\sigma}_{-i})$.

Every matrix game has a mixed-strategy Nash equilibrium [Nas51]. Mixed strategies have to be assumed for proving the existence of a Nash equilibrium, since there exist matrix games that have no equilibria in pure strategies.

In Bayesian games (see Section 2.2.1 above), players know only their own type (and have priors over others' types). Since other players' types are unknown, in equilibrium each player needs to form a best response against the expected strategy of each opponent, averaging over the reactions of all possible types of an opponent—we can formally define *Bayesian* equilibria

[14] A *best response* strategy of a player to his opponents' strategies is the strategy that would maximize his expected utility payoff, given the *strategy profile* vector of all players' strategies.

	Cooperate (do not confess)	**Deviate (confess)**
Cooperate (do not confess)	$5, 5$	$0, 6$
Deviate (confess)	$6, 0$	$1, 1$

Figure 2.4: The Prisoner's Dilemma game. If the players follow the unique equilibrium strategy profile prescribing confession, they get a payoff of 1 year of freedom (5 years imprisonment), while if they cooperate they gain 5 years of freedom (only 1 year of imprisonment).

for these games, as follows: A *randomized (mixed) strategy profile* for a Bayesian game is any strategy profile $\boldsymbol{\sigma} = \left((\sigma_i(a_i|\theta_i))_{a_i \in A_i}\right)_{\theta_i \in \Theta_i, i \in N}$ in the set $\times_{i \in N} \times_{\theta_i \in \Theta_i} \Delta(A_i)$ (with $\Delta(A_i)$ representing the set of all possible probability distributions over actions for player i) [Mye91]. In $\boldsymbol{\sigma}$, the number $\sigma_i(a_i|\theta_i)$ is the probability that player i would use action a_i if he was of type θ_i, and the *randomized strategy* for type θ_i of player i is $\sigma_i(\cdot|\theta_i) = (\sigma_i(a_i|\theta_i))_{a_i \in A_i}$ [Mye91]. Then, a *Bayesian (or Bayes-Nash) equilibrium* of the Bayesian game is any $\boldsymbol{\sigma}$ such that, $\forall i \in N$ and $\forall \theta_i \in \Theta_i$,

$$\sigma_i(\cdot|\theta_i) \in argmax_{\tau_i \in \Delta(A_i)} \sum_{\boldsymbol{\theta}_{-i} \in \Theta_{-i}} p_i(\boldsymbol{\theta}_{-i}|\theta_i) \sum_{a \in A} \left(\prod_{j \in N-i} \sigma_j(a_j|\theta_j) \right) \tau_i(a_i) u^i(a, \theta)$$

In words, this means that player i adopts (for each one of his possible types) the mixed strategy that is most rewarding, given his expectation regarding the types of opponents and their corresponding adopted mixed strategies.

The importance of the Nash equilibrium lies in the fact that, if the predicted behaviour of all the players in a game does not satisfy a Nash equilibrium, then there must be at least one player whose expected utility could be improved just by re-educating him to simply pursue his own best interests, without any other social change. So, a Nash equilibrium is an outcome that does not violate the assumption of *rational individual behaviour*. However, there exist situations where the pursuit of the individual best interest in a game, i.e., playing a Nash equilibrium, may lead to outcomes that are bad for all the players. A very well-known example of this is the *Prisoner's Dilemma* game (Figure 2.4). Moreover, the presence of *multiple* equilibria in a game gives rise to the *equilibrium selection* problem: if players choose strategies that belong to different Nash equilibria, the resulting strategy profile is not a Nash equilibrium [Mye91]; and, further, it would be preferable that the most efficient equilibrium for all players be played, and thus the issue of coordinating the players' actions so that they play this "better" equilibrium arises.

2.2.3 Repeated and Stochastic Games

A *repeated game* is a game made up from iterations of a single strategic form game. The strategy space of such a game is much richer than that of the single strategic form game. This is because a strategy for a player in a repeated game is a rule for determining his move in every round as a function of the history of moves that have been used at every preceding round [Mye91]. So, each player is able to choose a different stage-game strategy to play in each repetition, since he has access to state information about the number of the current iteration, and also about the other players' previous choices of stage-game strategies. However, in each iteration of the normal form game, players cannot know the actions chosen by other players *in that* iteration.

A repeated game may be either *finitely repeated* or *infinitely repeated*. In the first case, the players' payoffs are calculated at the end of the game, and each player may choose his optimal strategy by *backward induction*. In the case of infinite repetition, *discounting* of future payoffs is often[15] used: the value of a strategy after each choice of action in a particular game iteration is defined as the the payoff of that game iteration plus the sum of the payoffs of all future games, suitably discounted. In infinitely repeated games, players usually adopt strategies that are strongly dependent on the past behaviour of their opponents.[16]

Note that a repeated game deals only with one single state, since the same game setting is repeated in each iteration. Repeated games with *multiple* states are called *stochastic games* [Sha53]. A stochastic game is a tuple $\langle S, N, A_1, ..., A_n, p_T, r^1, ..., r^n \rangle$. S is a finite set of states, N is a finite set of n players, A_i is a finite set of actions available to player i, $p_T(s, \alpha, t)$ is a transition model that captures the probability of reaching state t after executing the *joint* action α at state s and $r^i(s, \alpha)$ is the real-valued payoff function for player i. A joint action is a vector of individual players' actions. Two-player stochastic games where $r^1(s, \alpha) = -r^2(s, \alpha)$ for all s and α are called *zero-sum* games, while when the sum of the payoffs is not restricted to 0 or any other constant the game is called *general-sum*. Figure 2.5 depicts a single-state zero-sum and a single-state general-sum game.

A strategy $\pi_i = (\pi_i^0, ..., \pi_i^t, ...)$ for a player i in a stochastic game is defined over the whole

[15] An average reward optimality criterion is sometimes used as well.

[16] Consider for example the well known *tit-for-tat* or *Grim trigger* strategies and their application to the repeated Prisoner's Dilemma game [Mye91]. A player following the tit-for-tat strategy would choose to cooperate with his opponent, until his opponent deviates, in which case he would deviate as well in the next repetition and as long as his opponent deviates, but would be "forgetful" as soon as his opponent started being cooperative again. A player following the Grim trigger strategy would be cooperative as long as his opponent is cooperative, but in the case his opponent deviates once, he would "punish" him by "confessing" forever.

	Rock	**Paper**	**Scissors**
Rock	$0, 0$	$-1, 1$	$1, -1$
Paper	$1, -1$	$0, 0$	$-1, 1$
Scissors	$-1, 1$	$1, -1$	$0, 0$

	Left	**Right**	**Down**	**Up**
Right	$25, 0$	$0, 0$	$0, -10$	$100, 100$
Down	$30, 0$	$10, 20$	$10, -20$	$-10, 20$

Figure 2.5: Two repeated games: The Rock-Paper-Scissors zero-sum game, where, for each joint action, the column player receives the negative payoff of the row player; and a general-sum game, where the sum of the players' payoffs per action is not zero.

course of the game. π_i^t is called the *decision rule* at time t. A decision rule tells the player what action to play in each state of the game, should the state be reached. In stochastic games, a strategy is often called a *policy*. Assuming (as we implicitly did when defining the decision rules above) that a policy relies on no history information, the policy is *Markovian*. If π_i^t is fixed over time then π_i is a stationary policy. A set of strategies for all n players, $\pi = \pi_1 \times ... \times \pi_n$, is called a strategy vector. The vector of strategies for all players except player i is denoted by π_{-i}. The objective of each player in a stochastic game is to maximize a discounted sum of rewards. It should be noted that there do exist equilibria solutions for stochastic games. However, this is a non-trivial result, proven by [Sha53] for zero-sum games and [FV97a] for general-sum games. Furthermore, the value of the Nash equilibria in a zero-sum game is always *unique*.

Given the analogies between the MDPs and stochastic games definitions, *stochastic games constitute a game theoretic framework that extends simple strategic games to MDP-like environments*. We will return to this issue shortly.

2.2.4 Learning in Games

Learning in games involves the modeling of the processes by which players change, in the course of time, the strategies they are using to play a game. A starting point for this could be to imagine some players playing a game repeatedly and trying to learn to anticipate the play of others by observation of past play. However, a difficulty that is inherent to the problem is that players should consider that their own current play might influence the future play of their opponents. For example, the repetition of an action over and over again can lead to the eventual adoption of a best response to that action by an opponent (a fact which under different game settings might well be beneficial or damaging to the first player). In any case, this process in which—possibly up to an extent—sophisticated and rational players of a game try to learn and play optimally over time can provide one explanation of equilibrium; an equilibrium can be considered to be the long-run result of this process [FL98].

Fictitious Play

One well-known and quite simple learning model for repeated games is *fictitious play* [Bro51]. According to fictitious play, the players observe the results of their (own) past encounters and then play a best response to the historical frequency of play.

In more detail, the simplest form of fictitious play requires that each player i keeps a count $C^j_{a^j}$, for each opponent j and $a^j \in A_j$, of the number of times agent j has used the action a^j in the past. For each opponent j, i assumes j plays action a^j with probability

$$Pr^i_{a^j} = \frac{C^j_{a^j}}{\sum_{\tilde{a}^j \in A_j} C^j_{\tilde{a}^j}}$$

Player i updates the counts —which reflect his beliefs regarding opponents' play— appropriately, and, at each iteration, he plays a best response to this empirical mixed-strategy profile.

One way to interpret fictitious play is to note that it corresponds to Bayesian inference when player i believes that the play of each one of his opponents corresponds to a sequence of i.i.d. multinomial random variables with a fixed but unknown distribution, and player i's prior beliefs over that unknown distribution take the form of a Dirichlet [FL98]. Weights are assigned to initial fictitious play beliefs, in order to represent preference to some strategies; this is analogous to setting the "prior counts" of the outcomes of a Dirichlet distribution.

One problem that is exhibited by the fictitious play model is its inherent discontinuity: a small change in the data can lead to an abrupt change in behaviour. This is a reason that triggered the development of stochastic variations of the traditional (deterministic) process of fictitious play; those variations enable the players to randomize when they are nearly indifferent [FL98]. *Smooth fictitious play* is a stochastic variation on fictitious play in which players use a smooth approximation to the best response, a best response distribution $\overline{BR}$, instead of the best response itself. It has been proved that smooth fictitious play converges to profiles that approach Nash equilibria in all games where those profiles are global attractors for the continuous-time smooth fictitious play process (i.e., the sequence of near-best responses to the empirical mixed-strategy profile) [FL98]. It's worth mentioning that there is a variation on (*smooth*) fictitious play, the "stimulus-response" model, that ignores information about the opponent and uses only own payoff information. This variation constitutes a reinforcement learning method.

Identical interest games form a special category of stochastic games in which the players' payoff functions are identical. Identical interest games exhibit the particularly interesting *ficti-*

	A	**B**
A	2	0
B	0	1

Figure 2.6: A simple fully cooperative game. It holds $u_1(\boldsymbol{\sigma}) = u_2(\boldsymbol{\sigma}),\ \forall \boldsymbol{\sigma}$. A game with a reward structure such as this one is commonly referred to as a "coordination" game.

tious play property; that is, every fictitious play process (i.e., every sequence of best responses to the empirical mixed-strategy profile) within a game with identical interests *converges in beliefs to equilibrium* [MS96]. A process converges in beliefs to equilibrium if, $\forall \epsilon > 0$, the sequence of beliefs (regarded as mixed strategies), is within distance ϵ from the set of equilibria after a sufficient number of iterations. A mixed strategy profile $\boldsymbol{\sigma}_\epsilon \in \Sigma$ is in ϵ-equilibrium if $\forall i \in N$, $u^i(\boldsymbol{\sigma}_\epsilon) \geq u^i(\tilde{\sigma}^i, \boldsymbol{\sigma}_\epsilon^{-i}) - \epsilon$, $\forall \tilde{\sigma}^i \in \Sigma_i$. Randomization over best responses or ϵ-best responses may be employed as a tie-breaking rule in order to ensure the convergence of a fictitious play process [Bou96a].

Identical interest games are also known as *fully cooperative games* (or "common value games", or "coordination games"). As all repeated games, a repeated fully cooperative game can be viewed as a degenerate case of a stochastic game having only one state. Since the game is fully cooperative, each player's reward is drawn from the same distribution reflecting the utility assessment of all players, and thus only one matrix is needed, listing only one utility in each cell (Figure 2.6).

Fully cooperative games have been broadly used in learning research dealing with agent cooperation. We will review several relevant approaches in subsection 2.2.5.

Rational Learning and Convergence to Equilibrium

Rational players are those who try to maximize their expected gains, using their beliefs about future behaviour of their opponents. To be able to *predict* the future behaviour of the opponents means to be able to give a nearly accurate forecast of the probability with which the opponents will take various actions. Can rational players really learn to play a repeated game? Can they achieve optimality in their play, and therefore end up in a Nash equilibrium? The answers to these questions are seemingly contradictory.

We have already seen that fictitious play in games of identical interest leads to Nash equilibrium [MS96]. Furthermore, Kalai and Lehrer show in [KL93] that if the prior beliefs of each player contain a "grain of truth", that is, they put positive probability on the actual repeated

game strategies of the opponent, then, by using *Bayesian updating* of these beliefs, players learn to predict with probability 1. Therefore, if the players choose at all times their best responses to their beliefs, they must eventually play according to a Nash equilibrium. There could be interpretations of this result that would lead one to believe that it is valid to claim that Nash equilibrium behaviour is a necessary long-run consequence of optimization by cautious players. As Nachbar points out in [Nac97] and [Nac01], however, this is not the case, since the "grain of truth" condition may be very difficult to satisfy in practice. He shows that in games with Bayesian players it is difficult to identify any plausible family of beliefs such that the players' best response strategies are in support of their beliefs.

In addition to that, Foster and Young [FY01] argue about the *impossibility of predicting the behaviour of rational agents*. They prove that rationality and prediction are incompatible and they do that without placing any restrictions on the players' prior beliefs, their learning rules, or the degree to which they are forward-looking. They show that there are very simple games of incomplete information in which the players never come close to playing a Nash equilibrium. The reason is that trying to predict the next-period behaviour of an opponent, a rational agent must take an action this period that the opponent can observe, which may cause him to alter his next period behaviour, thus invalidating the first agent's prediction. So, in order for the players to be good predictors, at least one must be intending to play a mixed strategy, and the other one must be able to predict this. If the players are myopic, they cannot learn mixed equilibria in the first place, no matter what their beliefs are. However, the incompatibility result between prediction and rationality is proved for the more general case of forward-looking players. There exist games at which the predicted probability of some action next period differs substantially from the actual probability with which the action is going to occur. Nevertheless, *an observer may conclude that the system is being led to an equilibrium*.

So, the results presented in the previously mentioned papers don't really contradict each other. The *prediction by the players* is what is problematic. To an observer, though, the average behaviour of the players may exhibit empirical regularities; the players' average behaviour may *mimic* Nash equilibrium from the observer's standpoint. However, this does *not* imply that individual players ever play Nash equilibrium strategies or learn to predict [FY01].

Finally, it is worth mentioning here that there exists a class of *no-regret* learning algorithms (e.g. [FV97b, GJ03]) that can be shown to converge in repeated games—not to Nash, but to related equilibrium notions, such as the *correlated equilibrium* (CE) solution concept. Unlike Nash, CE assumes dependencies among the agents' probabilistic strategies: a CE is a probability distribution over the joint space of actions, with the agents optimizing with respect to

one another's probabilities of actions, conditioned on their own [Mye91, GH03].[17] A no-regret learning algorithm generates non-deterministic actions for an agent, such that, over time, the strategies generated by the algorithm outperform any other fixed strategy, in terms of average cumulative payoff—thus, the agents experience no regret for following the algorithm's recommendations.

Convergence of gradient dynamics for general-sum repeated matrix games: A common class of algorithms within machine learning is the one which contains algorithms that proceed by gradient ascent or descent on some appropriate objective function. *Gradient ascent in expected payoff* can be used by players in order for them to adapt their behaviour while participating in a repeated game. Regretably, the game theory literature does not make any assertions about the convergence of strategies computed by gradient ascent.

However, Singh, Kearns and Mansour [SKM00] examined a simple gradient ascent method ("Infinitesimal Gradient Ascent"—*IGA*) and were able to exhibit some convergence results. According to their method, which assumes a full information game, a player adjusts his strategy after each game iteration (i.e., modifies the probability with which each action is played) in the direction of the current gradient of his expected payoff, with some step size; in this way, the strategy is adjusted so that the expected payoff is increased. The main finding of the IGA analysis is the proof that, if both players in such a game follow IGA, then their strategies will converge to a Nash equilibrium *or* the average payoffs over time will converge in the limit to the expected payoffs of a Nash equilibrium. However, at any moment in time the expected payoff of a player could be arbitrarily poor. This may make it difficult to evaluate the learner, and could also be problematic when applied with temporal differencing for multiple state games, which assumes that past expected payoffs predict future ones.

Following [SKM00], Bowling and Veloso [BV01a] improved the IGA approach so as to satisfy a stronger notion of convergence to Nash equilibrium. Instead of considering the step size taken to the direction of the gradient as constant, [BV01a] introduces a *variable* learning rate for gradient ascent, and uses the *"Win or Learn Fast" (WoLF) principle* in order to regulate the learning rate: the learning rate is modified so that the agent will learn cautiously when winning and quickly when losing. It is proved that when players are following the "WoLF-IGA" approach, both strategies and expected payoffs converge to Nash equilibrium (for 2-

[17] The CE solution concept implies that there exist communication possibilities between the agents and a "mediator" that "recommends" strategies to the agents, drawn by a distribution that is common knowledge [Mye91]. Because the set of CEs is a convex polytope, CE can be computed easily via linear programming.

player, 2-action, full-information iterated general-sum games).

2.2.5 Reinforcement Learning in Games

The previous subsection shows the strong bonds between game theory and learning, and shows that game theory can be combined with learning in a multiagent environment. We have so far focused on *strategy* learning. Now we describe the bonds between game theory and reinforcement learning, focusing on the multiagent reinforcement learning (MARL) problem from a stochastic games perspective. As a result, the immediate goal here will not be strategy learning per se, but rather the maximization of the agents' long term discounted rewards. We will define the MARL problem, making explicit reference to approaches that attempt to solve it under a principled game-theoretic framework. However, some non-game theoretic approaches to MARL will be presented here as well.

Multiagent Reinforcement Learning

The obvious way to think about the *multiagent* reinforcement learning problem is to consider it as the extension of the reinforcement learning problem. Under this perspective, a "naive" definition of the MARL problem could be the following: "Multiple agents exist in a common environment and try to achieve their long-term utility maximization goals while learning by using RL techniques". After all, RL seems to be well suited for multiagent systems where agents know little about other agents. Simply applying single-agent RL algorithms in multiagent environments may seem an effective way to provide an "answer" to the MARL problem.

However, the approach just mentioned constitutes rather only *one aspect* of the MARL problem, and of the attempt to "solve" it. If this approach is considered in isolation, one runs the danger of treating other agents of the system as part of the environment, ignoring the difference between responsive rational agents and passive environment. *A second aspect* of the multiagent reinforcement learning problem therefore emerges. It's the aspect that tries to deal with the question: *Does (and: "In which way? To what extent?") the co-existence of multiple agents within a setting affect their learning capabilities?*

In the following subsections we will present attempted "solutions" to the MARL problem dealing with both of these aspects. We should, however, state in advance that it is apparent to us that in a multiagent world an agent has to learn how to align its action choices with those of other agents, because the effects of one's actions are directly influenced by the actions of others. This could be achieved by game theoretic techniques within a stochastic games

framework; within this framework, the MARL problem can be more accurately defined. In game-theoretic terms, then, the definition of the MARL problem is as follows:

Definition 1. *The multiagent reinforcement learning (MARL) problem is the problem of multiple agents trying to maximize their expected discounted total reward, while co-existing within a stochastic game environment whose underlying transition and reward models are unknown.*

Reinforcement Learning in Stochastic Games

A description of the stochastic game framework was provided in 2.2.3. Here we will be dealing with RL in stochastic games. The goal of each agent is to learn a Markovian and stationary[18] though possibly stochastic strategy, that maps states to a probability distribution over the possible actions, such that the agent's discounted future reward will be maximized.

To expand a bit on the MARL problem definition, a multiagent world can be viewed as a game with multiple players. Each state in a stochastic game can be viewed as a matrix game with the payoffs for each joint action determined by the reinforcement given to each agent at this state for this joint action. After playing in the current state's matrix game and receiving the payoffs, the agents are transitioned to another state according to a distribution dictated by their joint action. As was mentioned before, there exist Nash equilibria solutions for stochastic games. It's quite valid, therefore, to expect that the long-term solution for the agents' reward maximization problem may coincide with learning to playing such an equilibrium, given that a Nash equilibrium is a self-enforcing solution concept that endorses "rational" individual behaviour. However, sub-optimal results (in terms of utility maximization) are quite possible in a multiagent world. This is because of the equilibrium selection problem, and the uncertainty regarding the strategies of the opponents that results to a multiagent extension of the classical exploration vs. exploitation problem: should an agent explore in order to gather more information about the strategies of its opponents, or should it just exploit its current knowledge regarding those strategies?

In a nutshell, we can make two observations that place multiagent reinforcement learning in this perspective:

Observation 1. *Stochastic games are an extension of MDPs to multiple agents, and an extension of matrix games to multiple states. Stochastic games essentially are n-agent MDPs.*

[18]The strategy is considered to be stationary and Markovian for simplicity purposes. This is the assumption commonly used in current literature.

Therefore, the solution to the multiagent reinforcement learning problem can be sought within the stochastic games framework.

and

Observation 2. *A multiagent reinforcement learning method should explicitly take other agents into account.*

Single-agent RL methods used in a multiagent world—no matter how effective they may be under specific domains—lack the potential of exhibiting optimal behaviour in terms of the declared goal of maximizing discounted reward, since they basically ignore the interactions between the various agents and reason about them only implicitly: modeling all other agents as part of the environment has the drawback that the agent cannot capture the fact that in most situations the behaviour of others is influenced by his own presence. Placing the problem under a game theoretic framework, on the other hand, enables one to explicitly monitor other agents and reason about their expected future behaviour which is certain to have an impact on the outcomes of agent interactions. Results presented in several of the papers we review here suggest the validity of the second observation above (e.g., [Lit94, BV01b, DFR98, Lit01, LR00]).

A Brief Review of Approaches to the Problem In a paper that was the first to introduce stochastic (a.k.a. "Markov") games as a framework for MARL, Littman [Lit94] describes a Q-learning-like algorithm for finding optimal policies in *2-player zero-sum* stochastic games. The algorithm is called *minimax-Q*; essentially, Q-learning is used by each player, with the basic difference that the value function is evaluated using a "minimax" approach. When using the minimax approach, a player uses a strategy designed to maximize his minimum payoff (i.e., the payoff he would receive should the opponent do his best to minimize it; minimax is a rather conservative approach).[19] The notion of the traditional Q-function is extended to maintain the value of joint actions, and linear programming is used to find the equilibrium of the games. The approach is well-suited to zero-sum games, given that in those games one agent's gain is the other agent's loss (thus, the interests of the agents are strictly opposite). Not surprisingly, in Littman's experiments minimax-Q learners did better than agents using standard single-agent Q-learning. A criticism [BV01b] that can be applied to minimax-Q is that it is not "rational",

[19] Actually, the minimax approach was well-studied within the framework of *learning automata*, long before [Lit94]. Learning automata are simple low memory machines for solving a selectional (reinforcement) learning problem known as the *n-armed bandit* [BF85].

in the sense that it will always learn and play the (unique) equilibrium, independently of the opponent's policy; even when, for example, the opponent is "irrational", the minimax-Q player will still be conservative.

Hu and Wellman [HW99, HW98] extended Littman's work, by dealing with the MARL problem within a *general-sum* (rather than zero-sum) stochastic games framework. They design a multiagent Q-learning method ("Nash-Q") under this framework and prove convergence to Nash equilibrium under specific conditions, the most restrictive of which limits the structure of the intermediate bimatrix games encountered during learning. A quadratic programming solution is used to find the equilibrium of the games, with each agent trying to learn the Q-values corresponding to playing a mixed Nash equilibrium strategy. Of course, in order for the equilibrium to be derived, each agent *needs to maintain Q-value tables for all the other agents*. Not only does this induce a computational burden, it also requires that each agent can observe the other agents' immediate rewards, which is a rather unrealistic assumption in a general-sum game. The algorithm does find an optimal strategy whenever there exists a unique Nash equilibrium in the game; however, to prove convergence to the equilibrium they assume that a one-stage game with multiple equilibria is never encountered during learning — a restriction *not* satisfied by *any* non-zero-sum or non-fully cooperative games. When more than one Nash equilibria exist, the algorithm is not guaranteed to converge, and cannot be useful by itself; it should be combined with techniques that help overcome the equilibrium selection problem. Finally, the same criticism of "irrationality" [BV01b] that was applied to Minimax-Q also applies to the Nash-Q. In any case, this work did contribute to establishing the theoretical foundations for applying RL to multiagent systems from a game theoretic perspective.

Another multiagent RL algorithm that attacks general-sum games is the "Friend-or-Foe Q-learning" (FFQ) algorithm [Lit01]. In FFQ, the learner is given the additional information of which of two kinds of opponents to expect: friends (opponents whose actions will help both agents to maximize their possible rewards) or foes (opponents whose actions are such that only an adversarial equilibrium may be reached, an equilibrium, that is, in which no player is hurt by any change of the other players behaviour).[20] In 2-player games, if the opponent is considered a friend, ordinary Q-learning in the combined action space of the agents is used; otherwise, a minimax-Q approach is adopted. In the latter case, the agent is guaranteed to achieve its learned value independent of the opponents' action choices. In n-player games, the formulation of the Q-value learning algorithm is a minimax-Q formulation — the agents

[20] In the general case, the algorithm is actually mixing friends and foes.

who are agent i's friends are assumed to work together to maximize i's value, while its foes are collectively trying to minimize that value. Only one Q-function needs to be maintained by each agent, unlike Hu and Wellman's approach. FFQ is guaranteed to converge, but the values learned by it do not necessarily correspond to those of a Nash equilibrium policy, unless the game is fully cooperative or zero-sum. Another issue is that the agents have to be told whether they are facing friends or foes. Furthermore, Friend-Q is in many cases incapable of achieving the highest possible learned value in the presence of multiple equilibria.

Greenwald and Hall's "Correlated-Q Learning" (CE-Q) [GH03] is a multiagent RL algorithm that is based on the correlated equilibrium (CE) solution concept discussed earlier. The set of CE contains the set of Nash equilibria, and thus CE-Q—which requires that the agents play strategies that belong to some calculated CE equilibrium—generalizes both FFQ and Nash-Q. Even though CE-Q learning is shown to empirically converge to CE equilibria (in experiments with 2-agent games), no proof for its convergence has been provided—in the presence of multiple equilibria, it too suffers from the equilibrium selection problem.

Bowling and Veloso [BV01b] introduce two desirable properties for a multiagent learning algorithm: *rationality*, i.e., convergence to best-response policy for stationary policies; and *convergence*, i.e., convergence to stationary policies. An RL algorithm is presented, *"WoLF-Policy Hill-Climbing" (WoLF-PHC)*, which has both properties in a variety of games; it is rational and it is *empirically* shown to converge to mixed policy equilibria. WoLF-PHC works by applying a variable learning rate and the "Win or Learn Fast principle" [BV01a] to another algorithm introduced in [BV01b] which is named *"Policy Hill Climbing" (PHC)* and performs hill-climbing in the space of mixed policies. PHC is essentially a Q-learning algorithm that maintains the current mixed policy; it is rational and can play mixed policies, but does not converge. WoLF-PHC is considered to be a state-of-the-art MARL algorithm, shown to perform well in a variety of experimental domains; however, no theoretical guarantees of convergence yet exist.

In one of the rare contributions to MARL research deriving from the game theory community, Jehiel and Samet [JS01] address the problem of learning to play games "by valuation". The notion of *valuation* (i.e., the assignment of numeric values to different moves in the game) is used to reflect the desirability of the moves to the players. This is another way, of course, to talk about the quality of a state-action pair. A multiagent value iteration-like approach is presented, but the setting is drawn with the use of game-theoretic terminology. An important contribution of the paper is the provision of convergence results for the RL process presented. Jehiel and Samet show that a player who has a winning strategy in a *win-lose* game (i.e. a game

with only two payoffs—a winning and a losing one) can be guaranteed a win in the repeated game that embeds iterations of the win-lose game, by updating the value of each state so that it coincides with the payoff obtained in this round and by simply following a greedy exploration strategy.[21] When a player has more than two payoffs, they present a learning procedure that associates with each state the average payoff in the rounds in which this node was reached, and uses an ϵ-greedy exploration policy. This approach of exploiting past experience is reminiscent of the "stimulus-response" variation of the smooth fictitious play, since it does not use information on the opponent, just own payoff information. They proceed to show that, when all players adopt this procedure, with some perturbations, then, strategies that are close to *subgame perfect* equilibrium[22] are played after some time, with probability 1. However, a single player who adopts this procedure can guarantee only his individually rational (i.e., his "maxmin") payoff, which is what he can be guaranteed even if all other players are disregarded. In addition, since the method treats separately the valuation for every state, it becomes unrealistic for large state spaces because then the chance of meeting a given node several times is too small.

Finally, Huang and Sycara present in [HS03] two RL algorithms for multiagent learning in extensive form games with complete information and a unique subgame perfect equilibrium, one of which—the "Multiagent Q-Learning" method—is somewhat related to the approach of Jehiel and Samet. The second of their methods, "Multiagent Learning Automata", uses the reward obtained at the end of a game episode to reinforce the strategy (by properly updating its probability of being chosen) followed at a node of the game—instead of reinforcing the values of node-action pairs. Both of these algorithms are proved to converge to the subgame perfect equilibrium of the game.

We have seen that MARL algorithms usually set the goal for the agents to find their policy in the game's equilibrium solution. One problem with this is that providing theoretic proofs for convergence to equilibrium is non-trivial [BV00]. More importantly, however, as has been accutely pointed out by some authors (e.g., [Bou96a, Bou99, SPG04]), convergence to an equilibrium is valuable *if and only if* it serves the goal of maximizing payoff. Things get really difficult when it comes to dealing (as is often the case) with multiple equilibria. Closed-form solutions, even though they seem to be providing the assertion of opponent-independent algorithms, do not suit a problem with multiple equilibria. This is because the Observation 2 above holds: the optimal policy of one agent depends on the policies of the opponents. In fact, equi-

[21] However, this is *not* equivalent to saying that *the player's strategy is what guarantees him the victory* [JS01].

[22] A subgame perfect equilibrium of an extensive form game is a strategy vector in which every player's action is optimal for this player in every subgame of the game [Mye91].

librium selection affects the policy and the value of the state, while the value of the state affects equilibria of the states that transition into it. The number of equilibrium solutions in stochastic games can grow exponentially with the number of states, and thus equilibrium selection after learning is not feasible.

A possible remedy to those problems is to observe the others and adapt to their behaviour. This could be achieved by smooth fictitious play and other opponent-modeling or opponent-dependent methods, such as a method that utilizes *joint action learners (JALs)* [CB98], i.e., agents that learn values for joint actions, as opposed to individual actions. We will return to those issues and to JALs shortly.

Cooperative and Coordinating Agents and Reinforcement Learning

There are cases when agents occupying a system need to cooperate and coordinate in order to achieve their goals. General problems involving the interaction of agents with identical interests that have to cooperate in order to achieve a common goal naturally arise, for example, in task distribution. It is easy to imagine a fully cooperative set of agents representing a user — and sharing, therefore, the user's utility function — which have to collectively act to the common desired end. It is not surprising, thus, that the application of learning to the problem of coordination in multiagent systems has become popular in both AI and game theory.

There exist basically three ways to achieve coordination in a multiagent setting. One obvious way is using communication among the agents (e.g., [Tan93, YS93, JG00]). Sycara et al. [SZ96, SDP+96] have developed the RETSINA (Reusable Environment for Task Structured Intelligent Network Agents) framework, an open multiagent system supporting communities of heterogeneous agents, which provides many challenges and opportunities for agent collaboration and coordination. The framework is distributed, with agents entering and leaving the system dynamically. The agents are frequently in need to decide how to collaborate with others in order to decompose and execute tasks, based on their capabilities. This is achieved through inter-agent communication, facilitated by the architecture of the RETSINA system, which allows the agents to take up different roles (in order to interact with users, gather and pass around information, or execute tasks).

A second way to coordinate is *through the introduction of "social conventions" or rules*. This is natural, since the multiagent coordination procedure is a highly "social" one: the needs of others in the environment cannot simply be ignored. Thus, there exists a corpus of work which tries to address questions that focus exactly on this "social" aspect of multiagent coordi-

nation. Related approaches, which make use of social rules—or, more generally, use the social environment to facilitate coordination, and also investigate the benefits of "social behaviour"—include [ST92, WW95, Mat94a, Mat94b, Bal97].

In this thesis we are mostly interested in approaches that achieve *coordination through the use of learning*. In [SSH94], for example, there is no use of information sharing; instead, reinforcement learning techniques are employed so that eventual coordination is achieved. The agents use standard Q-learning, and demonstrate an ability to coordinate while acquiring complementary problem-solving knowledge (i.e., they learned to perform policies that were complementary—they could be combined effectively in order to achieve the agents' goal). Avoidance of information sharing is certainly beneficial, since it enhances adaptability to changing and noisy environments. However, there exist drawbacks in this specific work, such as convergence to sub-optimal policies due to incomplete exploration of the state space, and slow convergence unless the system parameters were chosen with great care.

Another related approach is the "Coordinated Reinforcement Learning" framework presented in [GLP02], within which only limited communication is required between collaborating agents, in order for them to efficiently select an optimal joint action, without them explicitly considering every possible action; the latter would be impossible in an exponentially large action space. The agents have only partial access to the state description. *Structured communication and coordination* of agents is used in the core of *both* the learning algorithms and the execution architectures (policy search phase) of the RL methods presented in the paper—communication needs are reduced by solving a "coordination graph" that is constructed exploiting the local structure of Q-functions corresponding to the agents. Learned policies can be executed in a distributed manner. However, the issue of potential changes in the structure of the problem—that would require agents to be inserted or deleted dynamically—is not addressed.

None of the aforementioned approaches addresses the problem of achieving coordination of multiple reinforcement learners from a game theoretic perspective (even though game theory has in fact identified or dealt with questions related to the benefits of cooperation [Axe84]). Nevertheless, taking a game theoretic view at the problem of coordination, we could argue that it could be cast as a problem of performing equilibrium selection in a cooperative repeated game, whereas *coordinated action choice can be learned through repeated play of the game.*

Boutilier [Bou96b] has discussed the use of learning to achieve coordination introducing a framework of *Mutiagent Markov Decision Processes (MMDPs)*, which actually are *n-person stochastic games*. He showed that decomposition of sequential decision processes can be employed so that coordination can be learned (or imposed) locally, at the level of individual states.

He has also discussed elsewhere ([Bou96a]) both a Bayesian learning approach and a learning model resembling fictitious play using *likelihood estimates* of opponents' actions, in order to achieve equilibrium selection in the case of unobservable actions. Moreover, he points to a way to add conventions to the learning model through the use of *maximum likelihood estimates* with the eventual goal to drop learning altogether after a while and achieve convergence to a *conventional* equilibrium - reducing, thus, the computational burden associated with ongoing computation of best responses [Bou96a]. In addition, Boutilier develops in [Bou99] an extension of value iteration, that allows for the systems's state space to be expanded dynamically to account for the coordination protocol used; thus, the agents are able to decide to engage or avoid coordination problems based on expected value. However, the focus of this research is more on coordination mechanisms and less on the employment of reinforcement learning in coordination games.

Convergence to an Optimal Equilibrium No matter how successful their application has been in empirical studies such as those presented so far (e.g., [SSH94, Tan93]), standard single-agent RL models are not usually theoretically justifiable as convergent methods for multiagent coordination in general. On the other hand, even in work where general-sum stochastic games have in fact been used as a general MARL framework [HW98], the coordination problem has been simplified by assuming a unique equilibrium. However, when more than one equilibrium strategy exists, coordination becomes a challenge.

Claus and Boutilier proved in [CB98] that in cooperative repeated games, under a set of conditions—which are basically the requirements that an agent samples each one of its actions infinitely often and that its exploration strategy is exploitive—it is assured that *Joint Action Learners* eventually play a (deterministic) equilibrium strategy profile, but *not necessarily* the optimal one. This entails several advantages, when behaving well *while* learning is important (i.e., when the discounted infinite-horizon model of optimal behaviour is used). More recently, some model-free techniques have been proposed to enforce coordination to optimal equilibrium in identical interest games—that is, convergence to the equilibrium with the maximum reward (identical to both agents) [KK02, LR00, WS02]. These approaches, however, do not take at all into account the fact that convergence to optimal equilibrium strategies may imply a substantial cost to be payed by the agent ([LR00, WS02]), which is in contradiction to those methods' forestated goal of discounted accumulated reward maximization, or, in addition, are not theoretically well-founded and generalizable ([KK02]). We critique those papers further in Chapter 3.

2.3 Cooperative Game Theory: Coalition Formation

Cooperative game theory deals with situations where players act together in a cooperative equilibrium selection process involving some form of bargaining, negotiation, or arbitration [Mye91]. The problem of *coalition formation* is the fundamental area of study within cooperative game theory.

Let $N = \{1, \ldots, n\}$, $n > 2$, be a set of players (or "agents"). A subset $C \subseteq N$ is called a *coalition*, and we assume that agents participating in a coalition will coordinate their activities for mutual benefit.[23] A *coalition structure* (CS) 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 problem* can be seen as being composed of the following activities [SL97]: (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. The aforementioned activities interact with each other: the agents should reach an agreement on those issues through negotiations. Furthermore, it should be noted that one distinguishing feature of cooperative game theory is the ability of the agents to *negotiate effectively*, meaning that if there were a feasible change in the strategies of the members of a coalition (or a feasible change in the coalition structure) that would benefit the negotiating agents, then they would actually agree to make that change [Mye91]. Finally, as mentioned in Section 1.2 above, and as will become more apparent later in this thesis, coalition formation can be viewed under both a cooperative (when the focus is on the final result of negotiations) and a non-cooperative standpoint (when the focus is on the negotiation process itself—i.e., on the coalitional bargaining problem); thus, one can refer (as we do in this thesis) to "cooperative coalition formation" or to "non-cooperative coalition formation".

There exists an ever-growing corpus of literature dealing with the coalition formation problem, both from a purely game-theoretic (e.g. [Aga97, BS00, CDS93, DS98, Eva97, HMC96, MW95, Oka96, KR02, PR94, SV97, SBWT99, SB99, Yan03]) and from a more AI-related (e.g., [AL04, MCW04, KG02, KST03, KST04, SK98, SL04]) point of view. Coalition formation ideas have been applied to problems as diverse as multilateral bargaining and resource allocation (e.g., [KST03, KST04, SL04, DJ06]), agent coordination for task execution (e.g., [SSJ97]), grid computing and e-business (e.g., [AL04, NPC+04, PTJ+05]), e-marketplaces

[23] Seeking "mutual benefit" does not imply that the agents are not *individually rational*—i.e., seeking to maximize their own individual payoffs by participating in coalitions. This will become more evident shortly.

(e.g., [YS01, LS02, LCRS03]), multisensor networks (e.g., [DDRJ06]) or even cryptography [ZR94]. However, most of the existing work does not deal with the problem of uncertainty (or, more specifically, *type uncertainty*) during coalition formation, as we do in this thesis. We will not be reviewing all related work in this section, but we will do so in subsequent chapters, as appropriate. Here we will just briefly present some basic coalition formation-related concepts.

2.3.1 Characteristic Function (Transferable Utility) Games

While seemingly complex, coalition formation can be abstracted into a fairly simple model, under the assumption of *transferable utility*, which assumes the existence of a (divisible) commodity (such as "money") that players can freely transfer among themselves. Thus, it is easy to to describe the possible *allocations* of utility among the members of each coalition, as it is sufficient to specify a single number denoting its *worth* (i.e., the total payoff available for division among its members).

This is the role of the *characteristic function* of a *coalitional game with transferable utility (TU-game)*: A characteristic function $v : 2^N \Rightarrow \Re$ defines the *value* $v(C)$ of each coalition C [vNM44]. Intuitively, $v(C)$ represents the maximal payoff the members of C can jointly receive by cooperating effectively. An *allocation* is a vector of payoffs $\boldsymbol{x} = (x_1, ..., x_n)$ assigning some payoff to each $i \in N$. An allocation is *feasible* with respect to coalition structure CS if $\sum_{i \in C} x_i \leq v(C)$ for each $C \in CS$, and is *efficient* if this holds with equality. The *reservation value* rv_i of an agent i is the amount it can attain by acting alone (in a *singleton* coalition): $rv_i = v(\{i\})$.

One important concept regarding characteristic functions is the concept of superadditivity. A characteristic function is called *superadditive* if any pair (C, T) of disjoint coalitions C and T is better-off by merging into one coalition: $v(C \cup T) \geq v(C) + v(T)$.[24]

When the transferable utility assumption is not in place, the coalitional games are called *non-transferable utility (NTU) games* [Mye91]. We will not be dealing with NTU games in this thesis.

2.3.2 The Core and Other Solution Concepts

When rational agents seek to maximize their individual payoffs, the *stability* of the underlying coalition structure becomes critical. A structure is stable only if the outcomes attained by the

[24] We note that we *do not* make any superadditivity (or any other additivity-related) assumption in the work presented in this thesis.

coalitions and the payoff combinations agreed to by the agents are such that both individual and group rationality are satisfied in some way. Research in coalition formation has developed several notions of stability, among the strongest being the *core* [Gil53, LR57, KR84]. The core of a characteristic function game is a set of *payoff configurations* $\langle CS, \boldsymbol{x} \rangle$, where each $\boldsymbol{x}$ is a vector of payoffs to the agents in coalition structure CS, which are such that no subgroup of agents is motivated to depart from its coalition in CS.[25] In other words, the core is the set of all possible allocations that can be "accepted" by all possible coalitions:

Definition 2. *Let CS be some coalition structure, and let $\boldsymbol{x} \in \Re^n$ be some allocation of payoffs to the agents. The* core[26] *is the set of payoff configurations*

$$core = \{\langle CS, \boldsymbol{x} \rangle | \forall C \subseteq N, \sum_{i \in C} x_i \geq v(C) \text{and} \sum_{i \in N} x_i = \sum_{C \in CS} v(C)\}$$

A core allocation $\langle CS, \boldsymbol{x} \rangle$ is both feasible and efficient, and 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, in many cases the core is empty, as there exist games for which it is impossible to divide the utility in such a way that the coalition structure becomes stable (as there might always be coalitions that could gain if they were given one more opportunity to negotiate effectively against the current configuration). Moreover, computing the core or even deciding on its non-emptiness is—in general—intractable [Rap70, Chv78, Tan91, DP94, SLA+99, CS03].

As mentioned in Section 1.2, dynamic coalition formation research is interested in the question of establishing endogenous formation processes that reach stable structures, such as structures in the core. Dieckmann and Schwalbe [DS98] provide such a dynamic formation process (a bargaining process that induces an underlying Markov process), which allows for the (conditional) destabilization, during some bargaining stage, of structures formed in previous stages, so that the dynamic process retains the potential to reach an absorbing, stable state. This destabilization is achieved through the random exploration of suboptimal bargaining actions (i.e., formation proposals and replies), and the process can be shown to converge to the (usual,

[25] We will sometimes refer to the core defined in Definition 2 as the "deterministic" core, since it assumes no form of uncertainty or stochasticity regarding partners or coalitional values. Analogously, we sometimes use the term "deterministic" to refer to the usual model of coalition formation (which admits no uncertainty).

[26] This core definition is basically the one provided in [SL97], and is also very similar to the one coined in [DS98]. It is more generic than other traditional core definitions which assume superadditivity, in that it considers the $\langle CS, \boldsymbol{x} \rangle$ configurations, rather than focusing only on the allocation of payoffs within the grand coalition (i.e., requiring that $\sum_{i \in N} x_i = v(N)$).

deterministic) core.

Suijs *et al.*[SBWT99, SB99] do not deal with coalition formation processes, but do describe a notion of the core concept under coalition value uncertainty—payoffs in their model are *stochastic*, and depend on the coalition action taken. To deal with payoff stochasticity, they use *relative* shares (i.e., shares described as percentages) for the allocation of the *residual* (or the "risk") of the stochastic coalitional value (i.e., actual payoff minus its expectation); however, the agents have *common expectations* regarding the coalitional values. [SBWT99, SB99] provide interesting theoretical results regarding the core concept under this restricted form of uncertainty. In Chapter 4, we will describe the approaches of [DS98, SBWT99, SB99] in some more detail.

In recent years, several papers in the game theory literature [PR94, MW95, HMC96, Eva97, SV97, Yan03] have tried to establish connections between the outcomes arising from equilibrium play during coalitional bargaining and the core of the underlying coalition formation problem. Broadly, the goal of this line of research is to show that the equilibrium payoff sets in particular coalitional bargaining games correspond to core allocations for the participating agents. The related results provide a further justification for the use of the core as a solution concept—i.e., they contribute to the *non-cooperative justification of the core*—and, more generally, describe forms of equivalence between cooperative and non-cooperative coalition formation solution concepts.

Apart from the core, there exist many other solution concepts, such as the *Shapley value* [Sha53] and the *kernel* [DM65]. The latter is a stability concept that combines individual rationality with team rationality, in the sense that it provides stability within a *given* coalition structure (and under a given payoff allocation): the kernel is a payoff configuration space in which each payoff configuration $\langle CS, \boldsymbol{x} \rangle$ is stable in the sense that any pair of agents i, j which are members of *the same* coalition in a specific CS are in equilibrium with one another, given payoff vector $\boldsymbol{x}$. Agents i and j are said to be in equilibrium if they cannot outweigh one another within their common coalition—in other words, neither of them can successfully claim a part of the other's payoff under the configuration $\langle CS, \boldsymbol{x} \rangle$. The kernel is *always non-empty*. In particular, for every CS for which there exists at least one allocation $\boldsymbol{y}$ such that all agents in CS receive at least their reservation value in $\boldsymbol{y}$, there also exists an allocation $\boldsymbol{x}$ such that the resulting configuration is in the kernel (we say that it is *kernel-stable*). However, note that the kernel merely determines the way payoffs should be distributed so that agents cannot outweigh their current partners given a specific CS. Thus, it is less generic than the core, while computing a kernel element is—in the general case—exponentially hard also. Blankenburg *et*

al.[BKS03] have recently coined a kernel stability concept under coalitional value uncertainty, introducing the *fuzzy kernel* for use in *fuzzy cooperative games*.[27]

We study in this thesis several cooperative and non-cooperative aspects of the coalition formation problem under the more general assumption of type uncertainty. Thus, in subsequent chapters, we introduce the *Bayesian core*, a core-like stability concept for coalition formation under type uncertainty (and payoff stochasticity), and also establish formation processes that enable convergence to stable coalition structures—under this model of uncertainty. Further, we deal with the problem of the non-cooperative justification of the Bayesian core, studying equilibria concepts for coalitional bargaining under type uncertainty, and establishing connections between their outcomes and allocations that lie in the Bayesian core. Finally, since learning can be of value to rational agents that seek to form rewarding coalitions in uncertain environments, we introduce a Bayesian RL framework to enable the agents to take informed decisions in scenarios of repeated coalition formation under uncertainty.

[27] In order to formally define the kernel, we have to define *the excess* of a coalition and *the surplus* of one agent over another:

Appendix Definition 1. *The excess of a coalition C with respect to a payoff configuration $(\vec{x}, CS)$ is defined as $e(C) = v(C) - \sum_{i \in C} x_i$. C is not necessarily a coalition in the specific CS, but may belong in any other coalitional stucture.*

Appendix Definition 2. *The surplus S_{ij} of agent i over agent j with respect to a payoff configuration $\langle CS, \boldsymbol{x} \rangle$ is defined by $S_{ij} = \max_{C | i \in C, j \notin C} e(C)$; in other words, it is the maximum of the excesses of all coalitions C that include i and exclude j, with C not in the current coalitional structure (since under the current coalitional structure agents i and j belong in the same coalition).*

We say that agent i outweighs agent j if $S_{ij} > S_{ji}$. If i outweighs j under $\langle CS, \boldsymbol{x} \rangle$, it can claim a part of j's payoff x_j. Individual rationality requires that $x_j > v(\{j\})$, where $v(\{j\})$ is the coalitional value of j in a singleton coalition. Two agents that cannot effectively outweigh one another are in equilibrium:

Appendix Definition 3. *Two agents $agent_i$ and $agent_j$ are in equilibrium, if one of the following conditions is satisfied:*

1. $S_{ij} = S_{ji}$
2. $S_{ij} > S_{ji}$ *and* $x_j > v(\{j\})$
3. $S_{ij} < S_{ji}$ *and* $x_i > v(\{i\})$

Now we are in the position to define the kernel as follows:

Appendix Definition 4. *A payoff configuration $\langle CS, \boldsymbol{x} \rangle$ is* K-stable *if $\forall i, j$-pairs of agents in the same coalition $C \in CS$ under $\boldsymbol{x}$, the agents i and j are in equilibrium. A payoff configuration is in the kernel iff it is* K-stable.

For a nice presentation of the kernel solution concept see, e.g., [SK99]; for a scheme used to compute kernel allocations see [Ste68].

Chapter 3

Bayesian MARL in Stochastic Games

In standard RL, the action selection problem an agent faces involves a tradeoff between *exploiting* what one knows about the effects of actions and their rewards and *exploring* to gain further information about actions and rewards that has the potential to *change* the action that appears best. In the multiagent setting, the same tradeoff exists with respect to action and reward information; but another aspect comes to bear: the influence one's action choice has on the future action choices of other agents. In other words, one can exploit one's current knowledge of the strategies of others, or explore to try to find out more information about those strategies. This is the *generalized exploration-exploitation tradeoff* in MARL.

Furthermore, the possible existence of *multiple equilibria* in the game in which the agents participate adds another aspect to this tradeoff: the agents should be able to coordinate their choice of equilibrium, or risk converging to undesirable equilibria.[1] This is the *equilibrium selection problem*. However, in a MARL setting, attempting to coordinate equilibrium selection may result to very poor rewards during the online learning period. This intensifies the need to address the tradeoff between long-term benefits and short-term costs, especially if, as is often the case in RL, the discounted reward criterion is used to evaluate performance.

In this chapter we develop a Bayesian, model-based MARL framework to tackle the issues above. We describe the solution to the generalized exploration-exploitation tradeoff as the solution to a system of Bellman equations over a belief state MDP. The Bayesian approach enables the agents to make informed rational decisions without requiring them to take explicit exploratory actions—rather, the *value of information* of the agents' actions, incorporated in the

[1] Notice that, viewing MARL under a game theoretic perspective (Definition 1), coordinating individual actions to some commonly desirable joint action is equivalent to choosing to play some equilibrium of the underlying game; when the commonly desirable action is the *most desirable* one, this is the problem of *coordinating equilibrium selection* so that some optimal equilibrium is played.

aforementioned equations, plays a critical role in determining an agent's policy.

However, as it is not feasible to provide an exact solution, we used computational approximations to tackle the problem. One of the methods we develop is the multiagent extension to a known Bayesian method (*VPI exploration*) introduced by [DFR98, DFA99].

We verify the value of our approach experimentally, applying it to the case of multiagent coordination, which is of particular interest because of the equilibrium selection problem. Our results show that the Bayesian approach outperforms other approaches (some of which enforce convergence to optimal equilibria).[2] Bayesian agents make cautious and informed decisions, exhibiting good online behaviour while learning.

We start by providing background on single-agent model-based Bayesian RL (Section 3.1), and a discussion of the multiagent coordination (equilibrium selection) problem (Section 3.2). Then we proceed to present our multiagent extension to Bayesian RL (Sections 3.3 and 3.4), and show experimentally that this can be used effectively to enhance the performance of agents facing the multiagent coordination problem (Section 3.5).

Parts of the research described in this chapter appeared originally in [CB03].

3.1 Single-Agent Model-Based Bayesian RL

Assume an agent is learning to control a stochastic environment modeled as a Markov decision process (MDP) which is a 4-tuple $\langle S, A, p_T, p_R \rangle$ with finite state and action sets S, A, transition dynamics p_T and reward model p_R (as described in 2.1.1). Assuming an infinite horizon and a discount factor $0 \leq \gamma \leq 1$, an agent's objective is to act so as to maximize the expected sum of his future discounted rewards $E[\sum_{t=0}^{\infty} \gamma^t r_t]$, where r_t is the reward at time step t.

In the RL setting, the agent does not have direct access to the model components p_T and p_R, so it must learn a policy based on its interactions with the environment. Any of a number of RL techniques—such as policy or value iteration—can be used to learn an optimal policy and its value $V^*(s)$ at each $s \in S$ [SB98]. However, while striving to learn the optimal policy, agents have to face the *exploration-exploitation tradeoff*: should one exploit what is already known by performing an action that currently appears best, or should one explore in order to gain further information about p_T and p_R and thus re-evaluate its perception of optimality of available actions? If the underlying uncertainty is not properly accounted for, then the agents risk exploring very unrewarding parts of the policy space.

[2] As others [SPG04] have also noted, "blindly" pursuing convergence to equilibria should not necessarily be the goal of MARL.

In model-based RL methods, the learner maintains an estimated MDP $\langle S, A, \widehat{p_T}, \widehat{p_R} \rangle$, based on the set of experiences $\langle s, a, t, r \rangle$ obtained so far. At each stage (or at suitable intervals) this MDP can be solved (exactly or approximately). Single-agent *Bayesian* methods [DFA99, Duf02, SL73] allow agents to incorporate priors to represent their *beliefs* over all the possible MDPs (models) that may be describing the environment, and explore optimally by updating these priors as they gain more knowledge.[3]

We consider the Bayesian framework to be the appropriate framework for dealing with *optimal learning*—acting so as to maximize performance while learning by striking an appropriate balance between short-term and long-term gains [Duf02]. The basic idea is that a Bayesian agent will model the uncertainty about the environment and take it into account when calculating value functions. In theory, once the uncertainty is fully incorporated into the model, acting greedily with respect to these value functions is the optimal policy for the agent, the policy that will enable him optimize his performance while learning. It is well known that *Bayesian exploration is the optimal solution to the exploration-exploitation problem*—meaning that there is no other method that can outperform the Bayesian solution in expectation, while using the same model space and same prior knowledge [Bel61, Mar67].

In practice, approximations to optimal Bayesian exploration have to be used. Even so, it has been demonstrated that Bayesian agents can effectively balance exploration of the environment with exploitation of actions [DFR98, DFA99]. When Bayesian model-based RL is used, the usual advantages of model-based RL apply: by learning a model the agent avoids costly repetition of steps in the environment, and the agent is able to use the model to reason about the effects of its actions so that the number of steps actually executed is reduced. In addition, Bayesian RL can be advantageous in that it does not ignore the agent's uncertainty about the dynamics of the environment, as common model-based approaches which keep point estimates of these dynamics do: by representing a distribution over possible models, the agent's uncertainty can be quantified, which can in turn be used to inform it as to what are the best actions to perform. Finally, despite the fact that Bayesian methods are commonly regarded as being time-consuming, it has been demonstrated that there exists a variety of techniques which allow them to overcome this criticism, for example via the efficient sampling of distributions [DFA99, Duf02, Pri03].

Bayesian methods assume some prior density P over possible dynamics D and reward distributions R, which is updated with each data point $\langle s, a, t, r \rangle$. This prior density describes the

[3]This draws on methods for Bayesian exploration in bandit problems[BF85].

agent's *belief state* regarding the world. Letting H denote the (current) state-action history of the observer, we can use the posterior $P(D, R|H)$ to determine an appropriate action choice at each stage. The formulation of [DFA99] renders this update tractable by assuming a convenient prior. Specifically, the following assumptions are made: (a) the density P is factored over R and D with $P(D, R)$ being the product of independent local densities $P(D^{s,a})$ and $P(R^{s,a})$ for each transition and each reward distribution (as with transition distributions $\Pr(s, a, s')$, reward distributions $\Pr(s, a, r)$ specify the probabilities for achieving each possible reward r, for each s, a state-action pair); and (b) each density $P(D^{s,a})$ and $P(R^{s,a})$ is a Dirichlet [DeG70]. The choice of the Dirichlet is appropriate assuming *discrete multinomials* as the transition and rewards models, for which Dirichlet priors are conjugate—so, the posterior can be represented compactly: after each observed experience tuple, the posterior is also a Dirichlet.

Thus, the posterior $P(D|H)$ can be factored into posteriors over local families, each of the form:

$$P(D^{s,a}|H^{s,a}) = z \Pr(H^{s,a}|D^{s,a})P(D^{s,a})$$

where $H^{s,a}$ is the history of s, a-transitions—captured by updates of the Dirichlet parameters—and z is a normalizing constant.[4] To model $P(D^{s,a})$ a Dirichlet parameter vector $\mathbf{n}^{s,a}$ is used, with entries $n^{s,a,s'}$ for each possible successor state s'.[5] (Similarly, to model $P(R^{s,a})$ a parameter vector $\mathbf{k}^{s,a}$ is used, with entries $k^{s,a,r}$ for each possible reward r.) The expectation of $\Pr(s, a, s')$ with respect to P is given by $n^{s,a,s'} / \sum_i n^{s,a,s_i}$. The updating of a Dirichlet is straightforward: given prior $P(D^{s,a}; \mathbf{n}^{s,a})$ and data vector $\mathbf{c}^{s,a}$ (where $c_i^{s,a}$ is the number of observed transitions from s to s_i under a), the posterior is given by parameter vector $\mathbf{n}^{s,a} + \mathbf{c}^{s,a}$.

To sum up, the Bayesian approach allows the natural incorporation of prior knowledge as a prior probability distribution over all possible MDPs; also, approximations to optimal Bayesian exploration can take advantage of this model, enabling the mass of the posterior to become progressively focused on those MDPs in which the observed experience tuples are most probable [DFA99]. The distribution maintained over possible MDPs, and the Q-values

[4] Similarly, $P(R^{s,a}|H^{s,a}) = z \Pr(H^{s,a}|R^{s,a})P(R^{s,a})$.

[5] The probability density function of the Dirichlet distribution of order L for a random variable X drawn from a multinomial distribution—with L corresponding to the number of values that X can take—is the following function of an L-dimensional vector $\theta = \langle \theta_1, ..., \theta_L \rangle$ with $\theta_i \geq 0$ and $\sum_{i=1}^{L} \theta_i = 1$:

$$f(\theta; \mathbf{n}) \sim \prod_{i=1}^{L} \theta_i^{n_i - 1}$$

where $\mathbf{n} = \langle n_1, ..., n_L \rangle$ is a parameter vector with $n_i \geq 0$ (and where θ_i corresponds to the probability with which the outcome x_i is drawn for X).

of each of these MDPs, induce a distribution over the Q-values at each s, a-pair. The Q-value distribution is then used for action selection. As [DFA99, DFR98] suggest, this can be done by incorporating the actions' *value of perfect information* [RN95] in the agent's exploratory policy.

3.1.1 Value of Perfect Information Exploration

The estimation of a distribution over MDPs at each stage enables one to capture the uncertainty about the model. Knowledge of this uncertainty can be exploited to improve exploration. The decision-theoretic idea of value of perfect information [RN95] is applied by [DFA99, DFR98] in this context; the attempt is to balance the expected gain from exploration (accumulation of more information that leads to improved policies) against the expected cost of doing a potentially suboptimal action. This exploration method is known as *VPI exploration* [DFA99, DFR98]:

Let $\mathit{EVPI}(s, a)$ be the expected value of perfect information about the quality of a state-action pair $q_{s,a}$, where $q_{s,a}$ is a possible value of the optimal Q-value function $Q^*(s, a)$ in one of the possible MDPs. These quantities are variables that depend on the agent's belief state. The $\mathit{EVPI}(s, a)$ coincides with the *expected gain* of performing a at state s given prior beliefs, and thus provides an upper bound on the myopic value of information for exploring a at s. (We present the details for an *EVPI* calculation method in Section 3.1.2.)

There is an *expected cost* incurred for the exploration of an action a. This is given by the difference between the expected value of a, $E[q_{s,a}]$, and the expected value of the current best action (given the agent's current belief state), $E[q^*_{s,a}]$. This can be written as $\mathit{Cost}(s, a) = E[q^*_{s,a}] - E[q_{s,a}]$. The expected values are derived by estimating the Q-value distributions using the maintained transitions and rewards models, as described in Section 3.1.2.

The value of exploration estimate $\mathit{Val}(s, a)$ of a at s can be defined as

$$\mathit{Val}(s, a) = \mathit{EVPI}(s, a) - \mathit{Cost}(s, a)$$

The VPI exploration method proposes that the agent will select the action with the maximum value of exploration estimate. $\mathit{Val}(s, a)$ is therefore used as a way of boosting the desirability of different actions. Choosing the action maximizing $\mathit{Val}(s, a)$ is equivalent to choosing the action that maximizes

$$E[q_{s,a}] + \mathit{EVPI}(s, a)$$

When the agent is confident in the estimated Q-values, the EVPI of each action is close to zero, causing the agent to always choose the action with the highest expected value.

3.1.2 Estimating Q-Value Distributions

One simple way to estimate the Q-value distributions is to use *naive sampling* [DFA99].

The approach consists of sampling k MDPs from the density describing the agent's belief state. Each MDP is solved by using value iteration. For each s and a pair then there exists a sample solution $q^1_{s,a}, ..., q^k_{s,a}$, where $q^i_{s,a}$ is the optimal Q-value given the ith MDP. Given these samples, the mean Q-value is estimated as:

$$E[q_{s,a}] \approx \frac{1}{\sum_i w^i} \sum_i w^i q^i_{s,a} \tag{3.1}$$

Similarly, the *EVPI* can be estimated by summing over the k MDPs:

$$\textit{EVPI}(s,a) \approx \frac{1}{\sum_i w^i} \sum_i w^i gain_{s,a}(q^i_{s,a}) \tag{3.2}$$

In the formulas above, w^i denotes the weight of each sample[6] and $gain_{s,a}(q^i_{s,a})$ denotes the gain from learning the value of $q^i_{s,a}$ (provided by solving that particular MDP). This gain is calculated as described below.

Suppose that, given the agent's current belief state, a_1 is the action with highest expected Q-value at state s and a_2 is the second-highest.

The *gain* associated with learning that the true value of the s, a-pair is in fact q, is defined as:

$$\textit{gain}_{s,a}(q) = \begin{cases} E[q_{s,a_2}] - q, & \text{if } a = a_1 \text{ and } q < E[q_{s,a_2}] \\ q - E[q_{s,a_1}], & \text{if } a \neq a_1 \text{ and } q > E[q_{s,a_1}] \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

Intuitively, the *gain* reflects the effect on decision quality of learning the true Q-value of a specific action at state s. In the first two cases, what is learned causes us to change our decision (in the first case, the estimated optimal action is learned to be worse than predicted, and in the second, some other action is learned to be better than the predicted optimal). In the third case, no change in decision at s is induced, so the information has no impact on the (estimated) decision quality.

[6] Samples may have different weights depending on the sampling method used. All weights can be set to 1, in the simplest case.

Finally, note that since parameter independence is assumed, each of the transition and rewards models can be sampled independently, and thus the sampling problem is reduced to sampling from "simple" posterior distributions. Furthermore, Dearden *et al.* propose various other techniques to make the sampling process more computationally efficient, such as the *importance sampling* technique which allows the reweighting of sampled MDPs (and their re-use, instead of collecting new samples at each step). For a detailed description of those techniques we refer to [DFA99].

3.2 Multiagent Coordination and Equilibrium Selection

Nash equilibria are generally viewed as the standard solution concept for stochastic games. However, it is widely recognized that the equilibrium concept has certain (descriptive and prescriptive) deficiencies. One important problem identified in Chapter 2 is the fact that games may have multiple equilibria, leading to the problem of equilibrium selection.[7] As an example, consider the simple two-player single-state identical interest (coordination) game called the *penalty game* [CB98], shown in Table 3.1 in standard matrix form. Here agent A has moves

	$a0$	$a1$	$a2$
$b0$	10	0	k
$b1$	0	2	0
$b2$	k	0	10

Table 3.1: The *Penalty Game*

$a0, a1, a2$ and B has moves $b0, b1, b2$. The payoffs to both players are identical, and $k < 0$ is some penalty. There are three pure equilibria. While $\langle a0, b0\rangle$ and $\langle a2, b2\rangle$ are the optimal equilibria, the symmetry of the game induces a coordination problem for the agents. With no means of breaking the symmetry, and the risk of incurring the penalty if they choose different optimal equilibria, the agents might in fact focus on the suboptimal equilibrium $\langle a1, b1\rangle$.

The existence of multiple "stage game" equilibria is again a problem that plagues the construction of optimal strategies for multi-state (stochastic) cooperation games. Consider another simple identical interest example, the stochastic *Opt-In or Out* game [Bou99] shown in Figure 3.1. In this game, there are two optimal strategy profiles that maximize sequential value.

[7] The equilibrium selection problem has drawn much attention in game theory [HS88] and is one prime motivation for theories of learning in games [FL98].

In both, the first agent chooses to "opt in" at state s_1 by choosing action a, which takes the agents (with high probability) to state s_2; then at s_2 both agents either choose a or both choose b—either joint strategy gives an optimal equilibrium.

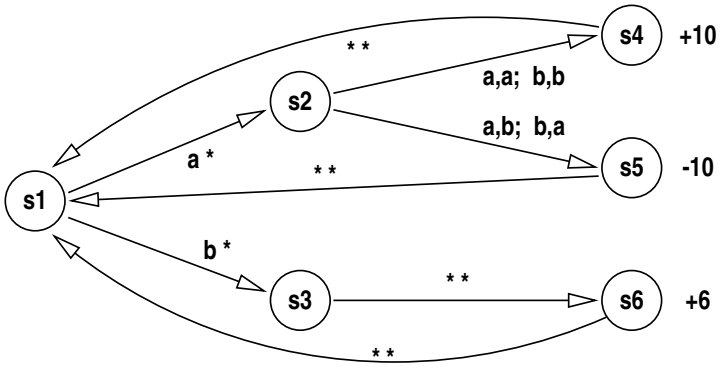


Figure 3.1: The Opt-In or Out stochastic game

Intuitively, the existence of these two equilibria gives rise to a coordination problem at state s_2. If the agents choose their part of the equilibrium randomly, there is a 0.5 chance that they miscoordinate at s_2, thereby obtaining an expected immediate reward of 0. On this basis, one might be tempted to propose methods whereby the agents decide to "opt out" at s_1 (the first agent takes action b) and obtain the safe payoff of 6. However, if we allow some means of coordination—for example, simple learning rules like fictitious play or randomization—the sequential nature of this problem means that the short-term risk of miscoordination at s_2 can be more than compensated for by the eventual stream of high payoffs should they coordinate. Boutilier [Bou99] argues that the solution of games like this, assuming some (generally, history-dependent) mechanism for resolving these stage game coordination problems, requires explicit reasoning about the odds and benefits of coordination, the expected cost of attempting to coordinate, and the alternative courses of action. For example, suppose the two agents use a randomization protocol for coordination, in which they randomly choose a potentially optimal action in the case of equilibrium conflicts at a given state, and—should they ever play a jointly optimal action (equilibrium)—they stick with that action at any subsequent visits to that state. The algorithm for identical interest stochastic games presented in [Bou99] would, in this case, determine that the optimal policy would have the agents opt in (and run the risk of miscoordinating a few times before eventually coordinating) if the discount factor is close enough to 1; but immediately opt out if the discount factor is too low.

As discussed in Chapter 2, Claus and Boutilier [CB98] proposed several MARL meth-

ods for repeated games. A simple joint-action learner (JAL) protocol learned the (myopic, or one-stage) Q-values of joint actions, using standard Q-learning updates. The novelty of this approach lies in its exploration strategy: (a) a fictitious play protocol estimates the strategies of other agents; and (b) exploration is biased by the expected Q-value of actions. Specifically, the estimated value of an action is given by its expected Q-value, where the expectation is taken with respect to the fictitious play beliefs over the other agents's strategies. When semi-greedy exploration is used, prescribing that actions with higher value are more likely to be played, this method will converge to an equilibrium in the underlying stage game.

One drawback of the JAL method is the fact that the equilibrium it converges to depends on the specific path of play. Certain equilibria can exhibit serious resistance—for example, the odds of converging to an optimal equilibrium in the penalty game above are quite small (and decrease dramatically with the magnitude of the penalty). Claus and Boutilier propose several heuristic methods (such as the *Optimistic Boltzmann* and the *Combined Boltzmann* methods [CB98]) that bias exploration toward optimal equilibria: that is, action selection can be biased toward actions that form part of an optimal equilibrium. In the penalty game, for instance, despite the fact that agent B may be predicted to play a strategy that makes the $a0$ look unpromising, the repeated play of the $a0$ by A can be justified by assuming that B *will* play its part of this optimal equilibrium—i.e., by making an *optimistic assumption*. This is further motivated by the fact that repeated play of $a0$ would eventually *draw* B toward this equilibrium.

This issue of learning optimal equilibria in identical interest games has been addressed recently in much greater detail. Lauer and Riedmiller [LR00] describe a Q-learning method for identical interest stochastic games that explicitly embodies this optimistic assumption in its Q-value estimates. Specifically, their update rule for playing an action a at state s defines that its Q-value estimate $Q_t(s, a)$ at time step t is only updated if the new value estimate $Q_{t+1}(s, a)$ is greater than $Q_t(s, a)$. By not allowing any lowering of the Q-values, [LR00] purposefully "neglect" to incorporate any negative effects of an agent's individual action into its corresponding Q-value, and thus the Q-values eventually converge to the maximum reward corresponding to their respective actions—and the agents' play into that corresponding to selectin of the optimal joint action. However, [LR00]'s approach cannot guarantee convergence to the optimal equilibrium in games with stochastic rewards.

Wang and Sandholm [WS02] *do* assume stochastic rewards when dealing with the problem of coordination in identical interest stochastic games environments. Their approach, labeled *Optimal Adaptive Learning*, also uses the optimistic assumption to guarantee convergence to

an optimal equilibrium in this more general class of games —even if the agents do not know the game's stochastic state transition model.

Kapetanakis and Kudenko [KK02] propose a method called *Frequency Maximum Q-value (FMQ)* for repeated games that uses the optimistic assumption to bias exploration, much like [CB98], but in the context of individual learners. The FMQ heuristic updates Q-values by taking into account the frequency with which an individual action produces its maximum corresponding reward encountered so far, and thus leads the agents' play towards optimal equilibria (since for an agent to achieve the maximum reward corresponding to one of its actions, the other agent must be playing the game accordingly). FMQ is a non-generalizable heuristic approach. Though it assures convergence to an optimal equilibrium in coordination games with deterministic rewards, or in games which have a specific structure that allows for the easy differentiation among equilibria based on their associated rewards, even if these rewards are stochastic (such as the penalty game), it cannot in general guarantee convergence in coordination games with stochastic rewards.

The pursuit to devise methods that ensure eventual convergence to optimal equilibria of repeated cooperation games is in some circumstances well justified. However, these methods do not account for the fact that—by *forcing* agents to undertake actions that have potentially drastic effects in order to reach an optimal equilibrium—they can have a dramatic impact on accumulated reward. The penalty game was devised to show that these highly penalized states can bias (supposedly rational) agents away from certain equilibria; yet optimistic exploration methods ignore this and blindly pursue these equilibria at all costs. Under certain performance metrics (e.g., average reward over an infinite horizon) one might justify these techniques.[8] However, when using the discounted reward criterion, the tradeoff between long-term benefit and short-term cost should be addressed.

As mentioned above, this tradeoff was discussed by [Bou99] in the context of known-model stochastic games. When coordination on a "good" strategy profile requires exploration in parts of policy space that are very unrewarding, the benefits of eventual coordination to an optimal equilibrium ought to be weighed against the cost (in terms of reward sacrificed while learning to play that equilibrium). In this chapter, we show that we can address the same tradeoff in the actual RL context—*assuming unknown (stochastic) reward and state transition models*—by formulating a Bayesian approach to model-based MARL. By maintaining probabilistic beliefs over the space of models and the space of opponent strategies, our learning agents can ex-

[8] Even then, more refined measures such as *bias optimality* [Put94] might cast these techniques in a less favourable light.

plicitly account for the effects their actions can have on (a) their knowledge of the underlying model; (b) their knowledge of the other agent strategies; (c) expected immediate reward; and (d) expected future behaviour of other agents. Components (a) and (c) are classical parts of the single-agent Bayesian RL model [DFA99]. Components (b) and (d) are key to the multiagent extension, allowing an agent to explicitly reason about the potential costs and benefits of coordination.

3.3 A Bayesian Model for Multiagent Reinforcement Learning

Here we develop a model that accounts for the generalized exploration-exploitation tradeoff in MARL. We adopt a Bayesian, model-based approach to MARL, much like the single-agent model described in [DFA99]. The value of information will play a key role in determining an agent's exploration policy. Specifically, the value of an action consists of two components: its estimated value given current model estimates, and the expected decision-theoretic *value of information* it provides (i.e., the ability this information has to change future decisions). We augment both parts of this value calculation in the MARL context. The estimated value of an action given current model estimates requires predicting how the action will influence the future action choices of other agents. The value of information associated with an action includes the information it provides about other agents's strategies, not just the environment model. Both of these changes require that an agent possess some model of the strategies of other agents, for which we adopt a Bayesian view [KL93]. Putting these together, we derive optimal exploration methods for (Bayesian) multiagent systems.

We assume a stochastic game G in which each agent knows the game structure, but not the reward or transition models. A learning agent is able to observe the actions taken by all agents, the resulting game state, and the reward received by himself (but *not* the rewards of others). Thus an agent's experience at each point in time is simply $\langle s, a, t, r \rangle$, where s is a state in which joint action a was taken, t is the resulting state, and r is the reward received by the agent.

A *Bayesian MARL agent* has some prior distribution over the space of possible models as well as the space of possible strategies being employed by other agents. These beliefs are updated as the agent acts and observes the results of its actions and the action choices of other agents. The strategies of other agents may be history-dependent, and we allow our *Bayesian agent (BA)* to assign positive support to such strategies. As such, in order to make accurate

predictions about the actions others will take, the BA must monitor appropriate observable history. In general, the history (or summary thereof) required will be a function of the strategies to which the BA assigns positive support. We assume that the BA keeps track of sufficient history to make such predictions. For example, should the BA believe that its opponent's strategy lies in the space of finite state controllers that depend on the last two joint actions played, the BA will need to keep track of these last two actions. If it uses fictitious play beliefs (which can be viewed as Dirichlet priors) over strategies, no history need be maintained.

The belief state of the BA has the form $b = \langle P_M, P_S, s, h \rangle$, where: P_M is some density over the space of possible models (i.e., games); P_S is a joint density over the possible strategies played by other agents; s is the current state of the system; and h is a summary of the relevant aspects of game history, sufficient to predict the action of any agent given any strategy consistent with P_S. Given experience $\langle s, a, t, r \rangle$, the BA updates its belief state using standard Bayesian methods. The updated belief state is:

$$b' = b(\langle s, a, t, r \rangle) = \langle P'_M, P'_S, t, h' \rangle \tag{3.4}$$

Updates are given by Bayes rule:

$$P'_M(m) = z \Pr(t, r | a, m) P_M(m)$$

and

$$P'_S(\sigma_{-i}) = z \Pr(a_{-i} | s, h, \sigma_{-i}) P_S(\sigma_{-i})$$

Here h' is a suitable update of the observed history (as described above). This model combines aspects of Bayesian reinforcement learning [DFA99] and Bayesian strategy modeling [KL93].

To make belief state maintenance tractable (and admit computationally viable methods for action selection below), we assume a specific form for these beliefs [DFA99]. First, our prior over models will be factored into independent local models for both rewards and transitions. We assume independent priors P_R^s over reward distributions (regarding each agent's *own* rewards) at each state s, and $P_D^{s,a}$ over system dynamics for each state and joint action pair. These local densities are Dirichlet and thus they can be represented using a small number of hyperparameters, and can be easily updated. For example, our BA's prior beliefs about the transition probabilities for joint action a at state s will be represented by a vector $\mathbf{n}^{s,a}$ with one parameter per successor state t. Expected transition probabilities and updates of these beliefs are as described in Section 3.1.

Second, we assume that the beliefs about opponent strategies can be factored and represented in some convenient form. For example, it would be natural to assume that the strategies of other agents are independent. Simple fictitious play models could be used to model the BA's beliefs about opponent strategies (corresponding to Dirichlet priors over mixed strategies), allowing ready update and computation of expectations, and obviating the need to store history in the belief state. Similarly, distributions over specific classes of finite state controllers could also be used. We will not pursue further development of such models,[9] since we use only simple fictitious play opponent models in our experiments below: Each agent i keeps a count $C^j_{a_j}$ for each opponent j and $a_j \in S_j$, with S_j being the opponent's strategy space at each state s of the stochastic game, of the number of times agent j has used the *individual* action a^j (at that state) in the past. For each opponent j, i assumes j plays a_j with probability $Pr^i_{a_j} = \frac{C^j_{a_j}}{\sum_{\tilde{a}_j \in S_j} C^j_{\tilde{a}_j}}$.

We provide a somewhat different perspective on Bayesian exploration than that described in [DFA99]. The value of performing an action a_i at a belief state b can be viewed as involving 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. Since each action gives rise to some "response" by the environment that changes the agent's beliefs, and these changes in belief can influence subsequent action choice and expected reward, we wish to quantify the value of that information by determining its impact on subsequent decisions.

EVOI need not be computed directly, but can be combined with "object-level" expected value through the following Bellman equations over the belief state MDP:

$$Q(a_i, b) = \sum_{a_{-i}} \Pr(a_{-i}|b) \sum_t \Pr(t|a_i \circ a_{-i}, b) \sum_r \Pr(r|a_i \circ a_{-i}, b)[r + \gamma V(b(\langle s, a, t, r\rangle))] \quad (3.5)$$

$$V(b) = \max_{a_i} Q(a_i, b) \quad (3.6)$$

These equations describe the solution to the POMDP that represents the exploration-exploitation problem, by conversion to a belief state MDP. These can (in principle) be solved using any method for solving high-dimensional continuous MDPs—of course, in practice, a number of

[9] However, we note that the development of tractable classes of (realistic) opponent models remains an interesting problem.

computational shortcuts and approximations will be required (as we detail below, in 3.4). We complete the specification with the straightforward definition of the following terms:

$$\Pr(a_{-i}|b) = \int_{\sigma_{-i}} \Pr(a_{-i}|\sigma_{-i}) P_S(\sigma_{-i}) \tag{3.7}$$

$$\Pr(t|a,b) = \int_{m} \Pr(t|s,a,m) P_M(m) \tag{3.8}$$

$$\Pr(r|b) = \int_{m} \Pr(r|s,m) P_M(m) \tag{3.9}$$

This formulation determines the optimal policy as a function of the BA's belief state. This policy incorporates the tradeoffs between exploration and exploitation, both with respect to the underlying (dynamics and reward) model, and with respect to the behaviour of other agents. As with Bayesian RL and Bayesian learning in games, no *explicit* exploration actions are required: acting greedily on the Bayesian Q-values guarantees optimal behaviour in terms of *expected* discounted accumulated reward.

Of course, it is important to realize that this model may *not* converge to an optimal policy for the true underlying stochastic game. Priors that fail to reflect the true model, or unfortunate reward samples early on, can easily mislead an agent, and direct him away from exploring sufficiently.

All the same, as we have seen, there are approaches that do ensure convergence to equilibria in the case of identical interest repeated games.[10] However, the cost in terms of discounted accumulated reward can be too high for such methods, "rushing" as they do towards equilibria without considering the dangers in their path.[11] It is precisely the "well-reasoned" behaviour exhibited by the Bayesian approach that allows an agent to learn how to behave well, avoiding drastic penalties when operating in environments that entail such dangers. The Bayesian agents willingly take the risk of converging to suboptimal policies, through due consideration of the learning process given their current beliefs about the domain. Even so, they often manage to find optimal strategies, as we shall see in Section 3.5. There, we also demonstrate that

[10] When learning is taking place, there are two aspects to convergence: convergence to the optimal solution and convergence to any policy at all. The approaches in question deal with the first aspect, considering as "optimal solution" the (eventual) play of an optimal equilibrium. However, as our experiments will demonstrate, this can be in sharp contrast with *optimal learning* as presented earlier in this thesis.

[11] Note also that most of those methods were designed for known-model or low stochasticity identical interest repeated games. As is admitted in the related literature (e.g., [KK02, LR00]), agents using those methods might fail to coordinate when operating in stochastic RL environments, or take an agonizingly long time before doing so—exacerbating their unrewarding online behaviour problem.

the Bayesian MARL algorithms we developed enhance the online performance of agents in coordination problems.

3.4 Computational Approximations

Solving the belief-state MDP defined in the previous section will generally be computationally infeasible. Therefore, we now propose two methods that can be used to approximate the optimal solution to the multiagent exploration-exploitation problem described by the belief-state MDP above.

3.4.1 Myopic *EVOI*

In specific MARL problems, the generality of a solution such as the one described by Equations 3.5 and 3.6—defining as it does a value for every possible belief state—is not needed anyway. Most belief states are not reachable given a specific initial belief state. A more directed search-based method can be used to solve this MDP for the agent's current belief state b. Here we present a form of *myopic EVOI* in which only immediate successor belief states are considered, and their values are estimated without using VOI or lookahead.

Formally, myopic action selection is defined as follows. Given belief state b, the *myopic* Q-function for each $a_i \in A_i$ is:

$$Q_m(a_i, b) = \sum_{a_{-i}} \Pr(a_{-i}|b) \sum_t \Pr(t|a_i \circ a_{-i}, b) \sum_r \Pr(r|a_i \circ a_{-i}, b)[r + \gamma V_m(b(\langle s, a, r, t\rangle))] \tag{3.10}$$

$$V_m(b) = \max_{a_i} \int_m \int_{\sigma_{-i}} Q(a_i, s|m, \sigma_{-i}) P_M(m) P_S(\sigma_{-i}) \tag{3.11}$$

The action performed is that with maximum myopic Q-value. Eq. 3.10 differs from Eq. 3.5 in the use of the myopic value function V_m, which is defined as the expected value of the optimal action at the current state, assuming a fixed distribution over models and strategies. Intuitively, this myopic approximation performs one step-lookahead in belief space, then evaluates these successor states by determining the expected value to BA w.r.t. a fixed distribution over models, and a fixed distribution over successor states. Henceforth, we will therefore be referring to this method as the *Bayesian One-Step Lookahead (BOL)* method.

The computation in Eq. 3.10 involves the evaluation of a finite number of successor belief states—$A \cdot R \cdot S$ such states, where A is the number of joint actions, R is the number of rewards, and S is the size of the state space (unless b restricts the number of reachable states, plausible strategies, etc.). Greater accuracy can be realized by BOL variants employing *multistage* lookahead—with the requisite increase in computational cost. Conversely, the myopic action can be approximated by sampling successor beliefs (using the induced distributions defined in Equations 3.7, 3.8 and 3.9) if the branching factor $A \cdot R \cdot S$ is problematic.

The final bottleneck in this approach involves the evaluation of the myopic value function $V_m(b)$ over successor belief states. The $Q(a_i, s|m, \sigma_{-i})$ terms are Q-values for standard MDPs, and can be evaluated using standard methods, but direct evaluation of the integral over all models is generally impossible. However, *sampling techniques* can be used to render the (approximate) evaluation of this integral possible [DFA99]. Specifically, some number of models can be sampled, the MDPs corresponding to the samples can be solved (using methods such as those presented in Section 2.1.1), and the expected Q-values estimated by averaging over the sampled results. One can thus use the following "algorithm" to evaluate the myopic Q-value of each individual action at belief state b:

(a) At belief state b, characterized by priors P_M and P_S, for each potential experience tuple $\langle s, a, r, t \rangle$ perform a one-step lookahead in belief space, resulting in successor belief state b' with updated P'_M, P'_S.[12]

(b) Sample a finite set of k models from each successor P'_M.

- Solve each one of the k sampled MDPs (using a standard method such as value iteration), with respect respect to the density P'_S over strategies. This results in state-action values $Q(a_i, s|m, \sigma_{-i})$ for each $a_i \in A_i$ in that (say the mth) sample MDP.
- Estimate the value $V_m(b')$ for this successor belief state as the value of the action with the maximum *average* $\overline{Q}(a_i, s|m, \sigma_{-i})$ value over all samples (this could possibly be a weighted average, as described in Eq. 3.1).

(c) The estimated $V_m(b')$ values are subsequently used in the evaluation of Eq. 3.10.

Various techniques for making this process more efficient can be used as well, including *importance sampling* (allowing results from one MDP to be used multiple times by reweighting)

[12] Alternatively, as already mentioned, we may just examine some samples of successor belief states, if $A \cdot R \cdot S$ is problematic.

and *repair* of the solution for one MDP when solving a related MDP [DFA99]. (We will not provide more details on these techniques, since we did not employ them in the experiments described in this chapter.)

For certain classes of problems, this evaluation can be performed directly. For instance, suppose a *repeated*, that is, *single-state (ss)* game (with stochastic rewards) is being learned, and the BA's strategy model consists of fictitious play beliefs. The immediate expected reward of any action a_i taken by the BA (with respect to successor b' derived by the observation of the reward generated by the joint action) is given by its expectation w.r.t. its estimated reward distribution and fictitious play beliefs. The maximizing action a_i^* with highest expected immediate reward at b',

$$a_i^* = arg\max_{a_i} \sum_{a_{-i}} \Pr(a_{-i}|b') \sum_r \Pr(r|a_i \circ a_{-i}, b')[r]$$

will (presumably) be the best action at *all* subsequent stages of the repeated game—and its expected reward $r(a_i^*)$ under the myopic (value) assumption that beliefs are fixed by b'. Thus, the long-term value at b' is

$$V_m^{ss}(b') = r(a_i^*)/(1-\gamma) \tag{3.12}$$

Then, expected reward for each action can readily be combined with the BA's fictitious play beliefs to compute the expected value of an action (over the infinite horizon) since the only model uncertainty is in the reward:

$$Q_m^{ss}(a_i, b) = \sum_{a_{-i}} \Pr(a_{-i}|b) \sum_r \Pr(r|a_i \circ a_{-i}, b)[r + \gamma V_m^{ss}(b')] \tag{3.13}$$

3.4.2 A Multiagent VPI Algorithm

The approaches above are motivated by approximating the direct myopic solution to the exploration POMDP. As explained in Sections 3.1.1 and 3.1.2, a different approach to this approximation was proposed in [DFA99], which estimates the myopic value of obtaining *perfect information* about the quality of a state-action pair. We adapt this approach in our setting.

Given an agent's belief state b, let the expected value of (any of) his $a_i \in A_i$ action be denoted by $\overline{Q}(s, a_i)$. Adapted to our multiagent setting to accomodate reasoning about others' strategies, a computational approximation to this *VPI exploration* employing *naive sampling* approach involves the following steps:

(a) A finite set of k models is sampled from the density P_M.

(b) Each sampled MDP j is solved (using, say, value iteration) with respect to the density P_S over strategies, giving optimal Q-values $q^j(s, a_i)$ for each $a_i \in A_i$ in that MDP.

(c) Then, the average Q-value $\overline{Q}(s, a_i)$ over all k MDPs is calculated (as in Eq. 3.1).

(d) For each a_i, we compute $\textit{gain}_{s,a_i}(q^j(s, a_i))$ for each of the k MDPs (as in Eq. 3.3).

(e) We define $\textit{EVPI}(s, a_i)$ to be the average of these k values (i.e., calculated as in Eq. 3.2).

(f) We define the value of a_i to be $\overline{Q}(s, a_i) + \textit{EVPI}(s, a_i)$ and execute the action with highest value.

Naive sampling can be more computationally effective than one-step lookahead (which requires sampling and solving MDPs from *multiple* belief states). The price paid is approximation inherent in the perfect information assumption: the execution of joint action a does not come close to providing perfect information about $Q(s, a_i)$. Henceforth, we will refer to this multiagent *Bayesian VPI* algorithm as *BVPI*.

3.5 Experimental Evaluation

We conducted a number of experiments with both repeated (single-state) and stochastic (multi-state) games to evaluate the Bayesian approach. We focus on two-player cooperative (coordination) games, largely to compare to existing methods for "encouraging" convergence to optimal equilibria. The Bayesian methods examined are the one-step lookahead (BOL) and naive sampling for estimating VPI (BVPI) algorithms described in Section 3.4. We want to evaluate the online, sequential behaviour of agents while learning, wishing to demonstrate that Bayesian agents address the tradeoff between accumulating short-term and long-term rewards effectively. Thus, the main metric we use in our experiments was the (average) *discounted accumulated reward* (over multiple experimental runs), as this metric provides a suitable way to measure both the cost being paid in the attempt to coordinate as well as the benefits of coordination (or lack thereof). Nevertheless, we also report on (average) undiscounted accumulated reward and convergence to optimal equilibria, as appropriate. In all cases, convergence to a policy was assumed if the policy remained unchanged for the last 5% of iterations in an experimental run.

In all experiments, the Bayesian agents use a simple fictitious play model to represent their uncertainty over the other agent's strategy. Thus at each iteration, a BA believes its opponent to play an action with the empirical probability observed in the past (for multistate games, these beliefs are independent at each state). BOL is used only for repeated games, since these allow the immediate computation of expected values over the infinite horizon (Eqs. 3.12 and 3.13). BVPI is used both in single-state and multi-state games, and in all cases *five* game models are sampled to estimate the VPI of the agents' actions.[13]

We compare our Bayesian methods against three methods that have guarantees for convergence to equilibria in repeated (single-state) games: against the *FMQ* algorithm [KK02] (which, as was explained in Section 3.2 above, biases exploration towards optimal equilibria, and is tailored-made to tackle exactly these types of problems); and against model-based variants of the *Combined Boltzmann Exploration (CBE)* and *Optimistic Boltzmann Exploration (OBE)* methods for JALs, described in [CB98]. These variants of JALs update Dirichlet models of rewards (in addition to having fictitious play beliefs regarding opponents' strategies), and use these models to update their Q-value tables appropriately. OBE agents take the optimistic view that others will act to match their choice of individual action (and thus play an action that corresponds to the joint action with greatest expected value), while CBE agents take into account the "potential" that an individual action has, by assessing the probability that this action will be matched by a partner's action to form a rewarding joint action [CB98]. Furthermore, for both single-state and multi-state games, we compare our approach against the *Win or Learn Fast-Policy Hill Climbing (WoLF-PHC)* algorithm presented in [BV01b]. As was mentioned in Chapter 2, *WoLF-PHC* is a more generic, model-free learning algorithm, which works with arbitrary, general-sum stochastic games, and has no special heuristics for equilibrium selection (nor any theoretical guarantees for convergence to equilibria). In all experiments, the various parameters of the aforementioned algorithms were empirically tuned to give good performance.[14] Finally, the two learning agents in any experiment are in each case of the same type (i.e., the settings are *homogeneous*, with the participating agents using the same algorithm).

[13] Assessing sample complexity would be worthwhile, but we don't explore such issues in this dissertation.

[14] Specifically, CBE and OBE used a temperature parameter T that was initially equal to 10000 and was decayed with a rate equal to 0.9, $\lambda = 1$, while $\rho = 0.5$ in the CBE case [CB98]. The parameters used for FMQ [KK02] were c = 10, $s = 0.006$, $T_{max} = 500$, $\lambda = 0.9$. For WoLF-PHC [BV01b], we used $\epsilon = 0.2$ (with a decay rate varying in $[0.99, 0.99998]$ depending on the setting), $\delta_w = 0.016$, $\delta_l = 0.032$ and $\alpha = 1$ with a 0.9991 decay rate.

3.5.1 Single-State (Repeated) Games

We start by describing our experiments with single-state repeated games. The strategy priors for the BAs are provided by uninformative, uniform "prior counts" for the fictitious play models. The agents' priors regarding the stochastic reward model are similarly uninformative, assigning uniform nonzero (expected) probability to every possible reward for each joint action, regardless of whether this reward is truly feasible or not (except for two experiments, as noted).

The Climbing Game

We first pitted our Baysian algorithms against FMQ, WoLF-PHC, OBE and CBE in the context of the *Climbing Game* [CB98].[15] The Climbing Game is depicted in Table 3.2 (where the *means* of the stochastic rewards are depicted for each pair of actions—each joint action gives rise to a number of distinct rewards). As in the case of the Penalty Game (Table 3.1), k is a negative penalty, and is used to "scare away" the agents from playing the optimal equilibrium $\langle a0, b0\rangle$. The game takes its name from the fact that rational agents tracking each others' moves (and adjusting their policies to the opponent's behaviour) will engage in a "climbing" process that will see them play actions $\langle a2, b2\rangle$, $\langle a2, b1\rangle$ and $\langle a1, b1\rangle$ in their attempt to "reach" $\langle a0, b0\rangle$ [CB98]. As Claus and Boutilier [CB98] argue, if the penalties are too high, chances are that the agents will never reach the optimal equilibrium, ending up playing the *suboptimal equilibrium* $\langle a0, b0\rangle$—and justifiably so.

	$a0$	$a1$	$a2$
$b0$	20	k	0
$b1$	k	16	10
$b2$	0	0	4

Table 3.2: The *Climbing Game*

The experiments we conducted in this domain were composed of 30 experimental runs, with 2000 iterations/run. In the first of our experiments, the penalty k was set to have an expected value of -20, and we set the discount factor γ to 0.95. Results in terms of (average) discounted accumulated reward are shown in Figure 3.2.[16] The Bayesian methods visibly out-

[15] Even though the Climbing Game and the Penalty Game are toy games, they provide strong intuitions regading the dynamics of coordination, and, for that reason, they have provided the main experimental domains for several papers in the past—mainly for non-stochastic rewards (such as [KK02, LR00]).

[16] When discounting is involved, we depict only the initial "interesting" segments of our graphs.

perform the rest (BOL ranks first and BVPI second). It is interesting that WoLF ranks third, even though the other (non-Bayesian) methods were tailored-made to tackle exactly this type of repeated coordination games. In terms of convergence behaviour, all methods converge to the optimal equilibrium quite often, with the exception of WoLF-PHC that mainly converges to the suboptimal equilibrium (Table 3.3).

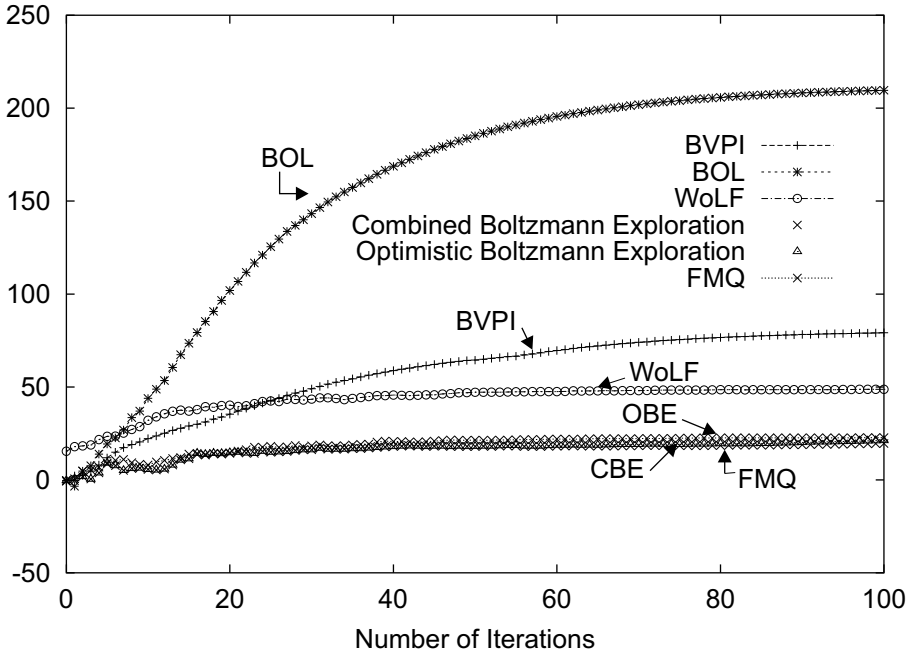


Figure 3.2: Climbing Game Results, $\gamma = 0.95$, $k = -20$; y axis is discounted accumulated reward (averaged over 30 runs).

	BOL	*BVPI*	*CBE*	*OBE*	*FMQ*	*WoLF-PHC*
OE	22	16	16	20	21	1
SE	8	2	14	10	8	29
NE	0	12	0	0	1	0

Table 3.3: Climbing game, $\gamma = 0.95$, $k = -20$: Number of runs converging to optimal equilibrium (OE), suboptimal equilibrium (SE) or non-equilibrium (out of 30 runs). Convergence to NE usually simply means the agents have not converged to playing some specific policy.

For interest, we repeated the experiment in this setting with a discount factor $\gamma = 0.999$, expecting that a discount factor close to 1 would diminish the advantage of the Bayesian methods and make the "convergent" methods (i.e., CBE, OBE and FMQ) look better in terms of discounted accumulated reward. The results were as shown in Fig. 3.3.

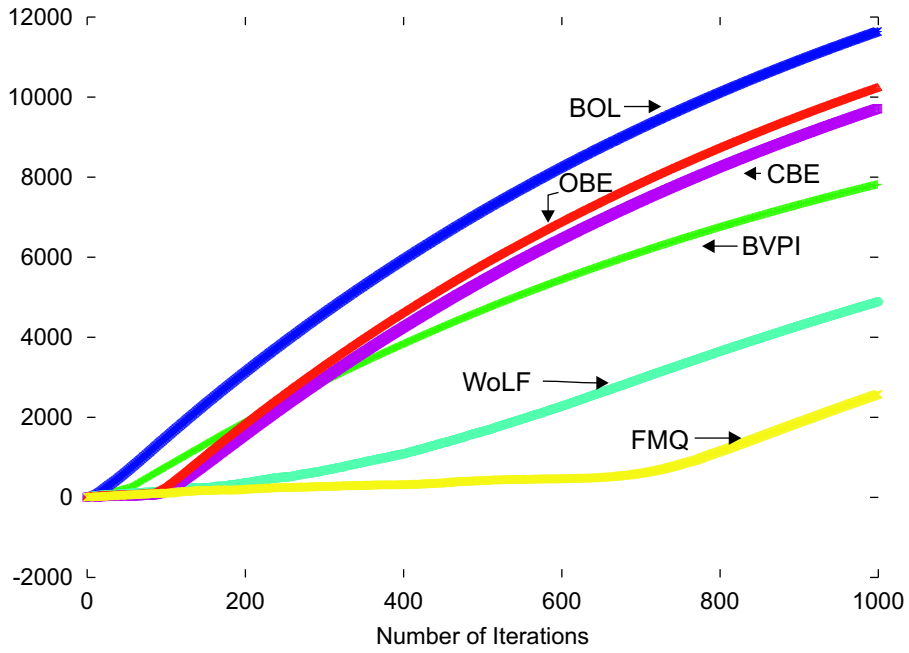


Figure 3.3: Climbing Game Results, $\gamma = 0.999$, $k = -20$; y axis is discounted accumulated reward (averaged over 30 runs).

Those results indeed present the convergent methods under a more favourable light. Nevertheless, BOL still ranked first in terms of discounted accumulated reward in this experiment (as it managed to converge to the optimal equilibrium $\langle a0, b0\rangle$ in 23/30 runs, and to the suboptimal equilibrium $\langle a1, b1\rangle$ in 7/30 runs). BVPI, however, even though it outperforms the OBE and CBE convergent methods in the initial stages of the experiment, it eventually ranks under them (it did converge to the optimal equilibrium in 15/30 runs, but did not converge at all in 10 runs). In contrast, OBE's and CBE's performance improves dramatically once they start converging to the optimal (20/30 times for OBE, 16/30 times for CBE; in the rest of the runs they converge to the suboptimal equilibrium). WoLF-PHC converged to the suboptimal equilibrium in all runs; even so, we can see in the graph that once WoLF agents stop exploring and settle for the suboptimal policy, their performance improves significantly. WoLF outperforms the FMQ method again, even though FMQ does converge to the optimal equilibrium in 21 of the runs and to the suboptimal in 9 of them. The problem for FMQ is that, due to the setting's stochasticity, it takes the agents quite some time before they converge to equilibrium. We note that the ranking of the methods remains the same after 2000 iterations: we show only 1000 iterations in the graph in order to clearly demonstrate the penalties suffered by the convergent methods in the initial phases of the game.

We conducted a third experiment in the Climbing Game setting, increasing the penalty k to -100 and also increasing the reward stochasticity (by increasing the variance of the reward received for each action). The discount factor was set to 0.95. To test one of the benefits of model-based RL, we provided informative priors for BOL, BVPI, CBE and OBE agents, giving the agents strong information about rewards by restricting the prior to assign (uniform) nonzero probability only to the small range of truly feasible rewards for each action.

Specifically, we wanted to test the validity of the hypothesis that the model-based approaches could benefit from the informative priors—in the sense that informative priors would help them overcome (to an extent, at least) the increased risks arising from increasing the penalty's value and the stochasticity of the domain. Nevertheless, we can see in Figure 3.4 that, in terms of discounted accumulated reward, this hypothesis was valid only for the Bayesian methods, which again top all others. Contrary to BVPI and BOL, and even though they employ informed priors, CBE and OBE again fall prey to the (increased) penalties, faring similarly to FMQ and being outperformed by WoLF.

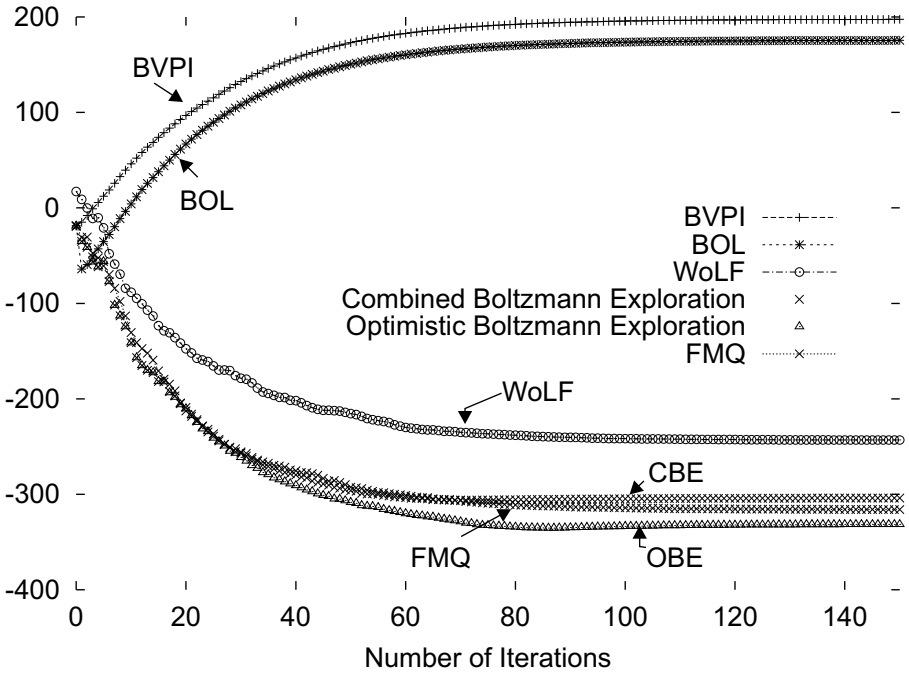


Figure 3.4: Climbing Game Results, $\gamma = 0.95$, $k = -100$; y axis is discounted accumulated reward (averaged over 30 runs).

Notice that, unlike the $k = -20$ case, BVPI now ranks first and BOL second. We attribute this (at least partly) to the fact that BVPI provides a more cautious approach than BOL in single-state games: in such games, BOL incorporates (rather optimistically) the myopic as-

sumption that the action with the highest expected reward in the immediate successor belief state will also be the best in all subsequent stages. However, this assumption is more likely to be flawed under increased stochasticity.

	BOL	*BVPI*	*CBE*	*OBE*	*FMQ*	*WoLF-PHC*
OE	8	0	7	30	15	2
SE	22	30	23	0	12	28
NE	0	0	0	0	3	0

Table 3.4: Climbing game, $\gamma = 0.95$, $k = -100$: Number of runs converging to optimal equilibrium (OE), suboptimal equilibrium (SE) or non-equilibrium (out of 30 runs). Convergence to NE usually simply means the agents have not converged to playing some specific policy.

In terms of convergence behaviour, most of the methods find it hard to converge to the optimal equilibrium, as shown in Table 3.4. However, the optimistic OBE agents do manage to converge to the optimal equilibrium in all runs in this scenario. This is because their informed priors allowed them to become confident in the high value of $\langle a0, b0 \rangle$, and both agents play under the—valid, in this case—assumption that the partner will match their choice of action. FMQ and WoLF-PHC, being model-free approaches, cannot benefit from informed priors. We can see however that WoLF's behaviour is similar to its behaviour with the lower penalty, in terms of convergence, while FMQ's convergence behaviour has visibly deteriorated (due to the increased penalty and the increased stochasticity of the domain). Still, it managed to converge to the optimal equilibrium 15 times—compared to 0 for BVPI and 8 for BOL. We should note here that BVPI agents top all others in terms of discounted accumulated reward, even though they never converge to the optimal equilibrium.

To conclude, it is noteworthy that, in both cases, WoLF-PHC outperforms FMQ, CBE, and OBE in terms of discounted accumulated reward—even though the later methods were all designed to tackle this specific type of problems. It is also noteworthy that the Bayesian methods managed to converge to the optimal equilibrium often, even though they come with no such convergence guarantees. Notably, the Bayesian methods strongly outperform the convergent methods in terms of accumulated discounted reward. The convergence of others comes at too high a price. This is confirmed by the results in the rest of our experiments.

The Penalty Game

We also conducted experiments within the Penalty Game setting (described in Section 3.2, with the game matrix given in Table 3.1). Again, the game is altered so that joint actions provide a

stochastic reward, whose mean is the value shown in the game matrix.

The first set of experiments in this setting uses a penalty k set to -20, discount factors of 0.75 and 0.95, and uninformative priors (as explained earlier). Results appear in Figures 3.5(a) and (b) showing the total discounted reward accumulated by the learning agents, averaged over 30 trials (with each trial composed of 5000 iterations).

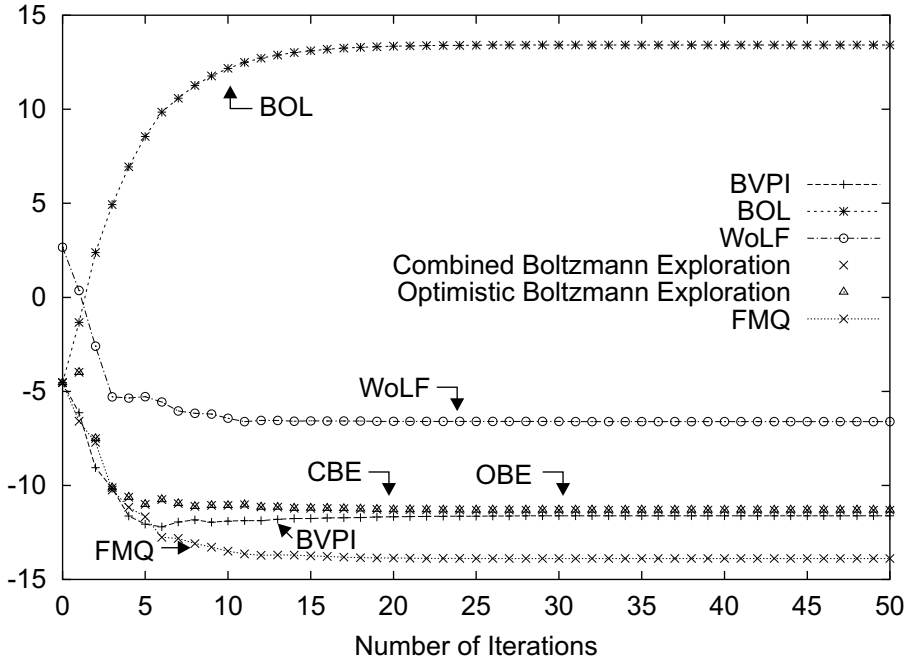
The results show that both Bayesian methods perform significantly better than the methods designed to force convergence to an optimal equilibrium. Indeed, OBE, CBE and FMQ converge to an optimal equilibrium (e.g., playing one of the $\langle a0, b0\rangle$ or $\langle a2, b2\rangle$ joint actions) in virtually all of their 30 runs (as shown in Tables 3.5 and 3.6), but clearly pay a high price.

WoLF-PHC does better than OBE, CBE and FMQ in both tests. In fact, this method outperforms the BVPI agent in the case of $\gamma = 0.75$. In that case, WoLF agents always converge to the optimal equilibria early on, while BVPI converges to a nonequilibrium—or fails to converge to a specific policy—5 times and to the suboptimal equilibrium (playing the $\langle a1, b1\rangle$ joint action) 4 times, as shown in Table 3.5. Nevertheless, the picture presented in Figure 3.5(a) regarding the performance of BVPI is a bit deceiving: in reality, the (average) behaviour of the VPI agents improved considerably after the first 20 iterations, but this cannot be shown clearly in Figure 3.5(a) due to the high discounting (0.75) used.[17]

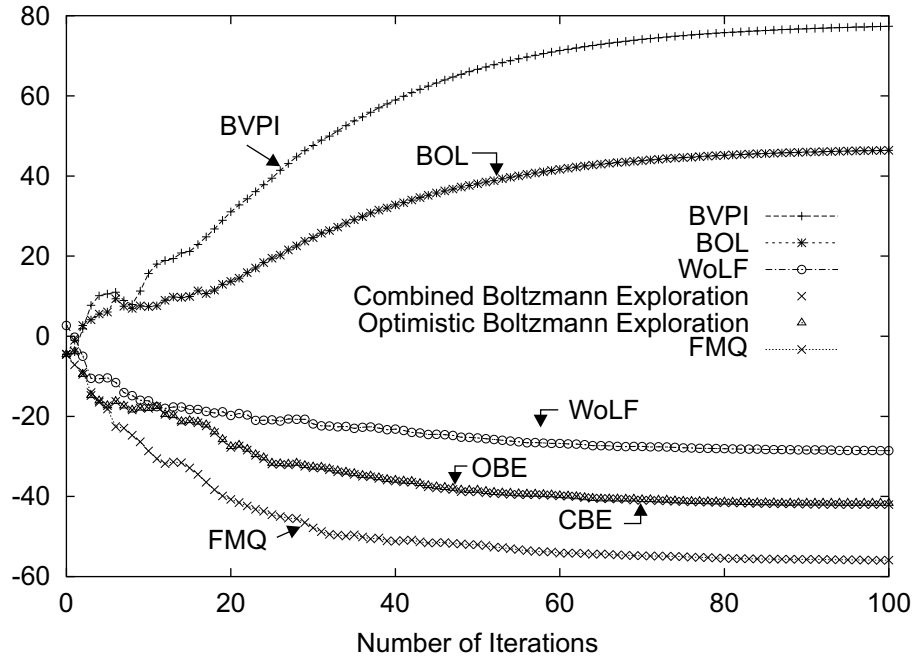
The exact picture in this game was that in some (5) of the runs, the BVPI agents suffered heavy penalization early on, and thus required more time to recover due to the small sampling size; in 3 of those runs however the agents eventually managed to converge to the suboptimal equilibrium—while in the other 2 they converged to playing a non-equilibrium strategy. Thus, it is not the lack of convergence to optimal equiliria that has a dramatic impact on the agents performance (this is verified by the rest of our experiments as well), but in the cases of "unlukcy samples" early on, the reward early in the trial is significantly less even if total reward over the entire trial is not. With heavy discounting, underperformance for even small initial periods cannot be overcome by good performance (with respect to total undiscounted reward) overall. Notice that $\gamma = 0.75$ means a reward received five steps from now is worth less than 25% of the same reward received now: this is heavy discounting, almost unrealistic for most environments.

We also conducted a third experiment in the Penalty Game setting, increasing the penalty k to -100 and also increasing the reward stochasticity (by increasing the variance of the reward

[17] BVPI did not do dramatically worse than WoLF-PHC in terms of *undiscounted* accumulated rewards, even in this case where several of the BVPI runs did not converge to the optimal or the suboptimal equilibrium. After 5000 iterations, BVPI agents accumulated a total (average over 30 runs) reward of $37,794.4$ (or, on average, a reward of 7.56 per iteration), while the WoLF agents collected a total (average) reward of $43,975.1$ (or, on average, 8.79 per iteration).



(a) $k = -20$, $\gamma = 0.75$, uninformed priors



(b) $k = -20$, $\gamma = 0.95$, uninformed priors

Figure 3.5: Penalty Game Results, $k = -20$; y axis is discounted accumulated reward (averaged over 30 runs).

	BOL	*BVPI*	*CBE*	*OBE*	*FMQ*	*WoLF-PHC*
OE	21	21	30	29	30	30
SE	9	5	0	0	0	0
NE	0	4	0	1	0	0

Table 3.5: Penalty game, k=-20, $\gamma = 0.75$: Number of runs converging to optimal equilibrium (OE), suboptimal equilibrium (SE) or non-equilibrium (out of 30 runs). Convergence to NE usually simply means the agents have not converged to playing some specific policy.

	BOL	*BVPI*	*CBE*	*OBE*	*FMQ*	*WoLF-PHC*
OE	14	25	30	29	30	30
SE	16	1	0	0	0	0
NE	0	4	0	1	0	0

Table 3.6: Penalty game, k=-20, $\gamma = 0.95$: Number of runs converging to optimal equilibrium (OE), suboptimal equilibrium (SE) or non-equilibrium (out of 30 runs). Convergence to NE usually simply means the agents have not converged to playing some policy.

received for each action). The discount factor was set to $\gamma = 0.95$. We provided informative priors for BOL, BVPI, CBE and OBE agents, giving the agents strong information about rewards by restricting the prior to assign (uniform) nonzero probability only to the small range of truly feasible rewards for each action. We wanted to test whether informative priors would allow the agents to counter the effects of increasing penalty and stochasticity in this setting.

Results are shown in Figure 3.6 (averaged over 30 runs). For interest, we also plot the results of the Bayesian methods with uninformed priors in the same graph. BVPI and BOL outperform all other methods in term of dicounted reward; however, CBE and OBE failed to take advantage of the informed priors, suffering big penalties in the initial learning stages, once again ranking after WoLF. Not surprisingly, the BVPI and BOL agents with informative priors do better than their "uninformed" counterparts; however, because of the high penalty (despite the high discount factor), they converge to the suboptimal equilibrium most of the time (23 and 22 times, respectively, as shown in Table 3.7).

As seen in Table 3.7, none of the runs of the OBE or FMQ managed to converge in this increased penalty/informed priors scenario. The model-based OBE agents are optimistic, and also quite confident of the value of the joint actions (due to informed priors). Thus, they alternated in choosing individual actions 0 or 2, without being able to coordinate in this game that (unlike the Climbing Game) has multiple (2) optimal equilibria. This is due to the fact that Optimistic Boltzmann exploration does not allow the agents to account for the strategy of the

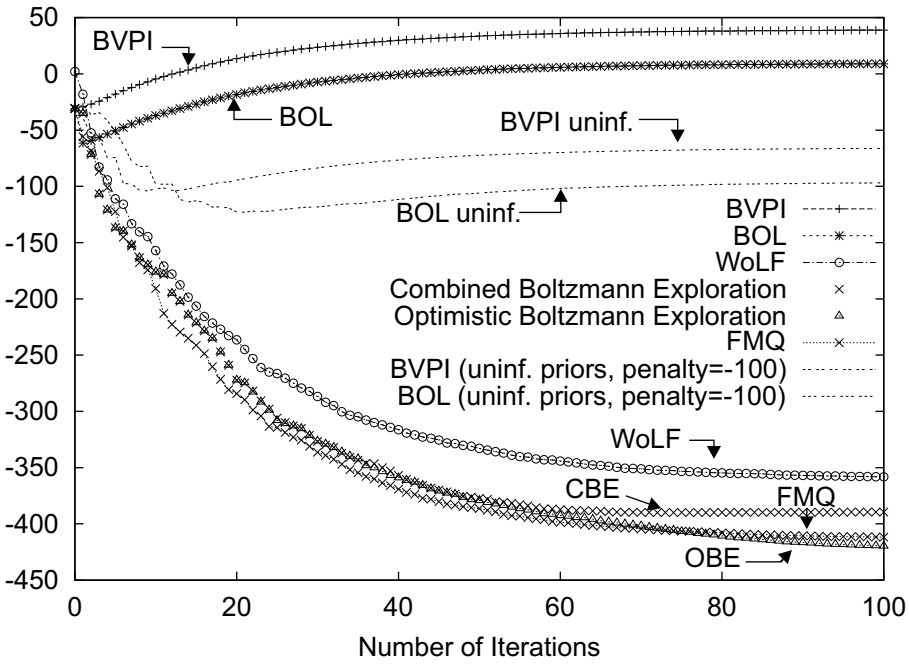


Figure 3.6: Penalty Game Results, $k = -100$, $\gamma = 0.95$, informed priors; y axis is discounted accumulated reward (averaged over 30 runs).

	BOL	*BVPI*	*CBE*	*OBE*	*FMQ*	*WoLF-PHC*
OE	8	7	28	0	0	29
SE	22	23	2	0	0	1
NE	0	0	0	30	30	0

Table 3.7: Penalty game, k=-100: Number of runs converging to optimal equilibrium (OE), suboptimal equilibrium (SE) or non-equilibrium (out of 30 runs). Convergence to NE usually simply means the agents have not converged to playing some policy.

other agent when estimating the value of their individual actions. As for the FMQ agents, they fell prey to the increased penalty and increased stochasticity (higher rewards' variance) of the domain, while not being able to benefit from informed priors (as they do not employ a model). Once again, it is worth noting that, in all cases, the agents employing the generic WoLF-PHC algorithm outperform agents employing other non-Bayesian methods in terms of discounted accumulated reward.

3.5.2 Multi-State Games

We also compared BVPI to WoLF-PHC in two multi-state coordination games. In both application settings, the state transitions and rewards are stochastic, and the agents have uninformative, uniform priors for the models describing state transitions, rewards and opponents' strategies.

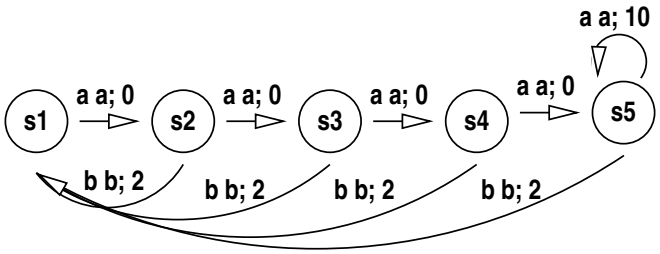
A Multiagent Chain World Domain

The first of our multi-state games is a version of the *Chain World* [DFA99] modified for multiagent coordination, and is illustrated in Figure 3.7(a).

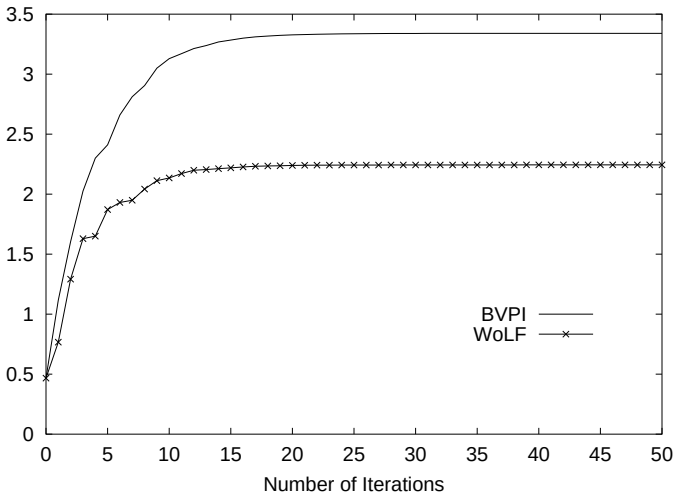
The optimal joint policy is for the agents to do action a at each state, though these actions have no payoff until state s_5 is reached. Coordinating on b leads to an immediate, but smaller, payoff, and resets the process.[18] Unmatched actions $\langle a, b\rangle$ and $\langle b, a\rangle$ result in zero-reward self-transitions (omitted from the diagram for clarity). Transitions are noisy, with a 10% chance that an agent's action has the "effect" of the opposite action. The original Chain World is difficult for standard RL algorithms, and is made especially difficult here by the requirement of coordination.

We compared BVPI to WoLF-PHC on this domain using two different discount factors, plotting the total discounted reward (averaged over 30 runs) in Figure 3.7(b) and (c). There were 50000 iterations per run. BVPI dominates Wolf-PHC in terms of online performance. BVPI converged to the optimal policy in 7 (of 30) runs with $\gamma = 0.99$ and in 3 runs with $\gamma = 0.75$, intuitively reflecting the increased risk aversion due to increased discounting. WoLF-PHC rarely even managed to reach state s_5, though in 2 (of 30) runs with $\gamma = 0.75$ it stumbled across s_5 early enough to converge to the optimal policy. It is obvious from the diagrams that the Bayesian approach manages to encourage intelligent exploration of action space in a way that trades off risks and predicted rewards; and we see increased exploration with the higher

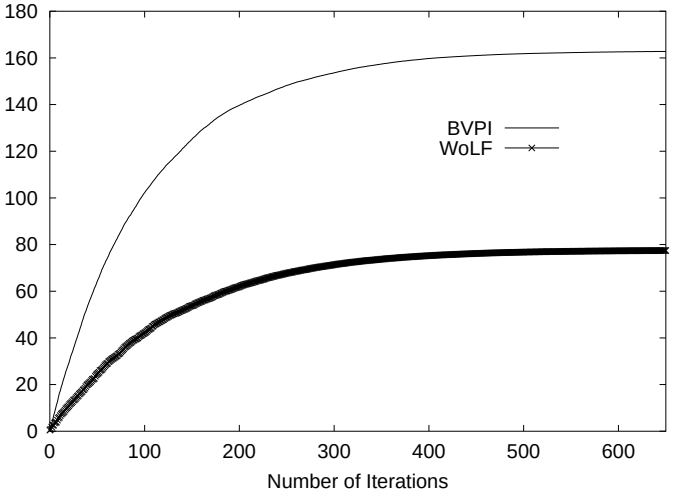
[18] As mentioned, the rewards are stochastic, with means shown in the figure.



(a) The *Multiagent Chain World*



(b) Results; $\gamma = 0.75$



(c) Results; $\gamma = 0.99$

Figure 3.7: *Multiagent Chain World*: Game and Results; y axis is discounted accumulated reward (averaged over 30 runs).

discount factor, as expected.

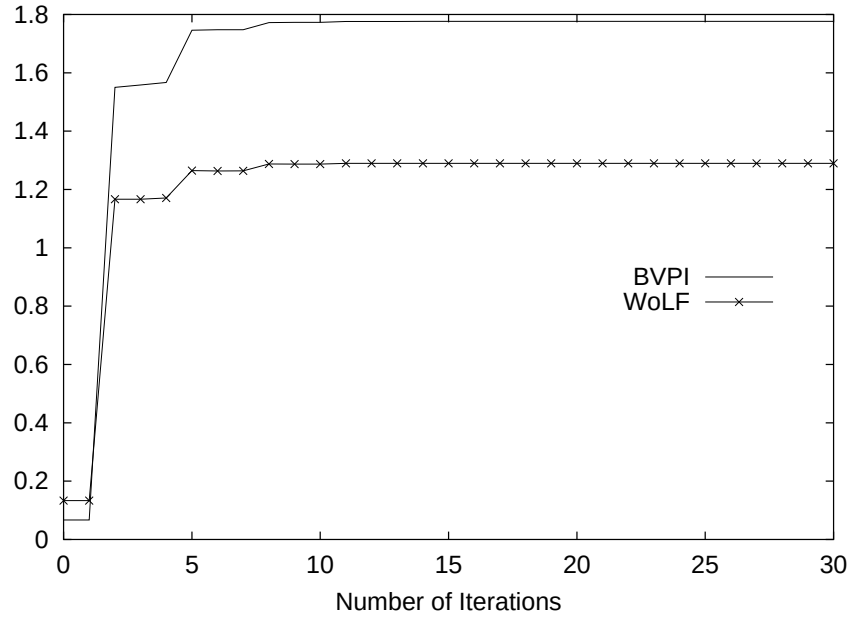
The Opt-In or Out Domain

The second multi-state game we experimented with is the "Opt-in or Out" game (Figure 3.1), discussed in Section 3.2. The transitions are stochastic, with the action selected by an agent having the "effect" of the opposite action with some probability. Two versions of the problem were tested, one with low "noise" (probability 0.05 of an action effect being reversed), and one with "medium" noise level (probability 0.11270167 of an action effect being reversed). With low noise, the optimal policy is as if the domain were deterministic (the first agent opts in at s_1 and both play a coordinated choice at s_2), while with medium noise, the "opt in" policy and the "opt out" policy (where the safe move to s_6 is adopted) have equal value.

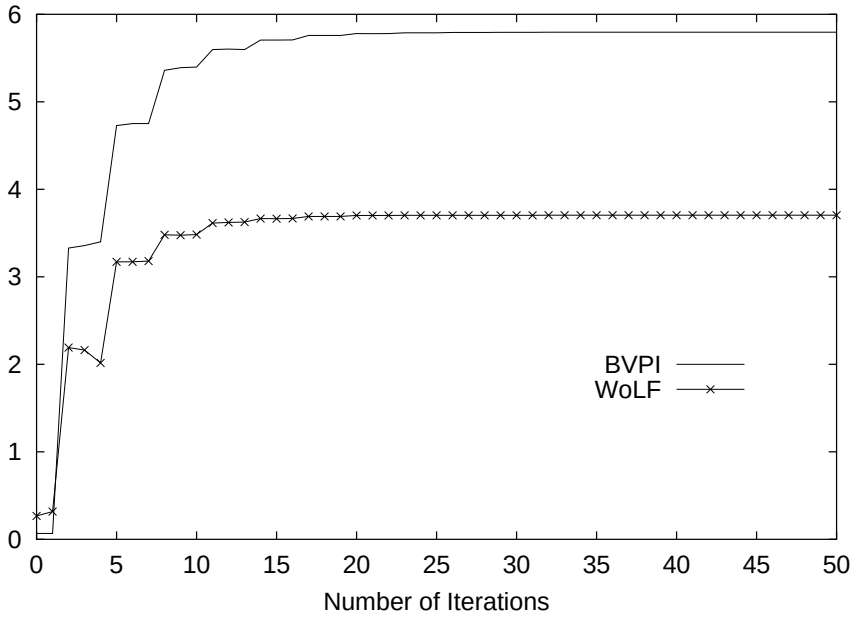
BVPI is compared to WoLF-PHC under three different discount rates, with low noise results shown in Figure 3.8 and Figure 3.9; and medium noise results in Figure 3.12 and Figure 3.13. Once again, BVPI dominates WoLF-PHC, in terms of dicounted reward averaged over 30 runs (with 5000 iterations per run).

The convergence results are presented in Figures 3.10(a) and 3.11(a). In the low noise problem, BVPI converged to the optimal ("opt-in and coordinate") policy in 18 (of 30) runs with $\gamma = 0.99$, in 15 runs with $\gamma = 0.75$ and 12 times with $\gamma = 0.5$. The WoLF-PHC agents converged in the optimal policy only once with $\gamma = 0.99$, but 17 times with $\gamma = 0.75$ and 10 with $\gamma = 0.5$ (probably because the lower discount factors helped them become more confident in this policy earlier). Notice, however, that even in the $\gamma = 0.75$ and the $\gamma = 0.5$ cases—when WoLF-PHC achieved convergence to the optimal policy several times—the average undiscounted accumulated reward was not significantly greater (or was even less) than its reward at the 0.99 case (when it converged to the optimal policy only once), as shown in Figure 3.10(b). This indicates that the WoLF-PHC agents encountered substantial difficulties in their attempts to coordinate while learning to play the optimal policy.

With medium noise, BVPI chose the "opt in" policy in 10 ($\gamma = 0.99$), 13 ($\gamma = 0.75$) and 9 ($\gamma = 0.5$) runs, but learned to coordinate at s_2 even when converging to the "opt out" policy. Interestingly, WoLF-PHC always converged on the "opt out" policy (recall that both policies are optimal with medium noise). Even in terms of undiscounted accumulated reward, BVPI always substantially outperforms WoLF (as shown in Figures 3.10(b) and 3.11(b)).



(a) $\gamma = 0.5$



(b) $\gamma = 0.75$

Figure 3.8: *Opt-In or Out* Results: Low Noise; y axis is discounted accumulated reward (averaged over 30 runs).

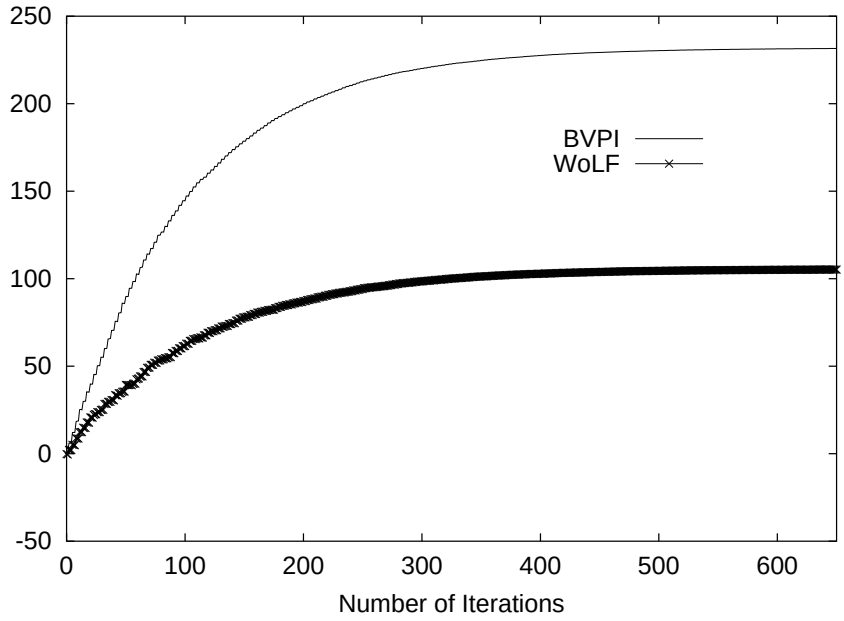


Figure 3.9: *Opt-In or Out* Results: Low Noise; $\gamma = 0.99$; y axis is discounted accumulated reward (averaged over 30 runs).

	BVPI	*WoLF-PHC*
$\gamma = 0.5$	12/30	10/30
$\gamma = 0.75$	15/30	17/30
$\gamma = 0.99$	18/30	1/30

(a) Number of runs converging to the "opt-in" (optimal) policy.

	BVPI	*WoLF-PHC*
$\gamma = 0.5$	11507.8	8912.1
$\gamma = 0.75$	11818.9	9217.2
$\gamma = 0.99$	12115.5	9107.5

(b) Average (undiscounted) accumulated reward (over 30 runs; 5000 iterations/run).

Figure 3.10: *Opt-in or Out* game, Low Noise; Convergence and Total Accumulated Reward.

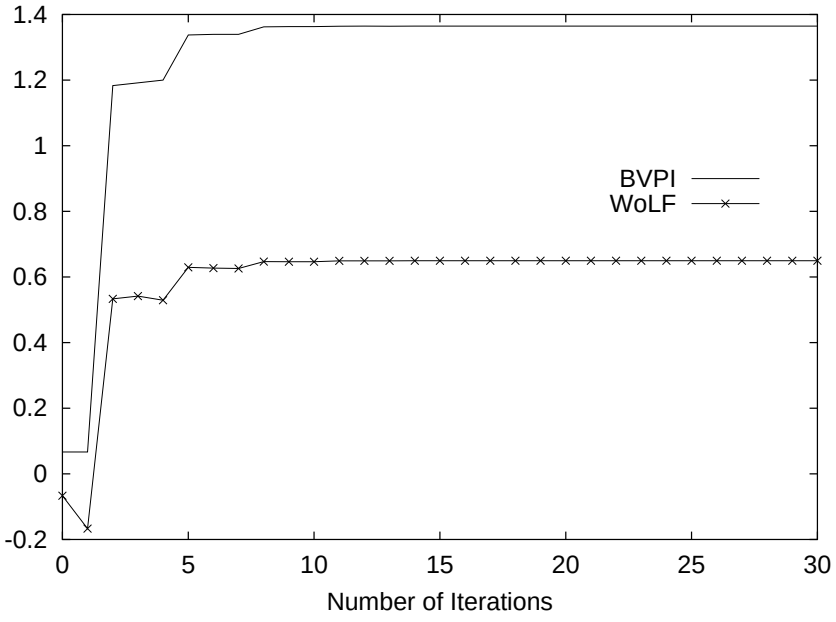
	BVPI	*WoLF-PHC*
$\gamma = 0.5$	9/30	0/30
$\gamma = 0.75$	13/30	0/30
$\gamma = 0.99$	10/30	0/30

(a) Number of runs converging to the "opt-in" policy (both policies are optimal).

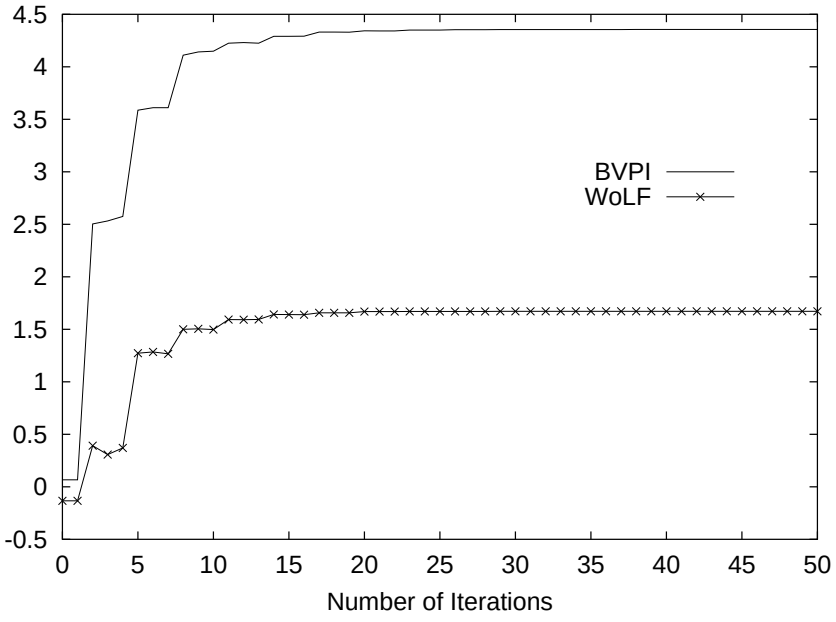
	BVPI	*WoLF-PHC*
$\gamma = 0.5$	9998.8	8325.2
$\gamma = 0.75$	10033.5	8414.9
$\gamma = 0.99$	10011.2	8565.8

(b) Average (undiscounted) accumulated reward (over 30 runs; 5000 iterations/run).

Figure 3.11: *Opt-in or Out* game, Medium Noise; Convergence and Total Accumulated Reward.



(a) $\gamma = 0.5$



(b) $\gamma = 0.75$

Figure 3.12: *Opt In or Out* Results: Medium Noise; y axis is discounted accumulated reward (averaged over 30 runs).

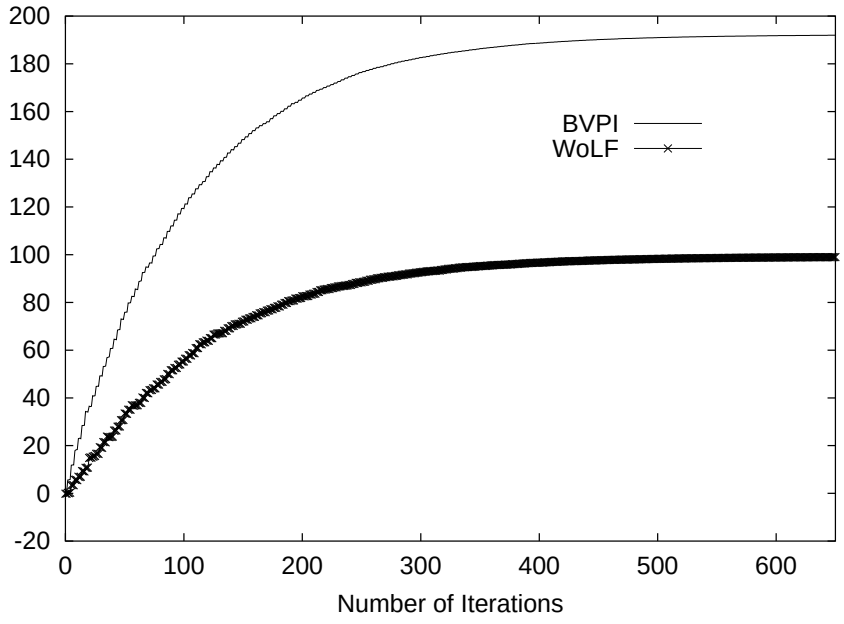


Figure 3.13: *Opt In or Out* Results: Medium Noise; $\gamma = 0.99$; y axis is discounted accumulated reward (averaged over 30 runs).

3.5.3 Discussion of Results

The experimental results presented demonstrate quite effectively that Bayesian exploration enables agents to make the tradeoffs described. Our results show that this ability can enhance online performance (reward accumulated while learning) of MARL agents in both single-state and multi-state coordination problems, when compared to heuristic exploration techniques that explicitly try to induce convergence to optimal equilibria. This implies that BAs run the risk of converging on a suboptimal policy; but this risk is taken "willingly" through due consideration of the learning process given the agent's current beliefs about the domain.

Still we see that BAs often find optimal strategies in any case. Key to this is a BA's willingness to exploit what it knows before it is very confident in this knowledge—it simply needs to be confident enough to be willing to sacrifice certain alternatives.[19] Apart from dominating other methods in terms of discounted accumulated reward, our experiments show that BAs perform well even in terms of undiscounted (total) accumulated reward, even when using uninformative priors.

When comparing the performance of the Bayesian methods (BVPI and BOL) to eachother,

[19] We note that the behaviour of Bayesian agents in multiagent settings, as observed in our experiments, matches their behaviour in single-agent settings—as reported in [DFR98, DFA99].

we note that the BVPI method seems to be the more robust of the two in the face of increased stochasticity (at least within the space of the single-state games, in which we pitted them against eachother). This is indicated by the results in Figure 3.4 and Figure 3.6, where BVPI is shown to outperform BOL (and all other methods). As mentioned earlier, we attribute this to the fact that BOL's myopic assumption (in single-state games), namely that the action with the highest expected reward in the immediate successor belief state will be the best action in all subsequent stages, is more likely to be flawed under increased stochasticity. Further, it is worth noting that BVPI achieves good sequential performance while converging less frequently than BOL to optimal equilibria. Nevertheless, given adequate computational power, one would expect a multistage-lookahead approach to do better than any myopic VPI-estimating algorithm—especially if the variance of the actual stochastic reward model is high (in which case it is not very realistic to expect that VPI can be adequately approximated with a small number of samples). On the other hand, of course, BVPI is the "computationally cheaper" Bayesian solution (as it does not require dealing with multiple successor belief states), and thus is a better candidate when computational power is an issue.

Though our results are definitely encouraging, our methods were only tested in domains limited in size and nature (i.e., cooperative). It is important to test our approach in more complex domains, including antagonistic ones, populated by large numbers of agents. When dealing with more complex domains, exploring the use of more elaborate sampling techniques, such as *importance sampling* and *repair*, would be definitely worthwhile—and the use of such sampling techniques is expected to be essential for our methods to scale. Further, though we did not explore this possibility here, it would certainly be of interest to assess sampling complexity, and study how varying the number of samples used would affect the performance of our algorithms. In general, we believe that efficient sampling is key to achieve scalability—as sampling is required given the intractability of the POMDP solution, even when adopting belief state lookahead approximations (since the number of successor belief states increases with the number of joint actions, rewards, and actual states).

In Chapter 7 we elaborate a bit more on ways to improve and test the scalability of our models and algorithms.

3.6 Conclusions

We have described a generic Bayesian approach to modeling MARL problems, that allows agents to explicitly reason about their uncertainty regarding the underlying domain and the

strategies of their counterparts. We have provided a formulation of optimal exploration under this model and developed appropriate computational approximations for Bayesian exploration in MARL. Further, our model-based Bayesian approach enables the agents to incorporate priors in their reasoning, with the obvious benefits in flexibility—though initially erroneous priors might lead to the inadequate exploration of the strategy space, and thus to convergence to suboptimal policies. However, this is a necessary tradeoff in "optimal learning", and key to achieving satisfactory sequential performance.

In addition to models of the environment, the agents maintain models of their opponents. Though we have experimented only with a simple opponent modeling technique—fictitious play—in this chapter, our formulation allows for the incorporation of more sophisticated opponent modeling solutions. Fictitious play, though simple, is probably adequate for the restricted class of repeated cooperative games we considered in our experiments here. However, more elaborate techniques are required as the games played become more complicated. For example, if a game in extensive form[20] is assumed to being played, one could imagine maintaining a prior over calculated equilibrium solutions corresponding to possible games potentially (given their reward uncertainty) being played by the agents—though the efficient and accurate calculation of such equilibria could be an issue. An agent could then sample that distribution to come up with a probabilistic assessment of an opponent taking an action. Along similar lines, assuming that the opponents' strategies could be represented by finite state machines (FSMs), each agent could maintain a prior over FSMs and use this to predict opponent behaviour. We discuss this issue in some more detail in Chapter 7.

Our experiments, though limited, clearly indicate that the Bayesian approach leads to more informed exploration for agents in multiagent settings. This results to better sequential performance, with the Bayesian methods dominating the heuristic approaches they were compared against in terms of discounted accumulated reward. Further, our Bayesian methods perform reasonably well in terms of convergence to optimal equilibria, even though they come without any such convergence guarantees.

In a nutshell, our work shows that Bayesian MARL methods allow the agents to learn how to behave well, while behaving well while learning.

[20] In Chapter 5 we elaborate on games in extensive form and their equilibrium solutions.

Chapter 4

Bayesian Coalition Formation

In this chapter, we provide a *Bayesian cooperative approach to coalition formation under uncertainty*. The creation of virtual organizations that have to interact under uncertainty regarding the *capabilities (types)* of potential partners provides a motivation for our research, suggesting potential applications in e-commerce: nowadays, there is an increasing need for open, decentralized computer systems that contain components representing distinct stakeholders with different aims and objectives [DJP03]. Moreover, *type uncertainty* in the context of coalition formation, which we coin here, poses interesting theoretical questions, such as the discovery of analogs of the traditional concepts of coalitional stability. Traditional stability concepts do not deal with the problem of uncertainty. In reality, however, uncertainty—regarding the types of potential partners and the effects of actions that coalitions might take—influences the decisions agents make in the coalition formation process, and the stability of the subsequently formed coalitions.

Moreover, traditional cooperative coalition formation disregards—to a large extent—the underlying bargaining process by which coalitions emerge. Nevertheless, increasingly, research on *dynamic* coalition formation has tackled both the dynamics of the process by which coalitions emerge, and the question of their stability. However, this research has not dealt extensively with the problem of uncertainty, either—and, perhaps surprisingly, it has not dealt with type uncertainty at all.

To tackle those realistic and interesting issues, in Section 4.2 we define a Bayesian coalition formation model that enables agents to have *expected* values about coalitions, given their uncertainty regarding partners' types. In this model agents must derive coalitional values by reasoning about the types of other agents *and* the uncertainty inherent in the actions a coalition may take (and the outcomes of those actions). To the best of our knowledge, ours is the first

coalition formation model to deal with both those forms of uncertainty at the same time, and is the first to deal with type uncertainty at all.

In Section 4.3 we define a stability concept under uncertainty, the *Bayesian Core (BC)*, presenting three versions of it: the *weak*, the *strict* and the *strong* BC—each one of them suitable for describing a different notion of stability under uncertainty. Then, in Section 4.4, we deal with the question of verifying the existence of stable coalitional configurations in our setting, and we provide an algorithm to decide whether the BC is non-empty. In Section 4.5 we present algorithms for *dynamic* coalition formation under uncertainty, thus linking the stability question with the formation question (under uncertainty). We prove that one of these algorithms, *Best Reply with Experimentation (BRE)*, leads to stable (strong BC) structures.[1] Finally, in Section 4.6 we present some simple experiments used to verify the convergence properties of our algorithms empirically.

Overall, we show that our framework and algorithms enable the agents to reach coalitional and payoff configurations that are stable given their beliefs regarding the types of others and the values of coalitions. We believe that these ideas could be of value for e-commerce and grid computing applications where *trust* among potential partners is an issue (see, e.g., [RRRJ07, TJJL06]), and also in general in environments where agents seek cooperative solutions to problems of resource sharing and task allocation.

We start this chapter by providing a brief review of related work in Section 4.1. Parts of the research described in this chapter appeared originally in [CB04] and [CMB07].

4.1 Related Work

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 here—uncertainty regarding both the knowledge of the types of others and the potential outcomes of coalitional actions. However, various coalition formation processes and some types of uncertainty have been studied. Here we present briefly some related work, upon which we draw.

Dieckmann and Schwalbe [DS98] recognize the need to deal with coalition formation in a dynamic context, combining the study of questions of stability with the explicit monitoring of the process by which the coalitions form. They describe a *dynamic* process of coalition forma-

[1] Some learning mechanism could prove to be valuable for tackling type uncertainty. However, we deal with this issue in subsequent chapters.

tion (a formation process that induces an underlying Markov process), but they do so under the usual deterministic model, which assumes full information regarding coalitional values. This process allows for exploration of suboptimal "coalition formation actions." At each stage of the process, assuming a given configuration of a coalition structure and associated allocations of payoffs (or "demands"), with some specified small random probability γ, *any* player independently may decide which of the existing coalitions to join, and states a (possibly different) own payoff demand. A player will join a coalition if and only if 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 prevailing in the beginning of the 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 coalitional structure and the demands of the other agents in the beginning of the period, and expect the current coalition structure and demand to prevail in the next period—which is not unrealistic if γ is small. (It is assumed that formed coalitions do not abandon the process, but the agents continue to be present and participate in the process until the end of all bargaining rounds. There are no explicit proposers or responders: rather, the process evolves by the agents adjusting their coalitions and demands as long as adjustments are feasible, given the configuration in place at each point in time.)

In some more detail, the process where all players adopt the best-reply rule corresponds to a finite Markov chain with state space

$$\Omega = \{\omega = (CS, \boldsymbol{d}) | CS \in \mathcal{SC}, \boldsymbol{d} \in \times_{i \in N} D_i\}$$

where $\mathcal{SC}$ is the space of all possible coalition structures and D_i corresponds to a finite set of demands[2] for player i. Letting $S(i)$ denote the coalition to which i belongs in any state ω, the transition probability from ω to ω' with corresponding (new) demand d_i' and coalition $S'(i)$ is then

$$\mathcal{P}_{\omega\omega'} = \prod_{i \in N} \gamma \beta_i(\omega' | \omega)$$

where β_i is defined by the best-reply rule as follows:

$$\beta_i(\omega' | \omega) > 0 \quad \textbf{iff}$$

$\{$ $d_i' = d_i(\omega)$, where $d_i(\omega)$ equals the maximum possible payoff for i given the other players'

[2]The demands are restricted to a finite set for reasons of computational tractability.

demands:

$$d_i(\omega) = \max_{S \in C \cup \{\emptyset\}} v(S \cup \{i\}) - \sum_{\substack{j \neq i \\ j \in S}} d_j$$

s.t. $d_i \in D_i$ and $S'(i)$ is a coalition in which d'_i can be achieved given the demands of its members }

In words, a state ω' is reached if and only if all "adjusting" players make their maximum feasible demands. This Markov process is shown to have at least one absorbing state. Now, if the players are assumed to *explore* with myopically suboptimal actions, the process is transformed to a related "best reply with experimentation" process.[3] [DS98] prove that if the core is non-empty, each core allocation corresponds to an absorbing state of this new process, and each absorbing state of this process can be associated with a core allocation. Furthermore, the process is proved to converge to a core allocation (i.e., an absorbing state) with certainty (if the core is non-empty). However, Dieckmann and Schwalbe's model does not explicitly allow for the agents to suggest and agree on coalitional actions to perform. Their work is influenced by the work of Agastya [Aga97], which is, unlike [DS98], confined to superadditive environments.

Konishi and Ray [KR02] study a somewhat related coalition formation process. Coalitions move to a new state (to a new coalition structure, accompanied by a corresponding payoff allocation) only if the move is profitable to all members of the coalition. The agents have common beliefs about the probability with which the state transitions may occur. There is no uncertainty regarding the payoff function or the partners' types.

Suijs *et al.* [SBWT99, SB99] 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. As was mentioned in Chapter 2, [SBWT99, SB99] use *relative* shares for the allocation of the residual of the stochastic coalitional values, and make the—in some cases unrealistic—assumption that agents have *common expectations* regarding expected coalitional values. These papers provide strong theoretical foundations for games with this restricted form of uncertainty, and describe classes of games for which the core of a SCG is non-empty—they basically focus on proving theoretical results about the existence of the core of such games, without modeling an explicit coalition formation process. Also, in contrast to our approach, no assumption of incomplete information about partners' types is made, and thus there is no direct translation of type uncertainty into coalition value uncertainty. However, [SBWT99] discusses the effect

[3] We omit the details here, but we note that this dynamic process is intuitively similar to the BRE process that we define later in this chapter—but with several important differences, as we will be explaining.

that different types of agents' risk behaviour might have concerning the existence of a core allocation within a specific class of SCG games.

Yamamoto and Sycara [YS01] and Li and Sycara [LS02] have also proposed versions of a core concept for use alongside coalition formation protocols enabling group buying and group bidding activities in e-marketplaces under incomplete information. However, these core versions refer to stability based on payoff sharing within only a single coalition, and not across coalitions. Nevertheless, Li *et al.* [LCRS03] propose a different core concept refering to stability across all coalitions in an e-marketplace where the buyers come together to profit from group discounts. This core characterizes stability based on the *reported utility* of the coalitions' members. This paper also interestingly presents a payoff division protocol (or "mechanism") that is empirically shown to maximize the social welfare while leading to stable outcomes and incentivising the buyers to truthfully reveal their valuations.[4] Though incomplete information is assumed in all these three papers, there is also use of a central "group leader" agent or a manager to which the agents report their preferences.

More recently, Blankenburg *et al.* [BKS03] has dealt with coalition formation in *fuzzy cooperative games*, introducing the *fuzzy kernel* stability concept. However, this work makes the assumption that all agents share the same understanding regarding the fuzziness of coalitional values, and it does not deal with the more general notion of type uncertainty—nor does it deal with the problem of coalitional action selection.

Finally, Yokoo *et al.* [YCS+05] have coined the *anonymity-proof core* and the *core for skills* concepts, to be used in contexts where the skills (or, types) of agents are private information. However, these concepts do not take into account the uncertainty or beliefs of the agents regarding the type of others. Rather than ensuring that coalitions in the core are deviation-proof, which is a central requirement from any core concept, these concepts rely on the agents reporting their types to a special "mechanism designer" agent, whose task is to implement a payoff allocation function that produces an outcome in the core. They then proceed to show that such a function exists if a set of axioms is satisfied. This work does not deal with any sort of decentralized coalition formation process.

[4] However, these results are empirical: as a matter of fact, the paper presents a negative, impossibility result on the question of existence of an "incentive compatible" mechanism for such coalitional games.

4.2 A Bayesian Coalition Formation Model

The need to address type uncertainty, reflecting an agent's uncertainty about the abilities of potential partners, is critical to the modeling of realistic coalition formation problems. For instance, if a carpenter wants to find a plumber and electrician with whom to build a house, his decision to propose (or join) such a partnership, to engage in a specific type of project, and to accept a specific share of the surplus generated should all depend on his (probabilistic) assessment of their abilities. To capture this, we start by introducing the problem of *Bayesian coalition formation* under type uncertainty. We then show how this type uncertainty can be translated into coalitional value uncertainty.

A *Bayesian coalition formation problem* (or game) 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:

Definition 3 (Bayesian coalition formation problem (BCFP)). *A Bayesian coalition formation problem (BCFP) is a coalition formation problem that is characterized by a set of agents, N; a set of types T_i for each agent $i \in N$; a set A_C of coalitional actions for each coalition $C \subseteq N$; a set $\mathcal{O}$ of stochastic outcomes (or states); with transition dynamics $Pr(s|\alpha_C, \boldsymbol{t}_C)$ denoting the probability of an outcome $s \in \mathcal{O}$ given that coalition C with members type vector $\boldsymbol{t}_C$ takes coalitional action α_C; a reward function $R : \mathcal{O} \longrightarrow \Re$; and agent beliefs B_i for each agent $i \in N$ comprising a joint distribution over types T_{-i} of potential partners.*

We now describe each of the BCFP 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". We let $T = \times_{i \in N} T_i$ denote the set of type profiles. For any coalition $C \subseteq N$, $T_C = \times_{i \in C} T_i$, 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 $B_i(\boldsymbol{t}_{-i})$ is the probability i assigns to other agents having type profile $\boldsymbol{t}_{-i}$. We use $B_i(\boldsymbol{t}_C)$ 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. We can think of A_C as the set of decisions available to C on how to deal with the underlying task at hand—or even a decision on what task to deal with. When an action is taken, it results in some outcome or *state* $s \in \mathcal{O}$. 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 capabilities of the team members). We let $\Pr(s|\alpha, \boldsymbol{t}_C)$ denote the probability of outcome s given that coalition C takes action $\alpha \in A_C$ and member types are given by $\boldsymbol{t}_C \in T_C$. Finally, we assume that each stochastic 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.

We can also define a BCFP *subgame* as follows:

Definition 4 (Subgame of a Bayesian coalition formation problem). *Let N be a set of agents, and $S \subseteq N$. The S-agent* subgame *of a Bayesian coalition formation problem (game) with N agents, is defined as the BCFP with agents S whose sets of types, beliefs, coalitional actions, outcomes, transition dynamics and reward function are the restriction of their corresponding elements in the N-agent problem.*

Thus, for example, an agent i in the S-agent subgame has the same beliefs regarding potential partners in S as it has in the N-agent game.

Now we turn to the problem of showing how the type (and action) uncertainty that is incorporated in a BCFP's definition can be translated into coalitional value uncertainty. In a BCFP setting, the *value* of coalition C with members of type $\boldsymbol{t}_C$ is:

$$V(C|\boldsymbol{t}_C) = \max_{\alpha \in A_C} \sum_s \Pr(s|\alpha, \boldsymbol{t}_C) R(s) = \max_{\alpha \in A_C} Q(C, \alpha|\boldsymbol{t}_C) \tag{4.1}$$

where, intuitively, $Q(C, \alpha|\boldsymbol{t}_C)$ represents the value (or quality) of coalitional action α to coalition C that is made up of members with types $\boldsymbol{t}_C$. $V(C|\vec{t}_C)$ therefore represents the (maximal) payoff that coalition C can obtain by choosing the best coalitional action. Unfortunately, this coalition value cannot be used in the coalition formation process if the agents are uncertain about the types of their potential partners (since any potential partners may have one of several types, any agent in any C would be uncertain about the type profile $\boldsymbol{t}_C$ of its members, and thus about the value $V(C)$). However, each agent i has beliefs about the value of any coalition based on its expectation of this value with respect to other agents's types:

$$V_i(C) = \max_{\alpha \in A_C} \sum_{\boldsymbol{t}_C \in T_C} B_i(\boldsymbol{t}_C) Q(C, \alpha|\boldsymbol{t}_C) = \max_{\alpha \in A_C} Q_i(C, \alpha) \tag{4.2}$$

where, intuitively, $Q_i(C, \alpha)$ represents the expected value (or, expected quality) of α to coalition C, according to i's beliefs. Note that $V_i(C)$ is not simply the expectation of $V(C)$ with respect to 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 (e.g, i may believe that k is extremely competent, while j may believe that k is incompetent; thus, i will believe that coalition $\langle i, j, k\rangle$ has a much higher value than j does). 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:

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

4.3 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 [SBWT99, SB99]. Stability is rendered 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 agents join a coalition with certain *relative payoff demands*. Intuitively, since the agents cannot expect to have an accurate estimate of the coalition payoffs (and, consequently, the payoff shares of coalition members), it is more natural for them to take into consideration relative demands; these correspond to the *perceived* "power structure" within the coalition and can be used for the allocation of unexpected gains or losses:[5]

Let $\boldsymbol{d}$ represent the payoff demand vector $\langle d_1, \ldots, d_n\rangle$, and $\boldsymbol{d}_C$ the demands of those agents in coalition C, assuming that these (actual) demands are observable by all agents. 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 gains 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 (maximum) *expected stochastic payoff* of some agent

[5]Incidentally, it should be clear given this transferable utility-based model that it is not true that the agents would prefer to participate in coalitions with skilled partners (since this would not necessarily increase their individual payoffs).

$j \in C$ is denoted

$$\bar{p}_j^i = r_j V_i(C)$$

with r_j being the relative demand of agent j given the stated demands of the agents in C, and $V_i(C)$ the value that i expects C to have (recall that $V_i(C)$ is defined as the maximum over coalitional actions). Similarly, if $i \in C$, i believes its *own* (maximum) 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 i's assessment of the best (expected reward-maximizing) action for C is not necessarily shared by the rest of the agents: they most probably have their own views on the issue (for example, j might believe that it is better for coalition C in which j and i belong to take action α_1, while i believes—because he has his own estimates regarding C's members capabilities—that it is better for C to perform α_2). Therefore, we suppose instead 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 $\bar{p}_j^i(\alpha, C) = r_j Q_i(C, \alpha)$. Finally, we let $\bar{p}_j^i(C, \boldsymbol{d}_C, \alpha)$ denote i's beliefs about j's expected payoff if it were a member of any $C \subseteq N$ with demand $\boldsymbol{d}_C$ taking action α:

$$\bar{p}_j^i(C, \boldsymbol{d}_C, \alpha) = \frac{d_j Q_i(C, \alpha)}{\sum_{k \in C} d_k} = r_j Q_i(C, \alpha) \tag{4.3}$$

In the same way, we define:

$$\bar{p}_i^i(C, \boldsymbol{d}_C, \alpha) = \frac{d_i Q_i(C, \alpha)}{\sum_{k \in C} d_k} = r_i Q_i(C, \alpha) \tag{4.4}$$

Intuitively, if a coalition structure and payoff allocation are stable, we would expect that no agent believes it will receive a payoff (in expectation) that is less than its reservation value. Further, Bayesian stability, given that beliefs may vary widely across the agents, may have several dimensions, such as the following: (a) 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, and (b) even if there exist agents that may believe that deviation can pay off, (Bayesian) stability will still depend on the beliefs of their potential partners.

Thus, we first define the *weak Bayesian core (BC)* of a BCFP as the set of coalitional configurations—each consisting of a coalition structure, a demand (or a relative demand) vec-

tor, and a coalitional action vector—that satisfy the above requirements in the following manner:

Definition 5 (weak Bayesian core). *Let $\langle CS, \boldsymbol{d}, \boldsymbol{\alpha} \rangle$ be a coalition structure-demand vector-action vector triplet, with C_i denoting the $C \in CS$ of which i is a member (and let $\boldsymbol{r}$ be the relative demand vector corresponding to $\boldsymbol{d}$). Then, $\langle CS, \boldsymbol{d}, \boldsymbol{\alpha} \rangle$ (or, equivalently, $\langle CS, \boldsymbol{r}, \boldsymbol{\alpha} \rangle$) is in the* weak Bayesian core *of a BCFP iff there is no coalition $S \subseteq N$, demand vector $\boldsymbol{d}_S$ and action $\beta \in A_S$ s.t. $\vec{p}_i^{\,i}(S, \boldsymbol{d}_S, \beta) > \vec{p}_i^{\,i}(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$, $\forall i \in S$, where $\boldsymbol{d}_{C_i}, \alpha_{C_i}$ is the restriction of $\boldsymbol{d}, \boldsymbol{\alpha}$ to the C_i coalition.*

In words, there exists no coalition all of whose members each believe that they (personally) can be better off in it (in terms of expected payoffs, given some choice of action) than they currently are (within the current weak Bayesian core configuration). The agents' beliefs, in every $C \in CS$, "coincide" in the weak sense that there is a payoff allocation $\boldsymbol{d}_C$ and some coalitional action α_C that is commonly believed to ensure a better payoff. This doesn't mean that $\boldsymbol{d}_C$ and α_C is what each agent believes to be best. But an agreement on $\boldsymbol{d}_C$ and α_C is enough to keep any other coalition S from forming. Even if one agent proposed its formation, others would disagree because they would not expect to become strictly better off themselves. Notice that *the (deterministic) core is a special case of the weak Bayesian core* of a game, where all the agents have perfect information regarding the types of others (and, therefore, regarding coalitional values).

In BCFPs with *continuous* payoffs, the transferability of utility implies that if a new coalition makes any agent strictly better off with respect to his beliefs *without* making other agents worse off with respect to their own beliefs, then it can make all members strictly better off (with respect to their beliefs) through a suitable adjustment of relative demands.[6] However, if we assume *finite* demands, this is no longer the case. We can then define a stronger version of the Bayesian core, by demanding that there is *no* agent who believes that there exists a coalitional agreement that can make it strictly better off while not hurting the other members of the coalition, according to their own beliefs:

Definition 6 (strict Bayesian core). *Let $\langle CS, \boldsymbol{d}, \boldsymbol{\alpha} \rangle$ be a coalition structure-demand vector-action vector triplet, with C_i denoting the $C \in CS$ of which i is a member (and let $\boldsymbol{r}$ be the relative demand vector corresponding to $\boldsymbol{d}$). Then, $\langle CS, \boldsymbol{d}, \boldsymbol{\alpha} \rangle$ (or, equivalently, $\langle CS, \boldsymbol{r}, \boldsymbol{\alpha} \rangle$) is*

[6] Intuitively, if a new deal X makes a member of, say, a 4-agent coalition strictly better off by 4ϵ (where ϵ small) than an old deal Y, without hurting the rest of the agents with respect to their beliefs, then there exists a deal Z under which all 4 members can be strictly better off by ϵ each, rather than under the deal Y.

in the strict Bayesian core *of a BCFP iff there is no coalition $S \subseteq N$, demand vector $\boldsymbol{d}_S$ and action $\beta \in A_S$ s.t., for some $i \in S$,*

$$\bar{p}_i^i(S, \boldsymbol{d}_S, \beta) > \bar{p}_i^i(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$$

and

$$\bar{p}_j^j(S, \boldsymbol{d}_S, \beta) \geq \bar{p}_j^j(C_j, \boldsymbol{d}_{C_j}, \alpha_{C_j})$$

$\forall j \in S$, $j \neq i$.

The stability condition is a bit different now: Intuitively, agents "allow" the formation of coalitions in which they may be *weakly* better off. Of course, in the continuous demands case the strict BC coincides with the weak BC—specifically, the $\bar{p}_j^j(S, \boldsymbol{d}_S, \beta) \geq \bar{p}_j^j(C_j, \boldsymbol{d}_{C_j}, \alpha_{C_j})$ condition above loses its significance. This is because, as was explained above, the continuity of the payoffs ensures that there are always ways to make the partners of the strictly better off agent strictly better off themselves, if they believe that the new deal will not hurt them. (This results to points that would have been in the *weak BC* in the finite demands case to not belong to the *weak BC* anymore in the continuous demands case—we will demonstrate this through an example shortly.) But in the finite demands' case, for which the strict BC is defined, these concepts are distinct and this new core concept is stricter, because we now demand that there is not even one agent that believes he will be better off in some S, with the others believing that S is not harmful. The "strict" core is stricter in the sense that it is a subset of the weak core:

Observation 3. *The strict Bayesian core is a subset of the weak Bayesian core.*

To see an example of this, and highlight the differences between the strict and the weak BC, consider the following scenario (in which we assume for simplicity that there exists only one coalitional action possible for all coalitions):

Example 1. *Assume a BCFP with finite demands, discretized by $\delta = 10\%$, with participating agents a, b, y, z and coalitions $C_1 = \langle a, b \rangle$, $C_2 = \langle y, z \rangle$ and $S = \langle a, b, y, z \rangle$ with payoff allocations $\boldsymbol{d}_{C_1}$, $\boldsymbol{d}_{C_2}$, $\boldsymbol{d}_S$ such that:*

$$p_a^a(C_1, \boldsymbol{d}_{C_1}) = 100, p_b^b(C_1, \boldsymbol{d}_{C_1}) = 100$$

$$p_y^y(C_2, \boldsymbol{d}_{C_2}) = 200, p_z^z(C_2, \boldsymbol{d}_{C_2}) = 200$$

and

$$p_a^a(S, \boldsymbol{d}_S) = 150, p_b^b(S, \boldsymbol{d}_S) = 100, p_y^y(S, \boldsymbol{d}_S) = 200, p_z^z(S, \boldsymbol{d}_S) = 200$$

with $\boldsymbol{d}_S = \{10\%, 10\%, 40\%, 40\%\}$ for a, b, y, z respectively, and $\boldsymbol{d}_{C_1} = \boldsymbol{d}_{C_2} = \{50\%, 50\%\}$ for their members; say that the corresponding beliefs of the agents for the values of these coalitions are $\upsilon_a(C_1) = \upsilon_b(C_1) = 200$ (i.e., agents a and b share the same beliefs for the value of C_1), $\upsilon_y(C_2) = \upsilon_z(C_2) = 400$ (similarly, y and z both believe that $\upsilon(C_2) = 400$), while $\upsilon_a(S) = 1500$, $\upsilon_b(S) = 1000$, $\upsilon_y(S) = 500$ and $\upsilon_z(S) = 500$; and say that for all other potential coalitions the estimated p_i^i values of any agent i are strictly less than those above (say zero).

Then, configuration $\langle\{C1, C2\}, \boldsymbol{d}_{C_1C_2} = \langle \boldsymbol{d}_{C_1} \circ \boldsymbol{d}_{C_2}\rangle\rangle$[7] is in the weak BC: Even though a believes that $p_a^a(S, \boldsymbol{d}_S) = 150$, and has an incentive to propose S, not all *agents in S believe that they are strictly better off (all others expect to receive the same payoffs as in $\langle\{C1, C2\}, \boldsymbol{d}_{C_1C_2}\rangle$). Thus, $\langle\{C1, C2\}, \boldsymbol{d}_{C_1C_2}\rangle$ is* stable *in this sense—if S is proposed, others will say "no" because they are not strictly better off.*[8] *(Also, trivially, if a would suggest any other payoff allocation in the coalition C_1, instead of $\boldsymbol{d}_{C_1}$, in which he would be better off—i.e., would expect a higher share—b would have disagreed because he would actually expect to be worse off in this two-agent coalition).*

However, $\langle\{C1, C2\}, \boldsymbol{d}_{C_1C_2}\rangle$ is not *in the strict BC. This is because* there exists *one agent, a, that believes he is strictly better off in S, and others believe they won't be harmed in S (and so, intuitively, assuming that a proposes S, it will form, given the stability requirement of the "strict" BC concept—thus, $\langle\{C1, C2\}, \boldsymbol{d}_{C_1C_2}\rangle$ is not stable).*

It is trivial to show that configuration $\langle\{S\}, \boldsymbol{d}_S\rangle$ belongs in both the weak and the strict BC. Thus, in this example the weak BC has two elements (as we assumed that all agents expect strictly less payoff in every other coalition), and the strict BC has only one element (and is a subset of the weak BC).

The definition of the strict BC allows us to get some interesting results for games with finite demands. We present these results later in the current and next chapter of this dissertation.

We can also define a different stability concept, which we call the *strong BC* (and which is defined both for finite and continuous demands). The strong BC requires that there is no agent who believes there is an agreement that can make it better off and that it expects all partners to accept based on (its subjective view of) their expected payoff. This differs inherently from the weak and the strict core in that the agent assesses its own beliefs about the value of an

[7] Notation $\boldsymbol{d} \circ \boldsymbol{x}$ denotes a vector that is a union of disjoint vectors.

[8] Notice that if the demands were not discretized as they are, $\langle\{C1, C2\}, \boldsymbol{d}_{C_1C_2}\rangle$ would *not* have been in the weak BC: there would have been a continuum of allocations for S that could have made *all* agents in S strictly better off than in $\langle\{C1, C2\}, \boldsymbol{d}_{C_1C_2}\rangle$, given their beliefs on the coalitional values; for example, one such allocation for S could have been the allocation $\{9.7\%, 10.1\%, 40.1\%, 40.1\%\}$ for agents a, b, y, z respectively.

agreement to its partners.

Definition 7 (strong Bayesian core). *Let $\langle CS, \boldsymbol{d}, \boldsymbol{\alpha} \rangle$ be a coalition structure-demand vector-action vector triplet, with C_i denoting the $C \in CS$ of which i is a member (and let $\boldsymbol{r}$ be the relative demand vector corresponding to $\boldsymbol{d}$). Then, $\langle CS, \boldsymbol{d}, \boldsymbol{\alpha} \rangle$ (or, equivalently, the $\langle CS, \boldsymbol{r}, \boldsymbol{\alpha} \rangle$ triplet) is in the* strong Bayesian core *iff there is no coalition $S \subseteq N$, demand vector $\boldsymbol{d}_S$ and action $\beta \in A_S$ s.t. for some $i \in S$*

$$\bar{p}_i^i(S, \boldsymbol{d}_S, \beta) > \bar{p}_i^i(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$$

and

$$\bar{p}_j^i(S, \boldsymbol{d}_S, \beta) \geq \bar{p}_j^i(C_j, \boldsymbol{d}_{C_j}, \alpha_{C_j})$$

$\forall j \in S$, $j \neq i$.

The strong BC describes a notion of stability that is more tightly linked to the agents' subjective views on the potential acceptability of their proposals and is thus more "endogenous" in nature. By comparison, stability in the strict BC concept (and weak BC) is somewhat distinct. In an element of the strict BC, there may be an agent i who believes it would be strictly better off in some other coalition, and who believes all of its proposed partners would be better off as well; but the coalition may be considered unacceptable to some proposed partner j (who might believe that it will be hurt—that is, that he will *not* be even *weakly* better off), since its beliefs about the value of the coalition are different than those of i.

The following is an obvious fact (for the case of finite demands, for which the stict BC is defined):

Observation 4. *If the agents' beliefs coincide, the strict and the strong Bayesian core coincide.*

(In that case, the $\bar{p}_j^i$ and $\bar{p}_j^j$ estimates coincide, and the concepts' definitions become exactly identical.)

We can also make the following observation[9]:

Observation 5. *Let $\langle CS_N, \boldsymbol{d}, \boldsymbol{\alpha} \rangle$ be an element of (any version of) the BC of a BCFP with agents N. If $S \in CS_N$, $L = N \setminus S$, $CS_L = CS_N \setminus S$ and $\boldsymbol{d}_L, \boldsymbol{\alpha}_L$ is the restriction of $\boldsymbol{d}, \boldsymbol{\alpha}$ to the agents in L, the tuple $\langle CS_L, \boldsymbol{d}_L, \boldsymbol{\alpha}_L \rangle$, which is contained in the $\langle CS_N, \boldsymbol{d}, \boldsymbol{\alpha} \rangle$ configuration, is an element of the BC of the corresponding BCFP subgame with L agents.*

[9]This observation will prove to be useful in Chapter 5 when proving a proposition linking bargaining equilibria to the BC concept.

Proof: Assume that the configuration in the L-agent subgame is such that $\langle CS_L, \boldsymbol{d}_L, \boldsymbol{\alpha}_L \rangle$ is not in the subgame BC. Then it is not possible that its extension configuration $\langle CS_N, \boldsymbol{d}, \boldsymbol{\alpha} \rangle$ is in the N-agent game BC (contradiction).

This is because if, for example, an agent i in the $L = N \setminus S$ subset believed that he could do better in some coalition other than his current C_i, he would have believed this when the S coalition was present, as well; in that case, $\langle CS_N, \boldsymbol{d}, \boldsymbol{\alpha} \rangle$ couldn't have been a BC element. □
Notice that Observation 5 holds for all versions of the BC.

4.4 Existence of the Bayesian Core

Noting that the (deterministic) core is a special case of the (weak) Bayesian core, it is easy to show that (any type of) the Bayesian core does not always exist:

Proposition 1. *There exist BCFP for which the Bayesian core (weak, strict or strong) is empty.*

Proof: If the beliefs of the agents regarding all coalition values coincide, then the strong and the strict Bayesian core coincide (Observation 4), and they are a subset of the weak Bayesian core (by Observation 3).[10]

With the beliefs of all agents coinciding, which is possible only if all agents know the true types of other agents (since each agent knows its own true type), the BCFP is equivalent to a characteristic function game and the weak Bayesian core coincides with the (deterministic) core of the game. Since there exist characteristic function games with empty core, and any such game can be recast as a BCFP with "perfect" beliefs, it follows that there exist BCFP for which the Bayesian core is empty. □

In deterministic settings, or in uncertain settings that at least make superadditivity and some common knowledge assumptions (such as [SBWT99, SB99]), it is possible to provide generic necessary and sufficient conditions for the existence of the core. Briefly, such conditions specify the relationships between the various coalitional values of (any) given setting. The satisfaction of those conditions can be provided as the solution to a linear programming problem—in deterministic (and superadditive) environments—and it defines what is described

[10]This argument holds assuming finite demands. If continuous demands are assumed (in which case the strict BC is not properly defined), it is easy to show (by contradiction) that "if the beliefs of the agents coincide, then any point in the strong BC also lies within the weak BC".

by the term *balancedness* of a coalitional game. [Bon63] and [Sha67] used the duality theory of linear programming to show that a deterministic coalitional game has a non-empty core *iff* it is "balanced".

Naturally, when dealing with type uncertainty, it would be impossible for one to provide generic conditions for the existence of the Bayesian core without taking the beliefs of the various agents into account—since the BC existence depends on them. It is unlikely, however, that such *generic* conditions can be identified without making assumptions about some degree of common knowledge,[11] or about some form of coalitional value (super)additivity.[12] The former assumption is a concession that could significantly restrict the generality of any corresponding coalition formation model.[13] At the same time, there exists no obvious (at least to us) way to define superadditivity in an uncertain environment without making any such concessions.

Nevertheless, it would be of interest to devise algorithmic methods to try and establish the existence of the Bayesian core (or its non-existence) in specific games. This is of obvious practical value, since if we manage to devise such an algorithmic method for specific classes of games, or for games of specific sizes, we essentially achieve to compute the Bayesian core stability solution without turning to exhaustive search.

Of course, even in a deterministic setting where there is no uncertainty, testing for the nonemptiness of the core is an intractable problem. In fact, even in superadditive games, where the coalition structure is simply the grand coalition N and we only need to find an allocation of payoffs to the agents, there exist results that show that the problem is NP-hard [Chv78, Tan91, DP94, CS03]. In the absence of superadditivity, as is our case, there are even worse lower bounds on the complexity of the problem. Sandholm *et al.*[SLA$^+$99] show that in order to find an approximately optimal coalition structure[14] (i.e., in order to be able to establish a bound from the optimum), exponentially many (*at least* 2^{n-1}, if n agents) coalition structures have to be searched. All these suggest that an efficient algorithm for verification of the BC's

[11] Or even, perhaps, about the presence of a "more informed" agent with the ability to act as as a central manager assigning agents to coalitions, or as a trusted party disseminating information. However, such assumptions would constitute serious departures from a decentralized coalition formation model such as ours.

[12] We note, however, that in Chapter 5 we establish a result that can be used for the theoretical verification of the existence of the Bayesian core without making any such concessions. However, we established that result (Corollary 1) in a particular bargaining setting, and under specific assumptions regarding the bargaining strategies of the agents.

[13] The lack of superadditivity makes our framework more general, as we tackle the problem of forming rewarding coalition structures in addition to deriving acceptable payoff allocations. Furthermore, letting agents choose their coalitions makes our coalition formation model more easily applicable to a wide range of economic problems, such as "local public good economies", where participating individuals care about the number and characteristics of people in their coalition [Woo99].

[14] This was in terms of social welfare.

existence is unlikely.

Here we show that for a relatively small number of agents it is feasible to check for the nonemptiness of the Bayesian core without employing a brute-force approach that simply searches over all coalition structures. We formulate the problem as a constraint satisfaction problem (CSP) where the constraints are polynomial equalities or inequalities (i.e., we formulate it as a polynomial constraint satisfaction program). One could then use existing algorithms that solve such CSPs (we refer, e.g., to [BPR03] for algorithms with worst case analysis of their running time and to [Stu02] for heuristic approaches). Moreover, this program offers a concise way to describe existence conditions under the (realistic) uncertainty assumptions of our model.

Before we present the polynomial program that solves our problem, we make some simplifying assumptions. First, we assume that for each coalition there is a finite number d of possible relative demand vectors that one could propose (i.e., there is a finite number of possible ways in which the agents will split the payoff of the coalition). From a practical perspective, d should be reasonably small. This is of course only a coarse discretization of demands; nevertheless, such a discretization is not uncommon in realistic settings.[15] Also, without loss of generality, assume that each coalition has k available actions (these could have been coalition-specific, but we consider them to be the same for each coalition for notational simplicity).

The CSP that we present below tests the nonemptiness of the *weak* Bayesian core and has four types of variables. Similar CSPs can be written for the strict and the strong *BC* too. For each coalition S, there is a binary indicator variable X_S which indicates whether coalition S will form in the coalition structure that we are looking for. We also have a variable ρ_i for each agent i that indicates the share that i will have in the coalition to which he belongs. Furthermore, let $Q_i(S, j)$ denote the payoff that coalition S gets if the jth action is taken (recall there are k actions available). Then for each coalition S and action j, $j = 1, ..., k$ we have an indicator variable α_j^S that indicates whether action j is taken or not (if coalition S forms). Finally, for each possible deviation from the core, say $\langle T, \boldsymbol{r}, \beta \rangle$, where $\boldsymbol{r}$ is a $|T|$-dimensional relative demand vector $\langle r_1, ..., r_{|T|} \rangle$ and β is one of the k available actions to coalition T, we have an auxiliary variable $Z_{T,\boldsymbol{r},\beta}$ whose role is to ensure that it cannot be the case that all agents gain more expected payoff if they deviate to T. When we write "$\forall T, \boldsymbol{r}, \beta$" in the program below, we intend "$\forall T \subseteq N$, $\forall \boldsymbol{r} = \langle r_1, ..., r_{|T|} \rangle$, $\forall \beta \in \{1, ..., k\}$".

[15] Consider, for example, the case of stakeholders controlling shares of a company: it is reasonable to assume that each stakeholder controls an integer-valued percentage of the shares.

$$X_S(1 - X_S) = 0 \qquad \forall S \subseteq N \tag{4.5}$$

$$X_S(1 - \sum_{i \in S} \rho_i) = 0 \qquad \forall S \subseteq N \tag{4.6}$$

$$\rho_i \geq 0 \qquad \forall i \in N \tag{4.7}$$

$$\sum_{S: i \in S} X_S = 1 \qquad \forall\, i \in N \tag{4.8}$$

$$\alpha_j^S(1 - \alpha_j^S) = 0 \qquad \forall S \subseteq N, j \in \{1, ..., k\} \tag{4.9}$$

$$\prod_{i \in T}[Z_{T,\boldsymbol{r},\beta} - (r_i Q_i(T, \beta) - \rho_i \sum_{S: i \in S} X_S \sum_{j=1}^{k} \alpha_j^S Q_i(S, j))] = 0$$

$$\forall T, \boldsymbol{r}, \beta \tag{4.10}$$

$$Z_{T,\boldsymbol{r},\beta} \leq r_i Q_i(T, \beta) - \rho_i \sum_{S: i \in S} X_S \sum_{j=1}^{k} \alpha_j^S Q_i(S, j)$$

$$\forall T, \boldsymbol{r}, \beta, i \in T \tag{4.11}$$

$$Z_{T,\boldsymbol{r},\beta} \leq 0 \qquad \forall T, \boldsymbol{r}, \beta \tag{4.12}$$

Proposition 2. *The above program is feasible iff the weak BC of the corresponding game is non-empty.*

Proof. Suppose that the program is feasible and consider a solution. Then constraints (4.5) and (4.9) ensure that the variables X_S and α_j^S are integer $0/1$ variables. Hence we can see them as indicator variables, indicating which coalitions were chosen by the solution and which action was taken (if $X_S = 1$ we consider that coalition S forms). The constraints (4.8) ensure that the coalitions that form make up a coalition structure; each agent belongs to exactly one of them. Constraints (4.6) and (4.7) ensure that for any coalition that forms, the shares ρ_i for $i \in S$ form a valid (relative) demand vector. The rest of the constraints ensure that there is no coalition T, demand vector $\boldsymbol{r}$ and action β that would make all agents of T better off. For a coalition T and an agent $i \in T$, let ϵ_i be the amount by which i's payoff changes if he deviates from the solution to the program to $\langle T, \boldsymbol{r}, \beta \rangle$. The constraints (4.10) and (4.11) make sure that the variable $Z_{T,\boldsymbol{r},\beta}$ is equal to $\min_i \epsilon_i$ because the expression $\rho_i \sum_{S: i \in S} X_S \sum_{j=1}^{k} \alpha_j^S Q_i(S, j)$ is equal to the expected payoff of agent i under the feasible solution of the program (recall that only one of the variables X_S with $i \in S$ is set to 1 and the rest are 0). To elaborate, for any i—and for any agreement $\langle T, \boldsymbol{r}, \beta \rangle$—the variable $Z_{T,\boldsymbol{r},\beta}$ is less than or equal to (thus, the minimum

of) the difference between what i could get from $\langle T, \boldsymbol{r}, \beta \rangle$ and what his expected payoff in the feasible solution is. Finally, the last constraint ensures that for any $\langle T, \boldsymbol{r}, \beta \rangle$, $min_i \epsilon_i \leq 0$, which means that there is no coalitional agreement that can make all agents strictly better off.

Therefore, the feasible solution of the program consists of such values for the X_S, α_j^S and ρ_i variables that a coalition structure CS is defined (made up from the S coalitions for which $X_S = 1$), along with corresponding actions α_S and relative demand vectors $\langle \rho_i \rangle$ corresponding to $\boldsymbol{d}_S$ demands for each coalition. This configuration is such that there is no coalition $T \subseteq N$, demand vector $\boldsymbol{d}_T$ and action $\beta \in A_T$ s.t. $\bar{p}_i^i(S, \boldsymbol{d}_S, \beta) > \bar{p}_i^i(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$, $\forall i \in S$. In words, the feasible solution defines a configuration that is in the weak BC, and thus the weak BC is non-empty.

The reverse direction is straightforward. Given a configuration in the weak BC, with a corresponding configuration of X_S, α_j^S and ρ_i variables, the constraints of the above program are satisfied (by the weak BC definition), and thus the program is feasible. □

The number of variables in the program above is $O(kd2^n)$, where k is the number of actions, d the number of demand vectors, and n the number of agents. Moreover, the degree of the polynomials above can be (at worst) n, due to constraint 4.10. Although the worst case running time guarantees for such a program would be as high as $n^{O(kd2^n)}$ according to [BPR96] (which clearly is prohibitively high for most realistic settings), the program can be solved heuristically[16] for small problem size [Stu02]; further, it might be possible to linearize the program and solve it with an appropriate technique [Gro02]. Thus, in addition to providing a concise presentation of the existence constraints, in the case of small games the program enables us to solve the BC non-emptiness problem faster than a brute-force search approach.

We note, however, that in the next chapter we establish a result that can be used for the theoretical verification of the existence of the Bayesian core—but in a specific bargaining setting, and under specific assumptions regarding the bargaining strategies of the agents.

[16]For example, the feasibility of the polynomial program can be tested using *semidefinite programming* (SDP) techniques, such as the Dth SDP relaxation technique [Stu02]. This uses semidefinite programming (and appropriate software tools, such as Matlab) to find an *infeasibility witness* of degree D such that a polynomial identity—consisting of the system's polynomials multiplied with polynomials that are sums of squares—equals to zero. This technique is based on the *Real Nullstellensatz* theorem [Ste74].

4.5 Dynamic Coalition Formation

We now propose a protocol for dynamic coalition formation under uncertainty. The protocol is related to the one presented in [DS98], but with several important differences: it deals with expected, rather than certain, coalitional values (translating type uncertainty into coalitional value uncertainty); it requires the agents to explicitly state proposals and responses (since the feasibility of the agents proposals is not common knowledge); it has to account for the *expected*—rather than certain—feasibility of formation proposals, defined given each agent's beliefs; and it allows for the proposal of a coalitional action during formation.

We incorporate the proposal of coalitional actions in the dynamic formation process, since the coalitions, once formed, will eventually have to act. Therefore it is natural that deciding on the coalitional action to take should be part of the formation process. It is precisely for the value derived by an anticipated action to be performed that an agent decides to join a coalition in the first place. Moreover, since the agents have different views of the types of others and the values of coalitions, there exists no obvious "optimal" action for a coalition to take, thus it has to be agreed upon during negotiations. Finally, the incorporation of coalitional actions in the dynamic formation process, links this process with the stochastic cooperative games introduced by [SBWT99, SB99].

The process proceeds in stages. At any point in time, we suppose there exists a structure CS, (actual) demand vector $\boldsymbol{d}$, and a set of agreed upon coalitional actions $\boldsymbol{\alpha}_{CS}$ (with one $\alpha \in A_C$ for each $C \in CS$).[17] We assume that formed coalitions do not abandon negotiations, but the agents continue to be present and participate in the process until the end of all bargaining rounds. Therefore, we can define the state of the coalition formation game at time t as $\omega^t = (CS^t, \boldsymbol{d}_{CS^t}, \boldsymbol{\alpha}_{CS^t})$. This state is assumed to be observable by all agents. With probability $\gamma = 1/|N|$, agent i is given the opportunity to act as the *proposer*,[18] that is, to propose a change to the current structure. We permit i the following options: it can propose to stay in its current coalition without any changes in his demand or the coalitional action; 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 some demand d_i and a suggested coalitional action.[19] The second option includes the possibility that i "breaks away" into a

[17] We might initially start with singleton coalitions with the individually maximizing action, giving each agent its reservation value.

[18] We only allow for one proposer per negotiation round. In [DS98], any number of players were allowed to adjust their demands and/or join different coalitions at each round.

[19] Notice that he proposer is stating only his own actual demand d_i, and this is then "translated" into a relative demand by each other agent, given the demand vector currently in place. Another possibility would have been to

singleton coalition. Formally,

Definition 8. *A proposal (or proposed coalitional agreement) by proposer i is a triplet $\langle C \cup \{i\}, d_i, \alpha\rangle$ made by i to coalition $C \in CS \cup \emptyset$, with i stating a demand d_i for itself, and proposing that $C \cup \{i\}$ takes action $\alpha \in A_{C\cup\{i\}}$.*

(Notice that $C \cup \{i\}$ above is *not* a disjoint union operator, that is, i is allowed to propose to stay in a coalition $C \in CS$ in which it was already a member.) If i proposes a change to the current structure/demand/action, then the new arrangement will occur only if all "proposed-to" coalition members (those $j \in C$) agree to the change. Otherwise, the current structure and agreements remain in force. During the coalition formation process the beliefs of the various agents are considered to be fixed.[20]

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

Specifically, when proposing to join a coalition $C \in CS$, i should make the maximum demand (d_i and α) that is *expected to be feasible* according to its beliefs, in other words, that it believes the other agents will accept. More precisely:

Definition 9. *The proposal $\langle C \cup \{i\}, d_i, \alpha\rangle$ made by i to coalition $C \in CS$ (demanding d_i for itself and proposing that $C \cup \{i\}$ takes action α) is* expected to be feasible *for i if:*

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

In words, the definition above says that i expects that any $j \in C$ will be at least weakly better off accepting i's proposal than currently is (according to what i believes about the expected payoff $\bar{p}_j^i(C, \boldsymbol{d}_C, \alpha_C)$ of j in C—i does not know what the members of C believe, but does have its own estimates of their current values $\bar{p}_j^i(C, \boldsymbol{d}_C, \alpha_C)$[21]). If $\langle C \cup \{i\}, d_i, \alpha\rangle$ is (ex-

have the proposer state a relative demand; in that case, however, he would have been obliged to state the whole relative demand vector for the coalition (as just stating his own percentage share would have been ambiguous). That would have been more demanding in terms of the inter-agent communication.

[20] We do study the problem of belief updating (after execution of either bargaining or coalitional actions) in Chapters 5 and 6.

[21] This poses the compelling question of how best to model one agent's beliefs about another's beliefs in this setting. We will return to this issue and the complications arising from it later on in Chapter 6 (Section 6.2).

pected to be) feasible for i, then i expects the members of C to accept this demand. Agent i can therefore directly calculate its maximum *realistic* demand with respect to C and action α:

Definition 10. *The* maximum expected feasible (or maximal realistic) demand $d_i^{max}(C, \alpha)$ *stated by agent i towards coalition C, along with a proposal to perform coalitional action α, is given by the following formula:*

$$d_i^{max}(C, \alpha) = \min_{j \in C} \frac{d_j Q_i(C \cup \{i\}, \alpha) - \bar{p}_j^i \sum_{k \in C \cup \{i\}, k \neq i} d_k}{\bar{p}_j^i}$$

where $\bar{p}_j^i = \bar{p}_j^i(C, \boldsymbol{d}_C, \alpha_C)$.

Intuitively, i has to make such a demand that would satisfy even the tightest of the constraints imposed by the rationality of the agents in C (i.e., the constraints $\forall j \in C$, $d_i \leq \frac{d_j Q_i(C \cup \{i\}, \alpha) - \bar{p}_j^i \sum_{k \in C \cup \{i\}, k \neq i} d_k}{\bar{p}_j^i}$ implied by Def. 9 above.) This can be used to restrict the payoff demand of i:

Assumption 1 (Adapted to our setting from [DS98]). *Let $0 < \delta < 1$ be a sufficiently small* smallest accounting unit. *When any i makes a proposal $\langle C, d_i, \alpha \rangle$ 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 $[\delta, d_i^{max}(C, \alpha)]$.*

(Notice that i can always "propose" any $d_i \in D_i(\emptyset, \alpha)$, with α being a maximizer of expected payoff in $\langle i \rangle$, to itself and thus attain his reservation value rv_i.) For each state ω in the game, agent i's strategy set is denoted by:

$$\Sigma_i(\omega) := \{(C_i, d_i, \alpha) | C_i = C \cup \{i\}, C \in CS \cup \{\emptyset\}, \alpha \in A_{C_i}, d_i \in D_i(C, \alpha)\}$$

.

The strategy space is finite, since the number of possible coalitions and coalition structures and the number of possible demands are all finite.

With this model in place, we now define the following formation process, called the *best reply (BR) process*. In the BR process, proposers are chosen randomly as described above, and any proposer i is required to make its *maximum expected to be feasible coalition formation proposal*, asking for the formation of a coalition performing such an action so that his expected payoff is maximized:

$$\max_C \max_{\alpha \in A_{C \cup i}} \max_{d_i \in D_i(C, \alpha)} \bar{p}_i^i(C \cup \{i\}, \boldsymbol{d}_{C \cup \{i\}} = \boldsymbol{d}_C \circ d_i, \alpha) \tag{4.13}$$

Notice that d_i is constrained to belong to $D_i(C, \alpha)$: agent i considers only proposals that he expects to be feasible—and makes the maximal of them. Notice also that the $d_i \in D_i(C, \alpha)$ maximizing $\bar{p}_i^i(C \cup \{i\}, \boldsymbol{d}_{C\cup\{i\}} = \boldsymbol{d}_C \circ d_i, \alpha)$ is exactly $d_i^{max}(C, \alpha)$, the maximum expected feasible demand for i in $C \cup \{i\}$ (assuming d_i^{max} is an integral multiple of δ above—we can always choose a δ for the $D_i(C, \alpha)$ interval so that this is the case).

Such a proposal is accepted only if all members of affected coalition are *no worse off in expectation* (with respect to their own beliefs). We also impose the following tie-breaking rule for proposals:

Assumption 2 (Tie-breaking rule for the BR process). *If there are several maximum expected to be feasible proposals for proposer i, i chooses among them with equal probability, if each of these proposals induces a coalitional agreement with a value greater than the $\langle S_i, \boldsymbol{d}_{S_i}, \beta\rangle$ currently in place for i, and in which i's demand is d_i; however, if the if these proposals induce agreements that have the same value to i as the $\langle S_i, \boldsymbol{d}_{S_i}, \beta\rangle$ already in place (i.e., if $\langle S_i, d_i, \beta\rangle$ is one of the maximum expected to be feasible proposals), then i will propose $\langle S_i, d_i, \beta\rangle$ to reach this same agreement $\langle S_i, \boldsymbol{d}_{S_i}, \beta\rangle$ again.*

Thus, in the case of ties there is a preference for staying in the same coalition if the maximum expected to be feasible proposals do not strictly improve i's payoff: the proposer only makes a different proposal if it expects it to be strictly better than the agreement currently in place for him. Notice that whenever the proposed agreement coincides with the one currently in effect, the proposal will be trivially accepted (as it was already in effect, which implies that the rest of the members in the coalition do not object to it).

Given the finite strategy space, a discrete-time finite-state Markov chain can be defined, with state space

$$\Omega = \{\omega = (CS, \boldsymbol{d}, \boldsymbol{\alpha}_{CS}) | CS \in \mathcal{SC}, \boldsymbol{\alpha}_{CS} \in \times_{C\in CS} A_C, \boldsymbol{d} \in \times_{i\in N} D_i\}$$

where $\mathcal{SC}$ is the space of all possible coalition structures. Let $S'(i)$ denote the coalition agent i belongs to in any state ω'. We can formally define the Markov chain that corresponds to the BR process as follows:

Definition 11 (Markov chain corresponding to the BR process). *The BR process corresponds to a Markov chain M with state space Ω and transition function $\mathcal{P}$. Let $S(i)$ denote the coalition player i belongs to in any state ω. Then, the probability of a transition from state ω*

to (at time step t) to state ω' (at time step $t+1$) with demand d_i' and coalition $S'(i)$ for i,[22] *with $S'(i)$ having agreed on action α, is given as*

$$\mathcal{P}_{\omega\omega'} = \sum_{i \in N} \gamma \beta_i(\omega'|\omega),$$

where β_i is defined by the best-reply rule as follows:

(i) $\beta_i(\omega'|\omega) = 1/k_\omega^i$*, if $S'(i) = C \cup \{i\}$ and $d_i' = d_i^{max}(C, \alpha)$, with C and α being maximizers of Eq. (4.13), $C \in CS \cup \{\emptyset\}$ with $CS(\omega)$ the coalition structure in place at state ω, ties resolved as in the rule above with k_ω^i being the number of potential proposals* actually considered *at state ω as expected payoff maximizers by i*[23]*, and the proposal to join $S'(i)$ with d_i', α being indeed feasible (i.e., is to be accepted*[24]*) given the private beliefs of $S'(i)$'s members;*

(ii) $\beta_i(\omega'|\omega) = 1$*, if $S'(i) = C \cup \{i\}$ and $d_i' = d_i^{max}(C, \alpha)$, with C and α being maximizers of Eq. (4.13), $C \in CS \cup \{\emptyset\}$ with $CS(\omega)$ the coalition structure in place at state ω, ties resolved as in the rule above*[25]*, but the proposal to join $S'(i)$ with d_i', α* not *being indeed feasible (i.e., it is to be rejected) given the private beliefs of $S'(i)$'s members, and $\omega' = \omega$; and*

(iii) $\beta_i(\omega'|\omega) = 0$ *otherwise.*

In order to discuss stability, we have to present the concept of *ergodic sets* in a Markov chain:

Definition 12 ([KS76]). *A set $E \subset \Omega$ is ergodic if for any $\omega \in E$, $\omega' \notin E$, $\mathcal{P}_{\omega\omega'} = 0$ and no*

[22] A successor state ω' that is different from ω is characterized by *a specific* i's demand d_i' being the only potential change in the demand vector, when joining a specific $S'(i)$ performing α. Thus, a more appropriate notation for ω' could perhaps have been $\omega'(d_i', S'(i), \alpha)$, but we do not use this for simplicity and generality. Of course, with $\mathcal{P}$ being a transition function, the sum of transition pobabilities over all successor states is 1.

[23] Notice that in accordance to the tie breaking rules, if there is more than one proposal maximizing i's expected payoff and one of them coincides with the configuration currently in effect for i, then i will propose exactly this one (not considering the rest) and thus $k_\omega^i = 1$ and $\omega' = \omega$.

[24] For one specific proposal of some agent i, there exists only one possible transition state ω'; there might have been several best replies, but for each there exists only one possible transition state ω'. Further, notice that given the rules of the process, the agent states his own demand only (and not the demands of others), so a proposal $\langle S'(i), d_i', \alpha \rangle$ through which a specific ω' can be reached from a specific ω at t is a proposal of a specific agent only. Thus, if $\omega' \neq \omega$, then $\beta_i(\omega'|\omega)$ is non-zero for *at most one* i in $\sum_{i \in N} \gamma \beta_i(\omega'|\omega)$ above.

[25] In this case *(ii)*, however, none of the expected payoff maximizing proposals coincides with the configuration currently in effect for i, so he makes a distinctly different proposal (one of the many possible maximizers, perhaps), and it is rejected.

non-empty proper subset of E has this property. Singleton ergodic sets are called absorbing states—$\omega \in \Omega$ is absorbing if $\mathcal{P}_{\omega\omega} = 1$.

Once the process has entered an ergodic set, it will remain in that set forever.

A stable coalition structure will result if the bargaining process defined above reaches an absorbing state, where no agent has an incentive to propose a different coalitional agreement given its beliefs and the existing coalition structure. The following lemma characterizes the absorbing states of the best-reply process:

Lemma 1. *Let C_i represent the coalition in which agent i belongs to in a state ω of the BR process, and let α_{C_i} be its adopted action. State $\omega = (CS, \boldsymbol{d}, \boldsymbol{\alpha}_{CS})$ is an absorbing state of the BR process if and only if the following condition is met:*

- *For all $i \in N$, for any proposal $\langle S \cup \{i\}, d_i, \beta\rangle$ of agent i such that $S \in CS \cup \{\emptyset\}$, $\beta \in A_{S\cup\{i\}}$, and $d_i \in D_i(S, \beta)$* are maximizers of Eq. (4.13)*, either of the following conditions (or both) hold:*

  (i) $\bar{p}_i^i(S \cup \{i\}, \boldsymbol{d}_S \circ d_i, \beta) \leq \bar{p}_i^i(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$

  (ii) $\exists j \in S$ *such that* $\bar{p}_j^j(S \cup \{i\}, \boldsymbol{d}_S \circ d_i, \beta) < \bar{p}_j^j(S, \boldsymbol{d}_S, \alpha_S)$

Proof: First we prove the "if" direction.

The condition guarantees that either

(i) the proposing agent i does not have the incentive to switch coalitions or propose a different allocation to his current coalition (in one step of the process), because he does not (realistically) expect this to be profitable, therefore he will not make such a proposal: Consider a maximum expected to be feasible (in one step) proposal $\langle S\cup\{i\}, d_i, \beta\rangle$ (with S, d_i, β being maximizers of Eq. (4.13)), and being such that $\bar{p}_i^i(S \cup \{i\}, \boldsymbol{d}_S \circ d_i, \beta) \leq \bar{p}_i^i(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$ holds. Given the tie-breaking rule above, even if $\bar{p}_i^i(S\cup\{i\}, \boldsymbol{d}_S\circ d_i, \beta) = \bar{p}_i^i(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$, i will still make the proposal $\langle C_i, d_i, \alpha_{C_i}\rangle$ which will be trivially accepted (since it was already in effect, and therefore is acceptable) and thus the state ω will not left. (We remind also that the demands during the BR process are made subject to expected feasibility: in one step, only the d_i^{max} demands are ever stated. Because of this, for any possible $\langle S\cup\{i\}, d_i, \beta\rangle$ proposal to $j \in S$ stated by i, it is always guaranteed that i expects the proposal to be feasible—i.e., that $\bar{p}_j^i(S\cup\{i\}, \boldsymbol{d}_S\circ d_i, \beta) \geq \bar{p}_j^i(C_j, \boldsymbol{d}_{C_j}, \alpha_{C_i})$.)

(ii) *or* if i has such an incentive (expecting his proposal to be profitable, and also feasible—the latter is guaranteed by the process, as explained above), then it is the case that his proposal $\langle S \cup \{i\}, d_i, \beta \rangle$ will be denied, by condition (ii), because there always exists some agent j who believes that he is worse off by accepting, since $\bar{p}_j^j(S \cup \{i\}, \boldsymbol{d}_S \circ d_i, \beta) < \bar{p}_j^j(S, \boldsymbol{d}_S, \alpha_S)$. Therefore, ω will not be left.

Therefore, if for any $i \in N$ and any maximal expected to be feasible proposals $\langle S \cup \{i\}, d_i, \beta \rangle$ stated by i either of conditions (i) or (ii) hold, then state ω will not be left, and thus it is absorbing.

Now we prove the opposite direction: Say that ω is absorbing, but *neither* of conditions (i) or (ii) hold. That is, $\exists i$ and (maximal expected to be feasible) proposal $\langle S \cup \{i\}, d_i, \beta \rangle$ s.t. $\bar{p}_i^i(S \cup \{i\}, \boldsymbol{d}_S \circ d_i, \beta) > \bar{p}_i^i(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$ **and** $\nexists j \in S$, s.t. $\bar{p}_j^j(S \cup \{i\}, \boldsymbol{d}_S \circ d_i, \beta) < \bar{p}_j^j(S, \boldsymbol{d}_S, \alpha_S)$ Then, the BR process guarantees that i will propose $\langle S \cup \{i\}, d_i, \beta \rangle$ since it gives him the maximum expected to be feasible (at this step) payoff, and every j in S will accept the proposal (since no j believes that it makes him worse off). Thus, ω will be left, therefore it is *not* absorbing, and we reached a contradiction.

Thus, we proved both directions of the lemma. □

Notice that the lemma's condition implicitly guarantees that an agent is getting at least his reservation payoff in an absorbing state (if not, condition (i) does not hold—the agent can always form a singleton coalition to guarantee his reservation value). In addition, notice that the lemma's condition is not only necessary but also sufficient for state ω to be absorbing: in state ω, either no agent will put forward a different proposal, or if he does his proposal will be denied.

We now make the following observation:

Observation 6. *If the coalition structure-demand vector-action vector triplet $\langle CS, \boldsymbol{d}, \boldsymbol{\alpha}_{CS} \rangle$ is in the strong or the strict Bayesian core, the state described by this configuration must be an absorbing state of the BR process.*

Proof: Let $\omega = \langle CS, \boldsymbol{d}, \boldsymbol{\alpha} \rangle$ be a state in the strong Bayesian core, and let C_i be the coalition in which i is a member in ω, with α_{C_i} its adopted action. Since ω is in the strong BC, there is no coalition $K \subseteq N$, demand vector $\boldsymbol{d}_K$ and action $\beta_K \in A_K$ s.t. for some $i \in K$ $\{\bar{p}_i^i(K, \boldsymbol{d}_K, \beta_K) > \bar{p}_i^i(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$ *and* $\bar{p}_j^i(K, \boldsymbol{d}_K, \beta_K) \geq \bar{p}_j^i(C_j, \boldsymbol{d}_{C_j}, \alpha_{C_j})$ $\forall j \in K$, $j \neq i\}$. Therefore, for any $i \in N$, there is no expected to be feasible for i proposal $\langle S \cup \{i\}, d_i, \beta \rangle$ (and thus, not even a maximal expected to be feasible $\langle S \cup \{i\}, d_i, \beta \rangle$—i.e., one with S, β, d_i maximizers of Eq. (4.13)) such that $\bar{p}_i^i(S \cup \{i\}, \boldsymbol{d}_S \circ d_i, \beta) > \bar{p}_i^i(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$. Therefore, for

any i and any maximal expected to be feasible proposal $\langle S \cup \{i\}, d_i, \beta \rangle$, condition (i) of lemma 1 is satisfied, and thus ω is absorbing.

Now, let ω be a state in the strict BC. Since it is a state in the strict BC, there is no coalition $K \subseteq N$, demand vector $\boldsymbol{d}_K$ and action $\beta_K \in A_S$ s.t., for some $i \in K$, $\{\bar{p}_i^i(K, \boldsymbol{d}_K, \beta_K) > \bar{p}_i^i(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$ and $\bar{p}_j^j(K, \boldsymbol{d}_K, \beta_K) \geq \bar{p}_j^j(C_j, \boldsymbol{d}_{C_j}, \alpha_{C_j}) \; \forall j \in K, \; j \neq i\}$. Therefore, for any $i \in N$ and any triplet $\langle S \cup \{i\}, d_i, \beta \rangle$ (either expected to be feasible by i or not), it will be either the case that $\{\bar{p}_i^i(S \cup \{i\}, \boldsymbol{d}_S \circ d_i, \beta) \leq \bar{p}_i^i(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$ or $\exists j \in S \cup \{i\}, \; j \neq i$, s.t. $\bar{p}_j^j(S \cup \{i\}, \boldsymbol{d}_S \circ d_i, \beta) < \bar{p}_j^j(S, \boldsymbol{d}_S, \alpha_S)$. Therefore, for any i and any maximal expected to be feasible proposal $\langle S \cup \{i\}, d_i, \beta \rangle$ that he can make, at least one of the conditions (i), (ii) of lemma 1 is satisfied, and thus ω is absorbing. □

However, the converse of the observation is not true, as a game may have multiple absorbing states, but an empty Bayesian core; or that a Bayesian core state may exist that is unreachable under the rules of the game. Intuitively, it is possible that the process has converged to a state that is absorbing given the rules governing the BR process—such as that an agent may propose only to coalitions in the current CS (or to itself) at each step of the game; nevertheless, the agents may well believe that there exist other better configurations, and it might well be the case that if they find themselves in them they will be in a Bayesian core state—however, these better configurations are unreachable under the rules of the game. Therefore, being in an absorbing state does not necessarily mean that the process is a Bayesian core state.

We now turn to the problem of guaranteeing convergence to stable, BC states, through dynamic coalition formation. It is not very likely that, under type and action-related uncertainty, one can identify a *decentralized* formation process[26] (such as the one proposed above) that could guarantee convergence to the weak or the strict Bayesian core (if the core is non-empty, of course—otherwise there is no possibility of convergence to it anyway). This is because those stability concepts allow agents that lie in stable core configurations to have private expectations for higher payoffs in other states because they are not taking into account the expected feasibility of their proposals, as discussed in Section 4.3.

To elaborate on this, what we would like is any decentralized process to be able to proceed and eventually converge to a core state (if it exists) based on the beliefs of each agent *separately*, and this cannot happen if the agents state demands without examining the overall degree of acceptability for the future configurations. For example, as long as some agent be-

[26]However, as we shall see in Chapter 5, one can identify *equilibria* that lead to the strict Bayesian core.

lieves that he can gain from some future configuration X, he can always try to destabilize a configuration by proposing something suboptimal to himself, in hopes of eventually reaching such a configuration. However, if he keeps doing so without taking into account whether the future configuration X is indeed desirable by others, convergence to a BC state is difficult to guarantee (even if such a commonly desirable "core" configuration does exist—but it may not be X).

The only way to guarantee convergence to such a strict or weak BC state is to adopt a randomized proposals protocol that furthemore would have required the agents to perform a test to see if convergence to a strict or weak BC state has been achieved, in order to stop making further proposals. Unfortunately, this would have required them to have access to the private beliefs of others, since the strict and weak BC stability conditions depend on the private expectations of the agents regarding their own payoffs. Thus, this process couldn't possibly be decentralized. It could be possible, however, that we identify such a converging process for the *strong* Bayesian core (which, as we saw, describes a more "endogenous" form of stability). However, the BR process is not a process that could be guaranteed to converge to the strong BC if it is non-empty:

To demonstrate this, suppose that the BR process has entered an absorbing state that is not in the strong Bayesian core. Then, at least one agent will believe that there exists a better (and feasible) configuration in a coalition that is not a member of the current CS. Unfortunately, this cannot be reached by using the best-reply rule, given that the state is an absorbing state of the BR process (where the agents are already receiving their maximum possible expected payoff under the current CS or any CS' reachable by the best-reply rule—see Lemma 1—and they are forbidden to adopt a suboptimal strategy in order to destabilize the prevailing CS).[27]

In order to overcome this obstacle, we consider a slight modification of the BR 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 (i.e., its expected value is such that it can provide him with strictly better expected payoff—we elaborate in the definition below), but which may *not* be derived starting from the existing CS, it can propose an *arbitrarily small* feasible demand in order to destabilize the current state in hopes of reaching a better structure.

The BRE process advocates that the best reply is chosen with probability $1 - \epsilon$, while some other proposal from the space of proposals expected to be feasible for i is chosen with

[27] Notice that we can use similar arguments to show that the BR process cannot guarantee convergence to the strict or the weak BC.

probability ϵ. Specifically, such another feasible proposal $\langle C', d_i, \alpha \rangle$ is chosen with probability

$$\epsilon/|\Sigma_i^{BRE}(\omega)| = \epsilon \cdot (1/|\mathcal{C}^i|) \cdot (1/A_{C'}) \cdot (1/\lceil \frac{d_i^{max}(C,\alpha)}{\delta} \rceil)$$

where $\mathcal{C}^i$ denotes the set of all $C \cup \{i\}$ (with $C \in CS \cup \{\emptyset\}$) coalitions in which i can participate, and $\Sigma_i^{BRE}(\omega)$ denotes i's strategy set for ω:

$$\Sigma_i^{BRE}(\omega) := \{(C_i, d_i, \alpha) | C_i = C \cup \{i\}, C \in CS \cup \{\emptyset\}, \alpha \in A_{C_i}, d_i \in D_i^{BRE}(C, \alpha)\}$$

where $D_i^{BRE}(C, \alpha)$ is the finite set of all integral multiples of δ in the interval $[0, d_i^{max}(C, \alpha)]$.

This can be viewed as explicit experimentation on behalf of the agents. Furthermore, any agent j that is part of a proposed-to coalition will choose to accept a demand from i that lowers its payoff with probability ϵ iff j believes there exists some potentially more rewarding coalition S and corresponding action α_S, with $i, j \in S$, such that there is hope for improving his expected payoff without harming the other members of S (i.e., $Q_j(S, \alpha_S) > \sum_{k \in S} \bar{p}_k^j(C_{CS}^k, \boldsymbol{d}_{CS}^k, \alpha_{CS}^k)$, where C_{CS}^k is the coalition of k in CS and $\boldsymbol{d}_{CS}^k$, α_{CS}^k its corresponding demand vector and selected action).

Equivalently, we can provide the following definition for the BRE process:

Definition 13 (Best Reply Process with Experimentation (BRE process)). *: In any state $\omega = (CS, \boldsymbol{d}_{CS}, \boldsymbol{\alpha}_{CS})$, let C_{CS}^k denote the coalition in which agent k belongs to in CS, and let $\boldsymbol{d}_{CS}^k, \alpha_{CS}^k$ be its respective demand vector and selected action. The BRE proceeds like the BR process, with the following differences:*

Whenever there exists a coalition $S' \notin CS$ with some action $\alpha_{S'}$ such that $Q_i(S', \alpha_{S'}) > \sum_{k \in S'} \bar{p}_k^i(C_{CS}^k, \boldsymbol{d}_{CS}^k, \alpha_{CS}^k)$, each agent i that believes it can be a member in such an S' chooses a "best reply" with probability $1 - \epsilon$, and takes each strategy $(S_i, d_i, \alpha) \in \Sigma_i^{BRE}(\omega)$ with probability $\epsilon/|\Sigma_i^{BRE}(\omega)|$ when he gets the opportunity to revise his strategy.

In addition, an agent j who has to reply to some proposer's i proposal, and that believes that an $S' \notin CS$ exists, *such that $Q_j(S', \alpha_{S'}) > \sum_{k \in S'} \bar{p}_k^j(C_{CS}^k, \boldsymbol{d}_{CS}^k, \alpha_{CS}^k)$, with $j \in S', i \in S'$ chooses to accept only if he is not worse off in expectation (based on his beliefs) with probability $1 - \epsilon$, and chooses to (unconditionally) "accept"[28] with (some small) probability ϵ.*

The BRE process has some reasonable properties. First we note that the absorbing states

[28]Notice that we allow for "random unconditional acceptance" *only* if the agent believes that *both* itself and the proposer can be members of the S' coalition. This will be enough to ensure convergence to the strong BC if it is not empty, in Proposition 4 that follows.

of the process coincide with strong Bayesian core allocations.

Proposition 3. *The set of demand vectors associated with an absorbing state of the BRE process coincides with the set of strong Bayesian core allocations. Specifically,* $\omega = \langle CS, \boldsymbol{d}_{CS}, \boldsymbol{\alpha}_{CS} \rangle$ *is an absorbing state of the BRE process iff* $\langle CS, \boldsymbol{d}_{CS}, \boldsymbol{\alpha}_{CS} \rangle \in$ *strong* BC.

Proof: If a state ω is in the strong BC, no agent believes that he can gain either by switching coalitions or by changing his demand. That is, no agent i believes that there exists S', $\boldsymbol{d}_{S'}$ and $\beta \in A_{S'}$ s.t. $\bar{p}^i_i(S', \boldsymbol{d}_{S'}, \beta) > \bar{p}^i_i(C^i_{CS}, \boldsymbol{d}^i_{CS}, \alpha^i_{CS})$ and $\bar{p}^i_j(S', \boldsymbol{d}_{S'}, \beta) \geq \bar{p}^i_j(C^j_{CS}, \boldsymbol{d}^j_{CS}, \alpha^j_{CS})$, which implies that no agent i believes that there exists a coalition S' and action β such that $Q_i(S', \beta) > \sum_{k \in S'} \bar{p}^i_k(C^k_{CS}, \boldsymbol{d}^k_{CS}, \alpha^k_{CS})$. If the latter was the case, then there could have always been an allocation in S' such that, according to i's beliefs, i is strictly better off in expectation and the others weakly better off. Therefore, no proposing agent experiments and no responding agent accepts a proposal unconditionally: the process follows the BR rules, and any proposing agent proposes the agreement already in place (given the tie-breaking rule, and the fact that this agreement is one of the agreements maximizing his expected feasible payoff, as there is no other agreement s.t. $\bar{p}^i_i(S', \boldsymbol{d}_{S'}, \beta) > \bar{p}^i_i(C^i_{CS}, \boldsymbol{d}^i_{CS}, \alpha^i_{CS})$), and the agreement is trivially accepted by all responders (it was already in place, therefore it is an acceptable agreement, as it does not make the responders worse-off), and thus state ω is never left, so it is absorbing.

Suppose now that $\omega = (CS, \vec{d}_{CS}, \boldsymbol{\alpha}_{CS})$ is an absorbing state of the BRE process that is not in the strong BC. Since it is not in the strong BC, then there is an i that believes there exists S', $\boldsymbol{d}_{S'}$ and $\beta \in A_{S'}$ s.t. $\bar{p}^i_i(S', \boldsymbol{d}_{S'}, \beta) > \bar{p}^i_i(C^i_{CS}, \boldsymbol{d}^i_{CS}, \alpha^i_{CS})$ and $\bar{p}^i_j(S', \boldsymbol{d}_{S'}, \beta) \geq \bar{p}^i_j(C^j_{CS}, \boldsymbol{d}^j_{CS}, \alpha^j_{CS})$, which means that i believes that $Q_i(S', \beta) > \sum_{k \in S'} \bar{p}^i_k(C^k_{CS}, \boldsymbol{d}^k_{CS}, \alpha^k_{CS})$. Consequently, with probability ϵ, at least i will experiment, potentially asking for zero payoff[29] (having the opportunity to form a singleton, subject to no-one else's acceptance). Thus, there exists a positive probability that ω will be left, and therefore ω cannot be absorbing (contradiction). □

Proposition 3 does not guarantee that a BC allocation will actually be reached by the BRE

[29] Asking for less than i's reservation value is allowed in order to cover the possibility that the non-BC absorbing state already contained i in a singleton coalition, having a demand equal to his reservation value. In contrast to what is true for the deterministic full information case [DS98], such an absorbing state may exist, since the demand of i may be such that is unacceptable to others to form a coalition with him (i.e., it's not certain that if i believes that there exists a better S, others in S will agree to form a coalition with i given the currently observed demand vectors; therefore, for i to be certain to destabilize the current structure effectively, it has to be allowed to ask for less than its reservation value, thus "luring" others into forming a coalition with him in the future). If this was not possible, then a non-BC state might not be ever left—being absorbing without being in the strong BC.

process. However, we guarantee the following:

Proposition 4. *If the strong Bayesian core is non-empty, the BRE process will converge to an absorbing state with probability one.*

Proof: The proof is analogous to the proof for the deterministic coalition formation model [DS98]. The basic idea is that when the strong BC is not empty, all ergodic sets reached by the BRE process are singletons, therefore the BRE process will converge to an absorbing state.

Thus, we shall take the following steps to show that all ergodic sets are singletons. Suppose that there exists an ergodic set $E \subset \Omega$ with $|E| \geq 2$. We will establish a contradiction by showing that E contains a state from which there is a path to an absorbing state.

Proposition 3 ensures that none of the states in E involve a strong BC[30] configuration. This follows from the fact that all strong BC configurations are absorbing states, and ergodic sets are minimal—they cannot contain other ergodic sets. As a consequence, for each state $\omega = (CS, \boldsymbol{d}^E, \boldsymbol{\alpha}^E) \in E$ there exists at least one agent i that believes that $\exists S \ni i$, with $S \notin CS$, some demand vector $\boldsymbol{d}_S$, and some action $\alpha_S \in A_S$ s.t. $\bar{p}_i^i(S, \boldsymbol{d}_S, \alpha_S) > \bar{p}_i^i(C_{CS}^i, \boldsymbol{d}_{CS}^i, \alpha_{CS}^i)$ and $\bar{p}_j^i(S, \boldsymbol{d}_S, \alpha_S) \geq \bar{p}_j^i(C_{CS}^j, \boldsymbol{d}_{CS}^j, \alpha_{CS}^j)$, where C_{CS}^i, C_{CS}^j are i and j's coalitions in CS, and $\boldsymbol{d}_{CS}^i$, $\boldsymbol{d}_{CS}^j$, α_{CS}^i and α_{CS}^j are their respective demand vectors and selected actions.

So, there exists at least one agent i[31] whose beliefs are such that they imply $Q_i(S, \alpha_S) > \sum_{k \in S} \bar{p}_k^i(C_{CS}^k, \boldsymbol{d}_{CS}^k, \alpha_{CS}^k)$. Therefore, any such agent i will experiment with suboptimal strategies if he gets a chance to be the proposer. There is a positive probability that any such agent gets the chance to propose. Moreover, there is a positive probability that all those agents who experiment form the singleton coalition and demand zero payoff, picking the strategy $\langle \{i\}, 0, \alpha_i \rangle$—where α_i is some action available to i's singleton coalition. Thus, all states that can be defined from any $\omega \in E$ by replacing coalition $S(i)$—the coalition containing i—with $S \setminus \{i\}$ and adding $\{i\}$ in CS, and replacing d_i^E with 0 for all agents i who experiment with the singleton coalition can be reached with positive probability from ω. Denote the set of all such states by $Reach(\omega)$. It follows that $\bigcup_{\omega \in E} Reach(\omega) \subset E$, since all states in that set are reached with positive probability.

By the same argument as above, elements of $Reach(\omega)$ cannot involve BC allocations. Repeating the same procedure as above, replacing E by $Reach(\omega)$ for each $\omega \in E$ in the argument, we get a set $Reach^2(\omega)$ for each $\omega \in E$. Again, $\bigcup_{\omega \in E} Reach^2(\omega) \subset E$.

[30] Henceforth in this chapter, whenever we refer to "BC", we actually refer to the strong BC.

[31] If there is exactly one such agent, the non-empty BC will contain him as a singleton—otherwise the BC would have to be empty.

Continuing in the same way, after repeating this procedure a finite k number of times, the set $\bigcup_{\omega \in E} Reach^k(\omega)$ contains the state where either (a) each agent forms the singleton coalition or (b) some agents form singleton coalitions, and those who do not form singleton coalitions are playing best replies and do not experiment. More precisely, E contains a state $\omega' = (CS', \boldsymbol{d}', \boldsymbol{\alpha}')$ with the following property: Either $S'(i) = \{i\}$ and $d'_i = 0$ for all $i \in N$, or, if there are coalitions $S' \in CS'$ with $|S'| \geq 2$, then there exists an absorbing state $\omega^{BC} = (CS^{BC}, \boldsymbol{d}^{BC}, \boldsymbol{\alpha}^{BC})$ such that $S' = S^{BC}$ for some $S^{BC} \in CS^{BC}$, $\alpha_{S'} = \alpha_{S^{BC}}$ and $d'_i = d_i^{BC}$ for all $i \in S'$, for all S' with $|S'| \geq 2$. (As the BC is non-empty, an absorbing state exists. Further, as members of S' play best replies and do not experiment, their demands must be part of a BC allocation.)

Starting from ω', an absorbing state $\omega^{BC} = (CS^{BC}, \boldsymbol{d}^{BC}, \boldsymbol{\alpha}^{BC})$ will eventually be reached: Assuming that the reservation value of each agent is greater than zero, we establish that in state ω', each agent who forms a singleton coalition demanding zero payoff believes that there exists a potentially better coalition; thus, each agent i with $S(i) = \{i\}$ experiments with probability ϵ when it is his chance to move. Further, all agents who are in non-singleton coalitions in ω' do not experiment. Let $T^{BC} \in CS^{BC}$ represent a coalition in which several agents that are "single" in ω' will eventually find themselves into when in CS^{BC}. For each $T^{BC} \in CS^{BC}$, all agents that will belong in T^{BC}, progressively join each other starting from ω', making such demands (and proposing such coalitional actions) that their proposals are accepted and that the demands vector resulting eventually will coincide with the BC demands' vector—$\forall T^{BC} \in CS^{BC}, \forall j \in T^{BC} : d_j = d_j^{BC}$. There exists a positive probability that this will occur, since there exists a positive probability for the acceptance of each proposal that leads to an existing BC—given that agents can propose to join coalitions demanding suboptimal payoffs, and given that"random unconditional acceptance" is possible under the best-reply with experimentation process, for agents that believe there exists a potentially better (and expected to be feasible) coalition not in the current CS.[32]

Therefore, there exists a positive probability that the absorbing state ω^{BC} is reached when starting from ω'. This is a contradiction to ω' being an element of the ergodic set E. It follows that all ergodic sets reached by the best-reply with experimentation process are singletons, when the strong BC is non-empty. This completes the proof. □

Propositions 3 and 4 together ensure that if the strong BC is not empty then the BRE process

[32]It may be possible to remove the concept of "random unconditional acceptance" from the BRE process. In that case, however, one should probably employ some concept of superadditivity for the agents' beliefs.

will eventually reach a strong BC allocation, no matter what the initial coalition structure.

4.6 Some Simple Experiments

We report here on some simple experiments conducted in order to test the validity of our approach and verify the BRE's (and the BR's) convergence behaviour towards the strong BC empirically.

First, we examine the BR and BRE process in several "standard" coalitional games. The agents share common and correct beliefs about the coalitional values to mimic a standard characteristic function game (and there is only one action at hand for the agents).

One of them is the 3-player majority game [Mye91, DS98], where any majority of the 3 players (i.e., any pair of players) can get a coalitional payoff equal to that of the grand coalition. Specifically, in this game, the set of agents is $N = \{a, b, c\}$, $rv_i = 0$ for all $i \in N$, and $v(S) = 10$ for $|S| \geq 2$. The core of the game is empty, because no matter what deal two of the players have striken between them, the third player, say b, can always lure one of them, say c, into a deal that can be at least equally profitable for c (and therefore, the bargaining process could last for ever, if unlimited time is supposed). However, any state $\omega = \langle\{N\}, \boldsymbol{d}\rangle$ with $\sum_{i \in N} d_i = 10$—that is, any state ω in which the grand coalition has formed—is absorbing for the BR process (this is because if the grand coalition has formed, no agent can reach a coalition that gives him a strictly better payoff in just one step, since there is no such coalition to which it can propose and it will not propose to itself even if he gets 0 in the grand coalition, because of the tie-breaking rule of the BR process). Nevertheless, any such ω is unreachable for the BR process if it is started in a state where any of the players (in a 2-agent coalition) receives the entire payoff of 10. If that is the case, the process should exhibit non-singleton ergodic sets, with the players constantly changing allegiances and forming 2-agent coalitions with eachother.

The findings of our experiments confirmed these hypotheses. Specifically, in our experiment (which consisted of 30 runs of 1000 negotiation rounds each) we started the BR process in state $\langle CS = \{\langle a, b\rangle, \langle c\rangle\}, \boldsymbol{d} = \langle d_a = 10, d_b = 0, d_c = 0\rangle\rangle$. We observed that neither the BR nor the BRE process ever converged to any kind of stable configuration or to , but, rather, the agents kept regrouping into 2-agent coalitions.

Another game we experimented with was a game presented by [DS98], where $v(S) = 2$ for coalitions of size $|S| = 1$, $v(S) = 5$ for $|S| = 2$ and $v(S) = 8$ for $|S| = 3$. The set of agents in the game is $N = \{a, b, c\}$. We started the process in state $\langle CS = \{\langle a, b, c\rangle\}, \boldsymbol{d} = \langle 4, 2, 2\rangle\rangle$

which is absorbing for the best reply (BR) process, but is not a core (or BC) configuration, since agents b and c would be better off in coalition $\langle b, c\rangle$, which is unreachable given the rules of the BR process. Unsurprisingly, the BR process never left the absorbing state, while the BRE process *always* (that is, in 30/30 runs, each consisting of 1000 negotiation steps) converges to a configuration in the BC (i.e., in some configuration where the grand coalition forms, with two of its members receiving a payoff of 3 and the third receiving a payoff of 2). In 19 of those runs, the process converged in fewer than 50 rounds (typically less than 25).

We also tested a game having a non-empty BC, in which the beliefs of the three participating agents a, b and c *differed*, and each coalition had three actions (α, β and γ) available to it. The agents' beliefs regarding the Q-values of coalition-action pairs are as shown in table B.1 in Appendix B. Typical configurations in the Bayesian core for this game include $\langle\{\langle a, b\rangle, \langle c\rangle\}, \{\boldsymbol{r}_{ab} = \langle 79\%, 21\%\rangle, \langle \boldsymbol{r}_c = 100\%\rangle\}, \{a_{ab} = \alpha, a_c = \alpha\}\rangle$, $\langle\{\langle a, b\rangle, \langle c\rangle\}, \{\boldsymbol{r}_{ab} = \langle 75\%, 25\%\rangle, \langle \boldsymbol{r}_c = 100\%\rangle\}, \{a_{ab} = \alpha, a_c = \alpha\}\rangle$, and others like these—that is, agents a and b form coalition $\langle a, b\rangle$ (and perform action α), with a getting the biggest share (around 75%) of the payoff, and c is in a singleton (performing action α also). The initial coalition structure in our experiments was $\{\langle a, c\rangle, \langle b\rangle\}$, with both coalitions in the structure performing action γ, and the initial payoff allocation (actual demand allocation) was given as $\langle d_a = 250, d_b = 200, d_c = 350\rangle$.

The BRE process managed to reach a BC configuration in *all* 30 runs tested (with 1000 bargaining rounds each). The greatest number of negotiation rounds to convergence was 292. However, a BC configuration was typically reached in less than 100 rounds. In 5 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 (typically the $\{\langle a, b\rangle, \langle c\rangle\}$ structure), but not always with an appropriate payoff allocation.

For interest, in order to further support the rational behind the BRE and the BR process, we also tested a different formation process in this same setting (i.e., the 3-agent, 3-actions setting with beliefs that differ and expected values as in table B.1 in Appendix B). This new process is a *randomized* process which assumes that the agents *always* make *random* proposals if they believe that there exists a better configuration (i.e., proposals are as in the BRE, but with $\epsilon = 1$), and always unconditionally accept a proposal. Intuitively, this process allows the agents to make and accept random proposals as long as the Bayesian core has not been formed. Even though this process should allow the eventual creation of the Bayesian core (if it exists), it is clear that this would require a great number of bargaining rounds. Our results confirm

this hypothesis, even for this small problem: in 30 runs (each consisting of 1000 iterations), all of them converged to a BC configuration, but this typically required more than 200 iterations in comparison to BRE's fewer than 100 iterations (to be more specific, the average number of iterations to convergence for the completely randomized process was 270, with only 9 runs converging in less than 100 rounds).

As a final note, running time was not an issue with these small experiments. One experiment with 1000 negotiation rounds runs just within 0.06 seconds.[33] These processes provide an automated way to form stable coalitions, without the need of any human intervention.

4.7 Conclusions

In this chapter, we mainly dealt with the question of coalitional stability under uncertainty. To do so, we presented a Bayesian coalition formation model, the first ever to tackle type uncertainty; the model also accomodates action-related uncertainty, as well. We introduced the concept of the Bayesian core, and defined three versions of it, each corresponding to a different notion of stability. We dealt with the question of the non-emptiness of the Bayesian core, and provided an algorithm to decide the problem in small games. Then, we linked the stability question and the formation question by presenting two dynamic coalition formation processes, one of which was shown, both theoretically and empirically, to converge to the strong Bayesian core if it is non-empty.

Our framework and algorithms enable the agents to reach configurations that are stable, *given their beliefs*. We believe that this is a natural, reasonable way to approach the question of coalition formation and stability under uncertainty.

[33] In addition, one would expect that these processes can be decomposed and be run in parallel, with each agent residing on a different machine.

Chapter 5

Coalitional Bargaining under Uncertainty

In this chapter we provide a *Bayesian non-cooperative approach to coalition formation under (type) uncertainty*. Non-cooperative coalition formation research deals primarily with the *bargaining* process by which coalitions emerge, and the *equilibrium* solution concepts that characterize rational agents' behaviour. *Discounted* coalitional bargaining (e.g., [CDS93, Oka96]) provides a formation setting which emphasizes the need for strategic considerations, and makes the sequential decision making needed during formation more apparent. To the best of our knowledge, there exists no formal study of this problem *under uncertainty*. The introduction, in particular, of *type uncertainty* influences the decisions agents make during negotiations (and the stability of formed coalitions, as discussed in the previous chapter). Furthermore, it presents the opportunity to learn about the capabilities of others based on their behaviour during negotiations and by observing their performance in settings where coalitions form repeatedly.

Although the dynamic coalition formation processes discussed in the previous chapter were non-cooperative, driven by the agents' wish to maximize their individual payoffs, the emphasis was on the processes' properties regarding convergence to stable structures. Since the agents had the common understanding that the bargaining horizon was infinite, and that, importantly, the values to be gained were undiscounted and they would have the chance to participate in all bargaining rounds, the setting posed no emphatic need on studying sequential strategic considerations or equilibrium concepts. This is no longer the case if discounted coalitional bargaining is assumed.

After a review of related work in Section 5.1, in Section 5.2 of this chapter we define a *Bayesian* model for discounted coalitional bargaining under type uncertainty, formally defining the class of *Bayesian coalitional bargaining games (BCBGs)*. We proceed in Section 5.3 to describe the Perfect Bayesian Equilibrium (PBE) solution for coalitional bargaining under type

uncertainty. We then show that computing the PBE is intractable in practice. In that section we also define an equilibrium solution for BCBGs that assumes *fixed beliefs* of the agents during bargaining; this concept is an extension of the SPE (subgame perfect equilibrium) and a restriction of the PBE concept.

To combat PBE intractability, in Section 5.4 we present an algorithm that uses iterative coalition formation to heuristically approximate the equilibrium bargaining behaviour of the agents: the algorithm calculates the agents' strategies in each bargaining round assuming that the beliefs of the agents are to be held fixed for the remainder of the game (and making other relevant assumptions, such as using bounded lookahead when solving the game tree). Critically, however, the algorithm *does* use belief updates following every stage of bargaining (and thus facilitates learning of the potential partners' types). When used in an actual *repeated* coalition formation setting, we can combine this algorithm with RL-style belief updates following the execution of coalitional actions after each episode of coalition formation.

We then proceed in Section 5.5 to provide a *non-cooperative justification of the Bayesian core (BC) stability concept*: in the spirit of what others have done for non-stochastic models, we relate a cooperative stability concept under uncertainty with non-cooperative equilibrium play in (Bayesian) coalitional bargaining games. We achieve this by linking the (non-cooperative) equilibrium solution of the BCBG with the BC (cooperative) solution concept for Bayesian coalition formation, proving results that tie them to each other. Specifically, we prove that if the (strict) BC of a coalitional game is non-empty, then there exists an equilibrium of the corresponding bargaining game that produces a BC element; and also that if there exists a coalitional bargaining equilibrium (with certain properties), then it induces a (weak) BC configuration. In addition, as a corollary to this latter finding, we establish a sufficient condition for the existence of the (weak) Bayesian core.

Then, in Section 5.6 we experimentally evaluate our heuristic algorithm for Bayesian coalitional bargaining, demonstrating its advantages over a method that does not employ Bayesian rationality. As we shall see, Bayesian coalitional bargaining enables the agents to take rewarding, sequentially rational decisions while bargaining. We conclude the chapter by summarizing our findings in Section 5.7.

We note that parts of the research described in this chapter appeared originally in [CB07] and [CMB07].

5.1 Related Work

Much work can be found on coalitional bargaining, both in the economics and the AI literature. However, not much of this work assumes any form of uncertainty, or exploits the opportunity to combine learning with coalitional bargaining.

Okada [Oka96] suggests a form of discounted coalitional bargaining in the usual deterministic characteristic function games setting, where a random proposer chosen out of the set of participating agents puts forward a coalition and payoff allocation proposal at each round. If the proposal is accepted, the coalition abandons negotiations. Okada basically characterizes the subgame-perfect equilibria (SPE) of the discounted coalitional bargaining game he introduces, showing that if they are assumed to be *stationary* and the proposer is chosen *randomly* at each round, then there is no delay of agreement in equilibrium (i.e., any proposal made at round one is agreed upon by the interested agents immediately). The stationarity of the SPE means that the agents choose identical equilibrium strategies in subgames involving the same set of active agents—the proposals and responses of the agents in the t-th round of the game depends only on the set of players active at that round, and not on past history. The equilibrium proposals and responses of the players are given as solutions to a payoff maximization problem, assuming SPE stationarity, and that coalitional values are given and commonly known.

Chatterjee *et al.* [CDS93], unlike Okada, present a discounted bargaining model with a *fixed* proposer order, which results in a delay of agreement. Specifically, the basic result of their work was showing that—unlike what Rubinstein's [Rub82] seminal work on the alternating-offers model of bargaining has shown for two-person bargaining and multi-person unanimous bargaining—delay of agreement *can* occur in multi-person coalitional bargaining. Neither Chatterjee *et al.*'s nor Okada's model deals with type uncertainty or the selection of coalitional actions. Instead, both papers focus on calculating subgame-perfect equilibria. Furthermore, these results hold only in superadditive environments. We make no superadditivity assumption (in fact, no additivity assumption whatsoever) in our work. However, like [Oka96], our bargaining model assumes that the proposer in every round of negotiations is randomly chosen by nature.

Kraus, Wilkenfeld and Zlotkin [KWZ95] looked at the infinite horizon alternating offers model of bargaining where agents take the passage of time into account, and examined a case where agents need to negotiate about sharing a resource, and have incomplete information about each other. Thus, in order to decide on accepting or rejecting offers, the agents update beliefs regarding their opponents' type, given the opponents' response to previous offers. A

finite set of agent types is assumed, each type having a different utility function which depends on its resource usage. However, the negotiation in any given period is *bilateral*, since the assumption is made that no more than two agents need to share the same resource in a period.[1] Thus, there is no true coalitional bargaining by this model.

Two recent papers by Kraus, Shehory and Taase [KST03, KST04] deal with coalition formation in the "Request for Proposal" domain. In this domain, the agents come together to perform tasks comprised of subtasks; each substask should be performed by a different agent. The agents may not know the value of a subtask to another agent—they may not know the costs incured by performing it—but they *do* know the overall payoff associated with performing a task and the capabilities of the other agents, so "incomplete information" has a rather restricted meaning in this work. Also, these papers deal with coalitional-value uncertainty rather than partners' type uncertainty. Moreover, the bargaining algorithm used for formation, even though fitting for their specific framework, is somewhat unrealistic for more generic environments, as it employs an oracle-like agent which eventually decides on the coalitions to be created, making the most fitting (i.e., rewarding) matches among agents—even though this is done based on their suggestions. In addition, [KST03] assumes that all agents in a coalition will divide the gained surplus *equally* among them—an assumption that may violate individual rationality. However, [KST04] presents some other allocation strategies as well: allocation that is proportional to the agents' costs; allocation that is "kernel-stable" (i.e., in the kernel of the coalitional game); or combining these strategies with the use of *compromise* (i.e., the agents are willing to receive less payoff as long as they do get to form the coalition). Still, the focus is on maximizing social welfare rather than satisfying the agents' individual rationality. Nevertheless, later in this chapter, we will be comparing our heuristic algorithm with a (modified) version of their "kernel-stable allocation using compromise" approach. This is because, despite of its deficiencies and the fact that is better tailored to social-welfare maximization settings, it is a rare example of a successfully tested discounted coalitional bargaining method under some restricted form of uncertainty, which combines heuristics with principled game theoretic techniques.

As was mentioned in Section 2.3, in recent years much work ([PR94, MW95, HMC96, Eva97, SV97, Yan03]) has focused on the problem of providing a non-cooperative justification of the (deterministic) core, by establishing results linking the outcomes arising from equilibrium play in coalitional bargaining games to the core of the underlying coalition formation

[1] The agents play two different roles: one of them already has access to the resource and is using it during the negotiation process (gaining over time), while the other one is waiting to use the resource (it is losing over time).

problem. In establishing similar types of results for the Bayesian core, not only do we deal with uncertainty related to agent types (or abilities) and coalitional action effects, we do so without making additivity assumptions w.r.t. coalition value. Our model is thus richer and more realistic than those adopted in most previous work on this problem.

Of that work, the one that is most relevant to ours is that of Moldovanu and Winter [MW95]. They study bargaining games in extensive form, under the usual assumption of deterministic and commonly known coalitional values, but assuming non-transferable rather than transferable utility. The order of proposers in the game depends on the responders' replies: the first responder to refuse a proposal becomes the next *initiator*, and can propose a coalition and a payoff vector (out of a set of available payoff vectors) to its members—or, it can pass the initiative to another player. When all potential members accept a proposal, the formed coalition abandons the game. There is no assumption of any discounting of coalitional values over time. In this bargaining setting, the authors' focus of attention is on stationary, pure-strategy subgame perfect equilibria. They show that if a bargaining strategy profile is an *order independent equilibrium (OIE)* (an SPE that remains an equilibrium and leads to the same payoff allocation for any choice of proposer in a sequential coalitional bargaining game), then the resulting payoffs must be in the core—and conversely, if the coalition formation game has subgames with nonempty cores, then for each payoff vector there exists an OIE with the same payoff. The model of [MW95] differs from ours: it is deterministic, does not assume random proposers, and assumes superadditive, non-transferable utility. However, we do use a form of OIE in our work, generalized to incorporate an agent's (uncertain) beliefs about the abilities (or types) of potential partners and the lack of superadditivity.

5.2 Bayesian Coalitional Bargaining

While coalition structures and allocations can sometimes be computed centrally, in many situations they emerge as the result of some coalitional bargaining process among the agents, who propose, accept and reject partnership agreements. Once again, in order to address type uncertainty (which is inherent in many realistic situations), we assume the Bayesian coalition formation model introduced in Chapter 4 to be in place. Therefore, we assume an underlying Bayesian coalition formation problem. Recall that a BCFP, defined as in Section 4.2, is characterized by a set of agents, a set of types T_i per agent i, coalitional actions A_C per coalition C, outcomes $o \in \mathcal{O}$ corresponding transition dynamics $Pr(o|\alpha, \boldsymbol{t}_C)$, a reward function $R : \mathcal{O} \longrightarrow \Re$, and beliefs μ_i of each agent regarding the types of potential partners. The agents

do not know, but have *beliefs* about the types of their partners, and reason about the value of potential coalitional agreements given their beliefs about types.

We will now introduce a *Bayesian coalitional bargaining game* (or, a *BCBG*) that focuses on the strategic interactions of the rational agents while negotiating to address the BCFP. In other words, this bargaining game provides a non-cooperative view of the BCFP.[2]

The Bargaining Game

When the game starts each agent has knowledge of its own type only (we can assume that *nature*, announces each agent's type only to the agent itself—the type chosen among the types T_i according to a prior which is assumed to be common knowledge). Thus, the type profile $\boldsymbol{t}$ of the agents is specified, but any agent i observes only its own $t_i \in \boldsymbol{t}$. The game then proceeds in stages, with a randomly chosen agent proposing a coalition, a coalitional action and an allocation of payments to partners, who then accept or reject the proposal. The value of any potential coalitional agreement is discounted by $\delta \in (0, 1]$ over time (at each stage); this encourages agreement in earlier rather than later stages.

A set of *bargaining actions* is available to the agents. A bargaining action corresponds to either:

(a) some *proposal* $\pi = \langle C, \boldsymbol{d}_C, \alpha_C \rangle$ to form a coalition C with a specific payoff configuration $\boldsymbol{d}_C$ (specifying payoff shares[3] $d_i \in [0, 1]$ to each $i \in C$) and a suggested coalitional action α_C for C to perform; or

(b) the acceptance or rejection of such a proposal.

We make the assumption that for each coalition there is a finite number of possible demand vectors that one could propose—thus, the set of bargaining actions above is *finite*. Initially, all agents are active (i.e., capable of participating in the negotiations). At the beginning of stage s, one of the (say n) active agents i is chosen randomly with probability $\gamma = \frac{1}{n}$ to make a proposal $\pi = \langle C, \boldsymbol{d}_C, \alpha_C \rangle$ (with $i \in C$). Each other $j \in C$ either accepts or rejects this proposal. If all $j \in C$ accept, the agents in C are made inactive and removed from the game. Value $V_s(t_C) = \delta^{s-1} Q(C, \alpha_C | t_C)$ is realized by C at s, and shared among the agents according to $\boldsymbol{d}_C$. If any $j \in C$ rejects the proposal, the agents remain active (no coalition is formed). At the

[2] Actually, to be more precise, the agents face a different BCFP in each round of the bargaining game we will be defining, since the the agents' beliefs evolve over time.

[3] We use d to denote relative payoff demands in this chapter because we will be using r to denote "rejection" later on.

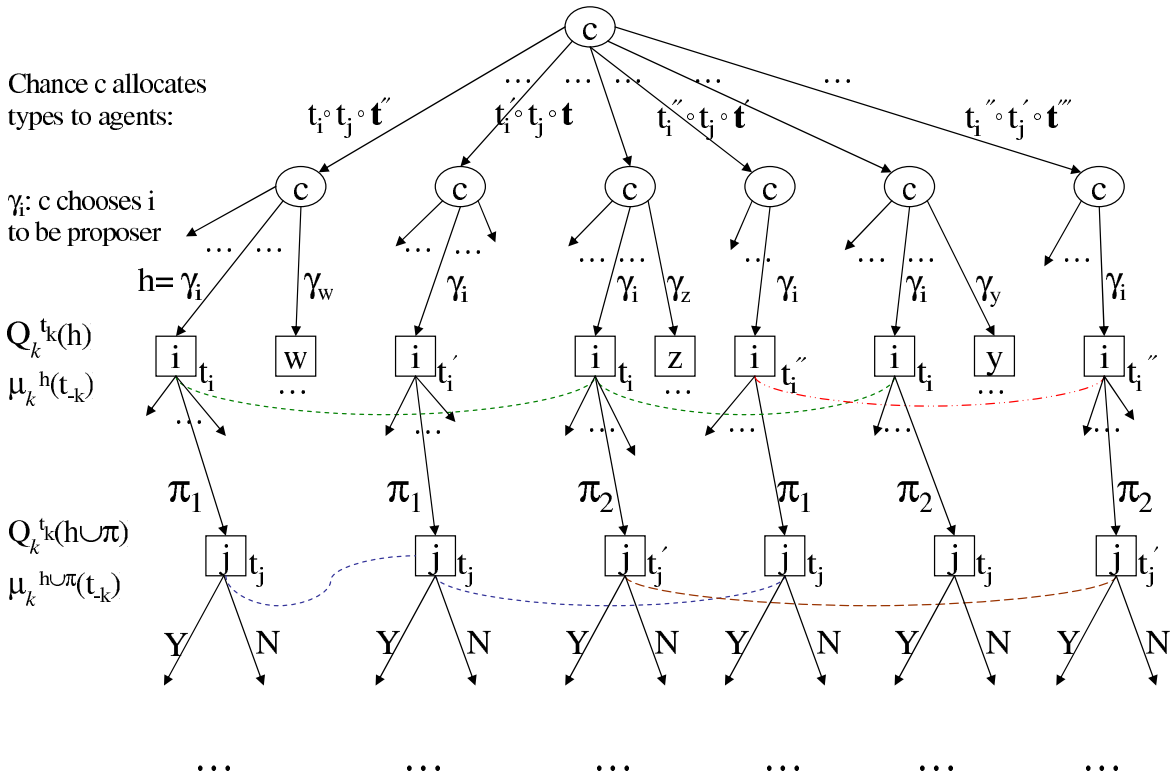


Figure 5.1: A BCBG game. Dashed lines join nodes that belong in the same *information set* (i.e., the agents have observed the same history of moves so far—see Definition 16). After each observed move in the game, any active agent has (updated) beliefs μ and also some (expected) value Q (from the continuation of the game). Actually, the μ^h shown here immediately after the nature's first choice of a proposer in the game correspond to the agents' prior. π denotes a proposal to some C—only the j responder is visible here.

end of a stage, the responses are observed by all participants, and the agents update their beliefs regarding others using Bayes rule. The game is finite-horizon: at the end of the final stage F, any i not in any coalition receives its discounted reservation value $\delta^{F-1} rv_i = \delta^{F-1} V(\{i\}|t_i)$—i.e., discounted singleton coalition value.[4] (If an infinite horizon is assumed, then, due to discounting, any such agent's payoff will be zero in the long run). Figure 5.1 depicts a BCBG game (showing the first levels of the game tree).We will provide a formal definition of the game shortly.

The coalitional bargaining game described above is a *Bayesian extensive form game with*

[4]Henceforth, whenever needed, we will be using $V(t_i)$ instead of $V(\{i\}|t_i)$.

observable actions [OR94] (where the actions correspond to *bargaining* actions as defined above). Briefly, an *extensive form game* is a tree-like representation of a multi-stage game, depicting the sequence of players' actions in the course of the game. The nodes of the tree correspond to decision-making points for the agents, and the branches correspond to players' actions [OR94, MCWG95]. When all players are fully informed regarding every move that has occured in the game, it is a game with *perfect information*. A *Bayesian* extensive form game *with observable actions* [OR94] "models a situation in which the only uncertainty is about an initial move of chance that distributes payoff-relevant personal information among the player, in such a way that the information received by each player does not reveal any information about any of the other players"—chance selects types for the players; however, the players are subsequently fully cognizant at all points of all moves taken previously. Formally,

Definition 14 (Bayesian extensive game with observable actions [OR94]). *A Bayesian extensive game with observable actions is a tuple* $\langle \Gamma, (T_i), (p_i), (u_i) \rangle$ *where*

- $\Gamma = \langle N, H, P \rangle$ *is an extensive form game with perfect information, where* N *is a finite set of players,* H *is the set of action histories (a* history *is an actions' sequence, each component of a history being an action taken by a player), and* P *is a function that assigns to each nonterminal history (to each member of* $H \setminus Z$*,* Z *being the set of terminal histories) a member of* $N \cup \{c\}$*. (*P *is the player function,* $P(h)$ *being the player who takes an action after the history* h*. If* $P(h) = c$ *then chance (or, "the nature") determines the action taken after the history* h*.) The finite set of actions available after a nonterminal history* h *is denoted* $A(h) = \{a : (h, a) \in H\}$*. The empty sequence* $\emptyset$ *is a member of* H*.*

and for each player $i \in N$

- T_i *is the set of possible types (determined by nature) for player* i*.*
- p_i *is a probability measure on* T_i *for which* $p_i(t_i) > 0$ *for all* $t_i \in T_i$*, and the measures* p_i *are stochastically independent. (*$p_i(t_i)$ *is the probability that player* i *is selected to be of type* t_i*.)*
- $u_i : T \times Z \to \Re$ *is a utility function—*$u_i(\boldsymbol{t}, h)$ *is player* i*'s payoff when the profile of types is* $\boldsymbol{t}$ *and the terminal history of* Γ *is* h*.*[5]

We can now formally define a BCBG as follows:

[5] In our scenario, this payoff corresponds to the coalition payoff share an agent receives after a coalitional agreement that involves him has been reached.

Definition 15 (Bayesian Coalitional Bargaining Game). *A* Bayesian coalitional bargaining game (BCBG) *with discount factor $\delta \in (0, 1]$ and a finite number F of stages (or rounds) can be represented as a Bayesian extensive game with observable actions G, whose $\Gamma = \langle N, H, P\rangle, (T_i), (p_i)$ are as in Defn. 14. Further,*

- *The prior from which the players' types are drawn is common knowledge.*
- *All players in G are said to be active unless they have formed a coalition and abandoned the game.*
- *The nature chooses a player to become a proposer at each round of the game with probability $\gamma = 1/n$, where n is the number of active players at that round.*
- *A history h in the bargaining game consists of a series of bargaining actions, encompassing proposals by proposers and responses (acceptances or rejections of proposals) by responders.*
- *Whenever $n > 0$ players are active after some history h in the game, these players face a BCFP (see Defn. 3) with n players, in which every coalition C possible to be formed has coalitional actions A_C in its disposal, leading to stochastic outcomes $o \in \mathcal{O}$; and in which every player i has beliefs $\mu_i(h)$ following h, comprising a joint distribution over types T_{-i} of potential partners.*
- *Whenever a history h of bargaining actions has a proposal $\pi = \langle C, \boldsymbol{d}_C, \alpha_C\rangle \in h$ as a member which is followed by acceptances by all players $j \in C$, coalition C forms and abandons the bargaining game, and h is said to be terminal. Further, any history h that ends in the final round F of the game with the rejection of a proposal is also said to be terminal.*
- *The payoff u_i of player i following terminal history h that terminates in round s with the formation of coalition C with member types profile t_C, and in which i is a member and receives agreed (relative) payoff share $d_i \in [0, 1]$, is given as $u_i(\boldsymbol{t}_C, h) = d_i\delta^{s-1}Q(C, \alpha_C|t_C) = d_i\delta^{s-1}\sum_{o\in\mathcal{O}} \Pr(o|\alpha, t_C)R(o)$. Finally, any active agent i of type t_i (i.e., an agent that has not formed a coalition) after the end of any terminal history in the last round of the game receives his discounted reservation value $\delta^{F-1}rv_i = \delta^{F-1}V(t_i)$.*

5.3 Equilibria for Bayesian Coalitional Bargaining Games

Here we will define equilibria concepts for BCBGs. However, we have to lay the ground for this by first introducing some basic background concepts for games in extensive form.

When the extensive form game is, as in our case, Bayesian with observable actions, the players have perfect information about the actions of others (since they can observe them), but, as we saw above, there is uncertainty about an initial move of chance (or "nature") that distributes payoff-relevant private information among the players (and, thus, defines the agents' "type"). In such a game, after every action of any agent j in the game, an *information set* of game nodes is defined for every agent i (of some nature-specified type) that has to act, at which the agent is uncertain as to which node exactly the game has reached (as he is uncertain about the initial assignment of types—see Figure 5.1). We define information sets as follows:

Definition 16 (Information set). *An* information set $I_i(h)$ *of some player* i *in an extensive form game is a set of decision nodes reached after history* h*, such that:*

- *Every node in the set belongs to only one player,* i*, who is the player who has to move after* h*.*
- *When play reaches information set* $I_i(h)$*, player* i *does not know with certainty which node in the set has been reached, but has probabilistic beliefs about this.*

In the case of an extensive form game with observable actions, the beliefs of any agent k after some history h which is followed by some action (say some bargaining action π in the case of BCBGs) of some agent i, are updated by the Bayes rule and depend only on k's initial beliefs, the (observed, apart from nature's move) history h before i's action and the action of i (e.g., see Figure 5.1). It is common in such a game to assume that each agent will adopt a suitable *behavioural strategy* [OR94], associating with each information set in the game tree at which it must make a decision a distribution over action choices (for each of its possible types):

Definition 17 (Behavioural strategy). *A behavioural strategy of player* i *of type* t_i *is a collection* $(\sigma_i^{t_i}(I_i))_{I_i \in \mathcal{I}}$ *of independent probability measures, where* $\sigma_i^{t_i}(I_i)$ *is a probability measure over* $A(I_i)$*, with* $A(I_i)$ *being the set of actions available to* i *at each information set* I_i *within the set* $\mathcal{I}$ *of* i*'s information sets in the game.*

A strategy of a player in an extensive form game needs not be behavioural, but it does have to specify *a complete contingent plan* of actions for the player [MCWG95]:

Definition 18 (Strategy in an extensive form game). *A strategy of player i in an extensive form game is a function that assigns an action in $A(h)$ to each nonterminal history $h \in H$ at which it is i's turn to move—H being the set of action histories and $A(h)$ the set of actions available to i at h.*

Thus, a strategy in an extensive form game can be pure (assigning one specific action to each nonterminal h), behavioural, or mixed. While a behavioural strategy specifies a probability measure over the actions available to i at each one of his information sets, a mixed strategy for i is a probability measure over player i's set of pure strategies.[6]

When an extensive form game is assumed to be a game of perfect information, the appropriate solution concept is that of a *subgame-perfect equilibrium (SPE)* [OR94, MCWG95]: an SPE is a strategy profile whose restriction to any subgame following any history in the game is a Nash equilibrium of the subgame. In other words, in SPE the agents play best responses to each other's strategies following any history in the game—the strategies specified in the SPE profile define an equilibrium in any subgame of the game. (Intuitively, a *subgame of an extensive form game* is the portion of the game tree that follows some history h of actions.) Thus, an SPE defines a sequentially rational behaviour for the agents.[7] To formally define the SPE, we first have to define the notion of a subgame formally (for any generic extensive form game):

Definition 19 (Subgame of an extensive form game [MCWG95]). *A subgame of an extensive form game Γ_E is a subset of the game having the following properties:*

i It begins with an information set containing a single decision node, contains all the decision nodes that are successors (both immediate and later) of this node, and contains only *these nodes.*

ii If decision node x is in the subgame, then every $x' \in H(x)$ is also, where $H(x)$ is the information set that contains node x. (That is, there are no "broken" information sets in a subgame.)

We now define the SPE concept as follows:

[6] In games of perfect information, behavioural and mixed strategies are *outcome equivalent*[OR94].

[7] The SPE notion helps to remove the possibility of having equilibria describing "incredible threats" by some agents. When the sequential nature of deliberations and future rationality and "optimality" of current and future behaviour is taken into account (as is the case in SPE), then the possibility of incredible threats is removed [OR94, MCWG95].

Definition 20 (Subgame-Perfect Nash Equilibrium [MCWG95]). *A profile of strategies $\sigma = \langle \sigma_1, \cdots, \sigma_N \rangle$ in a N-player extensive form game Γ_E is a subgame perfect Nash equilibrium (SPE) if it induces a Nash equilibrium in every subgame of Γ_E.*

However, when uncertainty is assumed and beliefs are introduced in the game, the SPE concept has to be extended to account for these changes. Thus, the preferred solution concept for a Bayesian extensive form game (and, thus, for a BCBG game as well) is that of a *perfect Bayesian equilibrium (PBE)* [OR94]. A PBE comprises a profile of behavioural strategies for each agent as well as a *system of beliefs* dictating what each agent believes about the types of its counterparts at *each decision node* in the game tree. The standard rationality requirements must also hold: the strategy for each agent maximizes its *expected* utility given its beliefs; and each agent's beliefs are updated from stage to stage using Bayes rule, given the specific strategies being played. In other words, the agents' strategies are interdependent with their beliefs: in equilibrium, each agent's strategy is a best response to the strategies of others, given that agent's beliefs; and each agent's beliefs (viewed at all decision nodes, after each update following an equilibrium action) have to be such that they support the (sequential) equilibrium strategies. Somewhat informally and intuitively, the PBE concept combines the SPE and Bayes-Nash (see Section 2.2.2) equilibrium concepts, by requiring that the agents' strategies are sequential best responses given the agents' beliefs after any possible history and appropriate update of beliefs.

Before defining the PBE formally, we define the concept of *sequential rationality* for a strategy profile in a Bayesian extensive form game with observable actions. First, however, we formally define a system of beliefs for a generic extensive form game:

Definition 21 (System of beliefs [MCWG95]). *A system of beliefs μ in extensive form game Γ_E is a specification of a probability $\mu(x) \in [0, 1]$ for each decision node x in Γ_E such that $\sum_{x \in H} \mu(x) = 1$ for all information sets H.*

Thus, such a system specifies, for each information set, a probabilistic assessment by the player who moves at that set of the likelihood of being at each decision node in the set, conditional upon play having reached that information set. In the case of a Bayesian extensive form game with observable actions, when the uncertainty is about the initial move of chance specifying the players' types, the system of beliefs can be thought of as specifying, for each player i, after any observed history h, the probability $\mu_i(h)(t_j)$ assigned by i to any other j being of type $t_j \in T_j$, s.t. $\sum_{t_j} \mu_i(h)(t_j) = 1$. Now we define sequential rationality:

Definition 22 (Sequential rationality [MCWG95]). *Let μ denote a system of beliefs specifying all players' beliefs at each one of their decision nodes after any history h in the game. Con-*

sider an information set $I_i(h)$ controlled by player i that moves after h; let $E[\upsilon_i|I_i(h), \mu, \sigma_{i(I)}, \sigma_{-i(I)}]$ denote player i's expected utility starting at $I_i(h)$ if his beliefs regarding the types of others at $I_i(h)$ are given by $\mu_i(h) \in \mu$, he follows (behavioural) strategy $\sigma_{i(I)}$ and other players use (behavioural) strategies $\sigma_{-i(I)}$ starting at $I_i(h)$.

A (behavioural) strategy profile $\sigma = \langle \sigma_1, \cdots, \sigma_N \rangle$ in a N-player Bayesian extensive form game with observable actions is sequentially rational for player i at information set $I_i(h)$ given the system of beliefs μ *if,*

$$E[\upsilon_i|I_i(h), \mu, \sigma_{i(I)}, \sigma_{-i(I)}] \geq E[\upsilon_i|I_i(h), \mu, \tilde{\sigma}_{i(I)}, \sigma_{-i(I)}]$$

for all $\tilde{\sigma}_{i(I)}$ in the set of i's (behavioural) strategies starting at $I_i(h)$.

If strategy profile σ satisfies the condition for any player at any of his information sets (i.e., if the condition is satisfied for all *information sets in the game), then σ is sequentially rational given belief system μ.*

In words, a strategy profile is sequentially rational if no player finds it worthwhile once one of his information sets has been reached to revise his strategy given his beliefs (as embodied in μ) about what the others' types are and the strategies of others.

We can now formally define the PBE for a Bayesian extensive form game with observable actions as follows:

Definition 23 (Perfect Bayesian Equilibrium [OR94, MCWG95]). *A profile of (behavioural) strategies and system of beliefs (σ, μ) is a Perfect Bayesian Equilibrium in the Bayesian extensive form game Γ if it has the following properties:*

(i) The strategy profile σ is sequentially rational given system of beliefs μ.

(ii) The system of beliefs μ is derived from the strategy profile σ through Bayes' rule whenever possible. That is, if i acts after history h, and his action $a = a_i$ is in the support of $\sigma_i(t_i)(h)$ for some t_i in the support of any $\mu_j(h)$ (any other player j's beliefs after h), then for any $t_i' \in T_i$ we have

$$\mu_j(h \cup a)(t_i') = \frac{\sigma_i(t_i')(h)(a_i)\mu_j(h)(t_i')}{\sum_{t_i \in T_i} \sigma_i(t_i)(h)(a_i)\mu_j(h)(t_i)}$$

Now, if the assumption is made that the BCBG is to be played without the agents revising their beliefs in the course of the game, that is, under an assumption of *fixed beliefs*, the PBE

solution concept has to be revised. The appropriate solution concept for such a BCBG is a *sequential equilibrium under fixed beliefs*. Before defining the SEFB, we define a fixed system of beliefs for a BCBG:

Definition 24 (Fixed system of beliefs). *A fixed system of beliefs μ for a BCBG Γ is a system of beliefs that specifies, for each player i, after any observed history h, the probability $\mu_i(h)(t_j)$ assigned by i to any other j being of type $t_j \in T_j$, s.t., independently of h, $\mu_i(h)(t_j)$ is fixed to some value $\mu_i(t_j) \in [0,1]$. That is, at any h in Γ, $\mu_i(h)(t_j) = \mu_i(t_j)$ (and $\sum_{t_j} \mu_i(t_j) = 1$).*

Now we define the SEFB concept:

Definition 25 (Sequential Equilibrium under Fixed Beliefs). *Let G be a BCBG. Then, a profile of (behavioural) bargaining strategies, one for each player in N, is a* sequential equilibrium under fixed beliefs (*SEFB*) *for G, if, for each $i \in N$ and each history h, i's strategy continuation after h is optimal, given the bargaining strategies of other players and assumed fixed beliefs μ_i for any i in the game, provided to each agent by its own prior.*

Equivalently, the strategy profile σ is an SEFB for G if it is sequentially rational given a fixed system of beliefs μ (corresponding to the agents' priors).

When the strategies above are pure we are talking about an *SEFB in pure strategies*.

The SEFB is therefore defined as an extension of SPE and a restriction of PBE equilibria. Unlike the SPE solution, SEFB incorporates beliefs (and the fact that different agents can have widely varying beliefs about the value of any coalition in our coalitional bargaining setting, given their own priors regarding the types of partners); the beliefs are merely held fixed throughout the bargaining process (unlike what the PBE solution assumes).

Notice that the SEFB may be a more appropriate equilibrium concept for bargaining games where the agents do not have full observability of the bargaining actions of all of their opponents, and are thus unable to continuously update their beliefs regarding all of them. For example, in many realistic settings the bargaining players may not be allowed to observe others' responses to proposals. Further, the fixed-beliefs assumption can be useful when one wishes to deal with approximating the game's bargaining equilibria, while the SEFB concept can be of use when—as we shall do—one tries to relate a stability concept for the formation problem to a bargaining equilibrium: in Section 5.5 we will argue that it is not possible for a static concept such as the Bayesian core to account for the belief dynamics present in bargaining game: stability can only be defined with respect to beliefs that have "settled" to specific values.

Nevertheless, it is still possible to formulate the exact PBE solution of the problem. We now proceed to do just that.

5.3.1 Formulation of the PBE solution

Here we formulate the constraints that must hold on both strategies and beliefs in order to form a PBE. We note that, to the best of our knowledge, no similar equilibrium formulation has been provided in the literature for any model of coalitional bargaining under any type of uncertainty.

Let σ_i denote a behavioural strategy for i, mapping information sets (or observable histories h) in the game tree at which i must act into distributions over admissible actions $A(h)$. If i is a proposer at h (at stage s), let $A(h) = \mathcal{P}$, the finite set of proposals available at h. Then $\sigma_i^{h,t_i}(\pi)$ denotes the (behavioural strategy) probability that i makes proposal $\pi \in \mathcal{P}$ at h given its type is t_i. If i is a responder at h, then $\sigma_i^{h,t_i}(y)$ is the probability with which i accepts the proposal on the table (says *yes*) at h (and $\sigma_i^{h,t_i}(n) = 1 - \sigma_i^{h,t_i}(y)$ is the probability i says *no*). Let μ_i denote i's beliefs with $\mu_i^{h,t_i}(t_{-i})$ being i's beliefs about the types of others at h given its own type is t_i. Throughout, $V_s(t_C)$ is the (discounted) value to coalition with specific agents' types t_C, if it forms at bargaining stage s (and performs a coalitional action as prescribed by the accepted proposal).

We define the PBE constraints for the game by first defining the values to (generic) agent i at each node and information set in the game tree, given a fixed strategy for other agents, and the rationality constraints on his strategies and beliefs. We proceed in stages.

(1) Let ξ be a proposal node for i at history h at stage s. Since the only uncertainty in information set h involves the types of other agents, each $\xi \in h$ corresponds to one such type vector $t_{-i} \in T_{-i}$; let $h(t_{-i})$ denote this node in h. The value to i of a proposal $\pi = \langle C, \boldsymbol{d}_C, \alpha_C \rangle$ at $h(t_{-i})$ is:

$$q_i^{h(t_{-i}),t_i}(\pi) = p_{acc}^{h(t_{-i})}(\pi) d_i V_s(t_C) + \sum_r p^{h(t_{-i})}(\pi, r) q_i^{\xi/\pi/r,t_i} \tag{5.1}$$

where: $p_{acc}^{h(t_{-i})}(\pi)$ is the probability that all $j \in C$ (other than i) accept π (this is easily defined in terms of their fixed strategies); d_i is i's payoff share in $\boldsymbol{d}_C$; r ranges over response vectors in which at least one $j \in C$ refuses the proposal; $p^{h(t_{-i})}(\pi, r)$ denotes the probability of such a response; (note that it is critical to account for this probability explicitly, as the precise vector of responses affects future beliefs, and thus continuation payoff); and $q_i^{\xi/\pi/r,t_i}$ denotes the *continuation payoff* for i at stage $s+1$ at the node $\xi/\pi/r$ (following n after proposal π and responses r). This continuation payoff is defined (recursively) below. The value of π at history h (as opposed to a node) is determined by taking the expectation w.r.t. possible types:

$$q_i^{h,t_i}(\pi) = \sum_{t_{-i}} \mu_i^{h,t_i}(t_{-i}) q_i^{h(t_{-i}),t_i}(\pi) \tag{5.2}$$

(2) Suppose i is a responder at node $\xi = h(t_{-i})$ in history h at stage s. As above, ξ corresponds to specific t_{-i} in h. W.l.o.g. we can assume i is the first responder (since all responses are simultaneous). Let $p_{acc}^{h(t_{-i})}(\pi)$ denote the probability that all other responders accept π. We then define the value to i of accepting π at ξ as:

$$q_i^{h(t_{-i}),t_i}(y) = p_{acc}^{h(t_{-i})}(\pi) d_i V_s(t_C) + \sum_r p^{h(t_{-i})}(\pi, r) q_i^{\xi/y/r,t_i} \tag{5.3}$$

where again r ranges over response vectors in which at least one $j \in C$, $j \neq i$, refuses π; $p^{h(t_{-i})}(\pi, r)$ is the probability of such a response; and $q_i^{\xi/y/r,t_i}$ is the continuation payoff for i at stage $s+1$ after responses r by its counterparts. The value of accepting at h is given by the expectation over type vectors t_C w.r.t. i's beliefs μ_i^{h,t_i} as above.

The value of rejecting π at $\xi = h(t_{-i})$ is the expected continuation payoff at stage $s+1$:

$$q_i^{h(t_{-i}),t_i}(n) = \sum_r p^{h(t_{-i})}(\pi, r) q_i^{\xi/n/r,t_i} \tag{5.4}$$

(where r ranges over all responses, including pure positive responses, of the others).

(3) We have defined the value for i taking a specific action at any of its information sets. It is now straightforward to define the value to i of reaching any other stage s node controlled by $j \neq i$ or by nature (i.e., chance nodes where a random proposer is chosen).

For an information set h_j where j makes a proposal, consider a node $\xi = h_j(t_j)$ where j is assumed to be of type t_j. Then, j's strategy $\sigma_j^{h_j,t_j}$ specifies a distribution over proposals π (determined given the values $q_j^{h_j,t_j}(\pi)$ which can be calculated as above, and j's type t_j). Agent i's value $q_i^{t_i,h_j(t_j)}$ at this node is given by the expectation (w.r.t. this strategy distribution) of its accept or reject values (or if it is not involved in a proposal, its expected continuation value at stage $s+1$ given the responses of others). Its value at h_j is then $Q_i^{t_i}(h_j) = \sum_{t_j} \mu_i^{h_j,t_i}(t_j) q_i^{t_i,h_j(t_j)}$. We define $Q_i^{t_i}(h_i)$ (where i is the proposer) as the value of his best possible proposal at this information set, calculated as in Case 1 above.

Finally, i's value at information set h that defines the stage s continuation game (i.e., where nature chooses proposer) is

$$q_i^{h,t_i} = \frac{1}{m} \sum_{j \leq m} Q_i^{t_i}(h_j) \tag{5.5}$$

where m is the number of active agents, and h_j is the information set following h in which j is the proposer.

(4) We are now able to define the rationality constraints. We require that the payoff from the

equilibrium behavioural strategy σ exceeds the payoffs of using pure strategies. Specifically, in PBE, for all i, $t_i \in T_i$, all h corresponding to i's information sets $I_i(h) \in \mathcal{I}$, and all actions $b \in A(h)$, we have:

$$\sum_{t_{-i}} \mu_i^h(t_{-i}) \sum_{a \in A(h)} \sigma_i^{h,t_i}(a) q_i^{h(t_{-i}),t_i}(a) \geq \sum_{t_{-i}} \mu_i^h(t_{-i}) q_i^{h(t_{-i}),t_i}(b) \tag{5.6}$$

We also add constraints for the Bayesian update of belief variables for any agent i regarding type t_j agent j performing a_j at any h (for all i, $t_i \in T_i$, all h and all a_j):

$$\mu_i^{h \cup a_j, t_i}(t_j) = \mu_i^{h,t_i}(t_j) \sigma_j^{h,t_j}(a_j) / \sum_{t_j^k \in T_j} \mu_i^{h,t_i}(t_j^k) \sigma_j^{h,t_j^k}(a_j) \tag{5.7}$$

Finally, we add the obvious constraints specifying the domain of the various variables denoting behavioural strategies or beliefs: they take values in $[0, 1]$ and sum up to 1 as appropriate.

This ends the PBE formulation. Note that this formulation constitutes a polynomial constraint satisfaction program (CSP), made up of the set of equalities and inequality constraints described above. This CSP encompasses μ-variables and σ-variables describing the beliefs and behavioural strategies of the agents; auxiliary q-variables (defined recursively through equations 5.1— 5.5) and denoting the value of bargaining actions; auxiliary p-variables used in the q-values definitions and denoting the probability of acceptance or rejection of proposals, and defined basically as products of behavioural strategies of responders (e.g., if π was addressed to C by i, and j, k of assumed types t_j, t_k respectively are the other members of C, $p_{acc}^{h(t_j,t_k)}(\pi) = \sigma_j^{h,t_j}(y)\sigma_k^{h,t_k}(y)$); and constants denoting relative demands (e.g., d_i), types (e.g., t_k) and pure bargaining actions (e.g., a_j) of the agents. Figure 5.2 summarizes the CSP formulation of the PBE.

Unfortunately, solving realistic versions of this program is, as we shall now demonstrate, practically impossible.

5.3.2 Complexity of the PBE Solution

The calculation of the PBE is rendered practically intractable because of a variety of combinatorial interactions evident in the constraints above. In several points in the program—which assumes random choice of proposer in each of several bargaining rounds—variables describing the strategies of an agent depend on the yet unknown strategies of others, which in turn depend on (updated) beliefs that themselves depend on strategies (see for example constraints 5.6

$$\sum_{t_{-i}} \mu_i^h(t_{-i}) \sum_{a \in A(h)} \sigma_i^{h,t_i}(a) q_i^{h(t_{-i}),t_i}(a) \geq \sum_{t_{-i}} \mu_i^h(t_{-i}) q_i^{h(t_{-i}),t_i}(b)$$

$$\forall i \in N, t_i \in T_i, h : P(h) = i, b \in A(h) \tag{5.8}$$

$$\mu_i^{h \cup a_j, t_i}(t_j) = \mu_i^{h,t_i}(t_j) \sigma_j^{h,t_j}(a_j) / \sum_{t_j^k \in T_j} \mu_i^{h,t_i}(t_j^k) \sigma_j^{h,t_j^k}(a_j)$$

$$\forall i \in N, t_i \in T_i;$$
$$\forall h \in H : P(h) = j, j \in N, a_j \in A(h), t_j \in T_j \tag{5.9}$$

$$0 \leq \mu_i^{h,t_i}(t_j) \leq 1 \qquad \forall i \in N, t_i \in T_i, h \in H, t_j \in T_j \tag{5.10}$$

$$\sum_{t_j \in T_j} \mu_i^{h,t_i}(t_j) = 1 \qquad \forall i \in N, t_i \in T_i, h \in H \tag{5.11}$$

$$0 \leq \sigma_i^{h,t_i}(a_i) \leq 1 \qquad \forall i \in N, t_i \in T_i, h \in H : P(h) = i, a_i \in A(h) \tag{5.12}$$

$$\sum_{a_i \in A(h)} \sigma_i^{h,t_i}(a_i) = 1 \qquad \forall i \in N, t_i \in T_i, h \in H : P(h) = i \tag{5.13}$$

and q-variables defined recursively as in Eq. 5.1— 5.5.

Figure 5.2: The CSP formulation for the PBE solution of a Bayesian coalitional bargaining game.

and 5.5). More specifically, many problems arise due to the fact that the agents have to calculate their future expected continuation payoff, which depends on the future actions of opponents, which in turn depend on beliefs updated in the future, given the yet uncalculated strategies in this and next rounds. Further, the program is not convex:

Proposition 5. *The constraint satisfaction program describing the PBE solution for a coalitional bargaining game is non-convex.*

(We provide a proof for this proposition in Appendix A.)

Nevertheless, the program *is* decidable: it is equivalent to deciding whether a system of polynomial equations and inequalities has a solution [BPR96]. The problem is decidable, but is intractable. [BPR96], for example, have provided an algorithm with exponential complexity for deciding this problem. Specifically, the complexity of deciding whether a system of s polynomials, each of degree at most d in k variables has a solution is $s^{k+1}d^{O(k)}$. To the best of our knowledge, this is the fastest algorithm to decide this problem—and is exponential to k.

In our case, assuming a random choice of proposer at each of F rounds, we can show that if α is the number of pure strategies, N the number of agents, T the number of types,

then $s = O((N\alpha(N-1))^F * N * T)$, $d = NF$ and $k = O((N\alpha(N-1))^F * N * T * \alpha)$. This is due to a variety of combinatorial interactions evident in the constraints above, creating as they do interdependencies between belief and strategy variables. More specifically, the degree of a round F polynomial is N (due to the need to multiply variables denoting acceptance probabilities of various agents together to come up with various $p_{acc}(\pi)$ probabilities of all agents accepting a proposal π), but the degree of a round one polynomial constraint will be NF (due to the need to take continuation payoffs into account when specifying the constraints). The number of constraints needed if a single stage of bargaining was to be used is $O(\alpha NT)$.[8] However, there exist F rounds, and at each one the proposer is chosen *randomly*: this necessitates that all appropriate constraints have to be specified for any possible choice of proposer at every stage. This requirement increases the number of polynomial constraints needed in the system by a factor of N^F. In fact, the number of all information sets in the game is $O((N\alpha(N-1))^F * N * T)$—the approximate number of information sets (observed histories) for each agent of each potential type[9] times the number of agents times the number of types—and we need to define constraints at each one of them. Finally, the number of variables k in the system is $O((N\alpha(N-1))^F * N * T * \alpha)$, because of the need to define at most α behavioural σ-variables at each information set—the number of μ-variables needed is less than that by a factor of at least $N^F * \alpha$ since agents need not update beliefs when nature selects a proposer (and the additional factor α above is not relevant here either). Therefore, solving the system can be guaranteed, but this can take

$$O(((N\alpha(N-1))^F NT)^{O((N\alpha(N-1))^F NT\alpha)+1}(NF)^{O((N\alpha(N-1))^F NT\alpha)})$$

time, which is obviously prohibitively high.

Notwithstanding the complexity of tackling the problem, there exist off-the-shelf (e.g., implemented in the MATLAB optimization toolbox) *iterative techniques for solving non-linear optimization problems*, such as those described in [JKP72]. Some of these techniques, commonly referred to as Sequential Quadratic Programming methods, might even guarantee super-

[8] Roughly, in *a single stage, with a fixed proposer* we need $O(\alpha)$ constraints for one specific type of the proposer, and $O(\alpha T)$ for all its possible types; we need $O(\alpha T 2(N-1))$ constraints for the responders; and $O((N-1)T\alpha)$ constraints to describe the Bayes rule updates of beliefs for each one of the $N-1$ responders after each possible proposal.

[9] The $(N-1)$ factor is an approximation: it corresponds to potentially $N-1$ responders responding to a proposal; we use $N-1$ instead of 2^{N-1} since we assume simultaneous responders' moves: therefore, the responders do not respond after observing others' responses, and we can assume that any subsequent proposer observes the vector of $N-1$ responses.

linear convergence to a solution point, in terms of the number of iterations involved in order to optimize the solution. Unfortunately, this does not necessarily guarantee that the process of calculating an intermediate solution is efficient in terms of time taken; in addition, these methods require certain conditions to hold, such as the requirement that the polynomials participating in the constraints are convex functions. (Unfortunately, as mentioned above, it is easy to show that our polynomial program is non-convex.) Despite these drawbacks, however, these methods might work well for restricted versions of our problem.

In summary, the formulation above characterizes the PBE solution of our coalitional bargaining game as a solution of a polynomial program. However, it does not seem possible that this solution can be efficiently computed in general. Nevertheless, this PBE formulation may prove useful for the computation of a PBE in a bargaining setting with a limited number of agents, types, proposals and bargaining stages. Perhaps more importantly, it is the first attempt to describe the equilibrium solution of a coalitional bargaining problem under uncertainty.

5.4 Coalitional Bargaining Heuristics

The calculation of the PBE solution is extremely complex due to both the size of the strategy space (as a function of the size of the game tree, which grows exponentially with the problem horizon), and the dependence between variables representing strategies and beliefs, as explained above. We present a *heuristic* algorithm that circumvents these issues to some degree by:

(a) performing only a small lookahead in the game tree in order to decide on a action at any stage of the game; and

(b) fixing the beliefs of each agent during this process.

This latter approach, in particular, allows us to solve the game tree by *backward induction*, essentially approximating heuristically the computatation of an SEFB equilibrium for this fixed-beliefs game. Note that while beliefs are held fixed during the lookahead (while computing an immediate action), they do get updated once the bargaining action (i.e., proposal or response) is selected and executed, and thus the beliefs do evolve based on the actions of others during bargaining. Furthermore, we allow *sampling* of type vectors in the computation to further reduce the tree size.

We now describe the algorithm in detail. Initially, the agents' types are assumed to be drawn from a common prior, known to all the agents. At any stage of the game, with a particular

collection of active agents (each with their own beliefs), the steps of the algorithm (summarized in Figure 5.3) are as follows:

1. Each agent constructs a game tree consisting of the next l rounds of bargaining, for some small *lookahead* l. This tree represents the game for the next l rounds, given the chosen proposer at the current round. (If less than l rounds remain, the tree is suitably truncated.) All active agents are assumed to have *fixed beliefs* at each node in this tree, corresponding to their beliefs at the current stage. Each agent computes its best response action for the current round using backward induction to approximate an SEFB equilibrium of this limited depth BCBG under fixed beliefs game. (We elaborate below.) Moreover, by solving the game tree, each agent estimates the current round best response action for each possible type of every other active agent. Furthermore, the agents *sample* partners' types when calculating the values of coalitions and proposals.

2. The proposer and each responding player execute the actions they computed as best in this reduced tree for the current round of bargaining. If a coalition is formed, it breaks away, leaving the remaining players as active.

3. All active agents *update their beliefs*, given the actual observed actions of others in the current round, and their expected best responses, using Bayesian updating (as in Eq. 5.7). Further, each agent keeps track of the belief updates that any other agent of a specific type would perform at this point.

4. The next bargaining round is implemented by repeating these steps until a complete coalition structure is determined or the maximum number of bargaining rounds is reached.

We note also that the algorithm can be combined with belief updates after observing the results of *coalitional* actions (in RL style).

We stress that this heuristic best-response algorithm *does not approximate the PBE solution*; getting any good bounds for a true PBE approximation would only be likely by assuming belief updating at *every* node of the game tree mentioned in Step 1. However, if our algorithmic assumptions are shared by all agents, each agent—using the method mentioned in Step 1 above—can determine their best responses to others' approximately optimal play, and thus their play resembles an SEFB equilibrium of the fixed-beliefs game. The use of this heuristic SEFB approximation algorithm in Step 1 above can be further motivated by the fact that—as we show in the next section of this chapter—SEFB play by the agents leads to stable (BC)

F: number of rounds;
l: lookahead $<< F$;
L: current round ($1 \leq L \leq F$);
N: set of active agents;
μ: agent types prior (common knowledge)
$\forall i : t_i \in T_i$ is private information

1. Nature selects proposer ψ; agents *solve* an l-rounds BCBG (if $L > F - l + 1$, then $l = F - L + 1$), using *sampling* of partners' types, and assuming *fixed beliefs* at the game tree nodes. Thus,
   - Each agent computes its best response action for the current round (i.e., the first level of the l-rounds BCBG tree).
   - Each agent i estimates the current round best response action for each possible type of every other active agent.
2. Bargaining actions executed: agent ψ states calculated best response proposal π to C and all other $\zeta \in C$ respond with their calculated best responses to π.
   - if all $\zeta \in C$ accept, C is formed and breaks away; $N = N \setminus C$
3. Any active agent i *update their beliefs* after observing agent j's bargaining action β; further, any agent i keeps track of beliefs μ_{t_ι} of any other active agent ι of type t_ι ($\mu^i_{t_\iota}$ denote those beliefs as updated by i):

$$\forall i, \iota \in N; \forall t_\iota \in T_\iota; \forall t_j \in T_j : \mu^i_{t_\iota}(t_j|\beta) = \mu^i_{t_\iota}(t_j)\sigma_j^{t_j}(\beta) / \sum_{t_j^k \in T_j} \mu^i_{t_\iota}(t_j^k)\sigma_j^{t_j^k}(\beta)$$

4. $L = L + 1$; Repeat steps 1–3 until a complete CS is determined or round F is over.

Figure 5.3: A heuristic best-response coalitional bargaining algorithm with belief updates.

coalitional configurations. Nevertheless, we note that it is not possible to get any good approximation bounds for this (Step 1) algorithm: it is only a heuristic method to approximate SEFB play (but it does behave well in practice). We will come back to this point after describing the method in some detail.

We assume that the agents proceed to negotiations that will last l rounds (corresponding to the algorithm's lookahead value l) under the assumption that all beliefs will remain fixed to their present values throughout the (Step 1) process. We will present the deliberations of agent i during negotiations. For *fixed* types t_{-i} of possible partners, drawn according to μ_i, i will reason about the game tree and assume fixed beliefs of other agents. (Agents *will* track of the updates of other agents' beliefs after this stage of bargaining; see Step 2 above). Then, i can calculate the optimal action of any t_j agent (including himself) at any information set by taking expectations over the corresponding tree nodes.

We begin our analysis at the last round l of negotiations. Consider any node ξ after history h where i of type t_i is a responder to proposal $\pi \in \mathcal{P}$ (see node $h(t_{-i})$ in Figure 5.4). At any

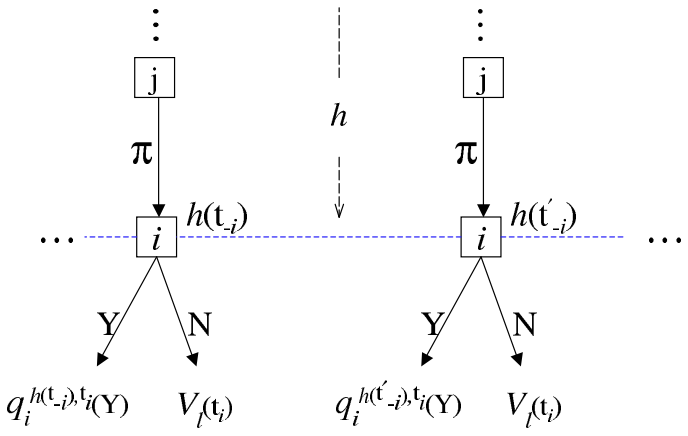


Figure 5.4: Evaluation of potential decisions for responder $i \in C$ of type t_i at information set h after proposal π proposed to C at the last round of negotiations. $h(t_{-i})$ denotes a node in this information set (following history h) where i assumes his opponents to have a specific type vector t_{-i}. The value of acceptance to i, at information set h, is $q_i^h(y) = \sum_{t_{-i} \in t_C} \mu_i^{h,t_i}(t_{-i}) q_i^{h(t_{-i}),t_i}(y)$. (The value of refusal (saying "no"—action n) is, trivially, $q_i^h(n) = V_l(t_i)$, with $V_l(t_i)$ denoting the discounted reservation value for i of type t_i at this last round l of negotiations.)

such ξ, i assumes a specific type vector for partners, and he expects a value for accepting (i.e., saying "yes"—action y) that is different to his (discounted) reservation value $V_l(t_i)$ only if all other responders accept the proposal as well:

$$q_i^{h(t_{-i}),t_i}(y) = \begin{cases} d_i V_l(t_C) & \text{if all } t_{-i} \in t_C \text{ accept} \\ V_l(t_i) & otherwise \end{cases} \tag{5.14}$$

(where $d_i \in \boldsymbol{d}_C$ is the share for i specified by the proposer as part of proposal π). However, to evaluate this acceptance condition, i would need to know the other responders' strategies (which in turn depend on i's strategy). Therefore, i will make the simplifying assumption that

all other responders j evaluate their response to π[10] by assuming that the rest of the agents (including i) will accept the proposal. Thus, any j with $t_j \in t_{-i}$ is *assumed* by i to accept if he (i.e., j) evaluates his expected payoff from acceptance as being greater than his (discounted) reservation payoff:[11]

$$d_j \sum_{t_{-j} \in t_C} \mu_j(t_{-j}) V_l(\{t_j, t_{-j}\}) \geq V_l(t_j) \tag{5.15}$$

With this assumption, i is able to evaluate the acceptance condition in Eq. 5.14 above, and so calculate a specific $q_i^{h(t_{-i}),t_i}(y)$ value.

At node $\xi = h(t_{-i})$, i can also evaluate his refusal value as $q_i^{h(t_{-i}),t_i}(n) = V_l(t_i)$ in this last round. Then, responder i's actual strategy at h—that is, his best response at information set h, containing all possible $h(t_i)$, $h(t_{-i})$ decision nodes—can be evaluated as the strategy maximizing i's expected value given μ_i^{h,t_i}:

$$\sigma_i^{h,t_i} = \arg \max_{r \in \{y,n\}} \{ \sum_{t_{-i} \in t_C} \mu_i^{h,t_i}(t_{-i}) q_i^{h(t_{-i}),t_i}(r) \}$$

If i is a *proposer* of type t_i deliberating at $\xi = h(t_{-i})$, the value of making proposal π is:

$$q_i^{h(t_{-i}),t_i}(\pi) = \begin{cases} d_i V_l(t_C) \text{ if } \sigma_j^{h,t_j} = y, \ \forall j \\ V_l(t_i) \quad otherwise \end{cases} \tag{5.16}$$

(i.e., i will get his reservation value unless all the responders of the specific type configuration agree to this proposal). Furthermore, i's expected value $q_i^{h,t_i}(\pi)$ from making proposal π to coalition C at h can be determined given μ_i^{h,t_i}. Thus, the best proposal that i of type t_i can make to coalition C is the one with maximum expected payoff: $\sigma_i^{C;h,t_i} = \arg\max_\pi q_i^{h,t_i}(\pi)$ with expected payoff $q_i^{C;h,t_i}$.

However, i can also propose to other coalitions at h as well. Therefore, the coalition C^* to which i should propose is the one that guarantees him the maximum expected payoff: $C^* = \arg\max_C \{q_i^{C;h,t_i}\}$. If P^* is the payoff allocation associated with that proposal, then the optimal coalition-allocation pair that t_i can propose in this subgame (that starts with i proposing at h)

[10] Recall that proposal $\pi = \langle C, \boldsymbol{d}_C, \alpha_C \rangle$ calls for the formation of a coalition C with a specific payoff shares configuration $\boldsymbol{d}_C$ and a coalitional action α_C. For notational simplicity, we do not explicitly link value notation or the shares d_i to their corresponding proposal here, nor do we explicitly refer to proposed coalitional actions. Instead, we use $V(t_C)$ as a shorthand denoting the value of a coalition assuming a type vector t_C for its members and a choice of coalitional action "as prescribed by proposal π".

[11] Note that j does not actually evaluate his acceptance value as in Eq. 5.15, but as in Eq. 5.14: agents do not evaluate their acceptance value by assuming that others will accept for sure. However, in order to evaluate Eq. 5.14, any agent assumes that others decide to accept or not using Eq. 5.15.

is: $\sigma_i^{*;h,t_i} = \{C^*, P^*\}$ with maximum expected payoff $q_i^{C^*;h,t_i}$. Finally, if there exists more than one optimal proposal for i, i randomly selects any of them (this is taken into account in agents' deliberations accordingly).

Of course, when the subgame starts, an agent i does not know who the proposer in this subgame will be; and i has only probabilistic beliefs about the types of his potential partners. Thus, i has to calculate his *continuation payoff* $q_i^{l:\xi,t_i}$ at stage l (that starts at node ξ) with m participants, in the way explained in the previous section. This is straightforward, as i can calculate his expected payoffs from participating in any subgame where some j proposes, given that any i can calculate the optimal strategies (and associated payoffs) for any j in this round l subgame.

Now consider play in a subgame starting in period $l-1$, again with the participation of m agents. The analysis for this round can be performed in a way completely similar to the one performed for the last round of negotiations. However, there is one main difference: the payoffs in the case of a rejection are now the continuation payoffs (for agents of specific type) from the last round subgame. We have to incorporate this difference in our calculations. Other than that, we can employ a similar line of argument to the one used for identifying the equilibrium strategies in the last period. Proceeding in this way, we define the continuation payoffs and players' strategies for each prior round, and finally determine the first round actions for any proposer i of type t_i or any responder j of type t_j responding to any proposal—employing backward induction.

Though this heuristic fixed-beliefs algorithm is intuitive and—as our experiments in Section 5.6 will demonstrate—does help the agents perform well in practice, it does not come with good approximation bounds for the SEFB solution: Using standard Chernoff bounds analysis, one can show that sampling is *not* harmful when agents calculate coalitional values in *the last* round (i.e., round l) of negotiations (under fixed beliefs). However, due to the varying agents' beliefs, even small differences in the calculation of the coalitional values (differences which are inescapable due to the use of sampling), when propagated up the tree may result to the adoption of strategies by the opponents that are totally different to the ones expected. This problem is mainly encountered when the values of coalitional agreements are very close to each other. Thus, the adoption of the strategy calculated by the algorithm cannot guarantee to an agent a value that is close to his actual equilibrium one. One possible "remedy" to this problem, would perhaps be to assume that all agents use the *same* samples—i.e., that they play according to a

(centrally computed) "common copy" of the algorithm.[12] However, we believe that this would not be a realistic assumption to make in this setting. Even though we cannot guarantee good theoretical bounds for the algorithm, we note that the algorithm is expected to (and does) have good behaviour in practice (as in many realistic situations the values of the various agreements are not expected to be very close to each other).

Finally, we note that the complexity of this heuristic fixed-beliefs algorithm is essentially the complexity of using backward induction to traverse a game tree with depth bound l: the tree branching factor is approximately $N\alpha N$, with each of the N proposers chosen randomly at each round to propose one of α bargaining actions to (in the worst case) N responders (which respond simultaneously). Since nature decides on one of the $|T|$ type profiles in the beginning of the game, the complexity of this heuristic algorithm is approximately $O((\alpha N^2)^l|T|)$. The complexity of the heuristic best-response bargaining algorithm in Figure 5.3 is then $O(F(\alpha N^2)^l|T|)$, assuming F rounds of bargaining. This running time can of course become very demanding, but it can also be kept within reasonable limits with the appropriate choice of the sampling size (when sampling type vectors), and lookahead value l.[13] Furthermore, in practice, as agents form coalitions and abandon the game the number α of bargaining actions is reduced (as there are fewer coalitions to which the agents can make proposals), with the subsequent benefits to running time.

5.5 A Non-Cooperative Justification of the Bayesian Core

As was pointed out earlier in this thesis, the core occupies a central position as a *cooperative* solution concept in coalition formation. Nevertheless, the core cannot itself capture the *non-cooperative* dynamics and player interactions that lead to its creation. These interactions are naturally described by equilibria notions that justify the choices of rational players.

It is natural, then, to raise the question of equivalence of these different solution concepts. As we saw in Section 5.1, there exists a corpus of literature that tries to analyze the non-cooperative underpinnings of the core, and relate it to the equilibria solution concepts of the bargaining games that lead to it. The aim of such research is to show that the non-cooperative (commonly SPE) equilibria of the bargaining game produce outcomes that lie in the core.

[12] For example, Kearns, Mansour and Singh [KMS00] use this assumption when computing near-Nash equilibria in stochastic games.

[13] As we mention in Chapter 7 of this dissertation, the development of less demanding algorithms based on this heuristic is also possible—but of course this would come at a price regarding the accuracy of the approximation.

To date, none of this work has taken type uncertainty (or any other form of uncertainty) into account. We, on the other hand, do take type uncertainty into account, and investigate the relationship between the equilibria outcomes of a Bayesian coalitional bargaining game and the Bayesian core configurations of its underlying Bayesian coalition formation game.

Of course, the Bayesian core—being a stability concept—implicitly assumes that the agents' beliefs are fixed to specific values in its definition. Certainly, the fact that an agent i makes a specific proposal to a set of other agents (or accepts or rejects a particular proposal) can influence j's beliefs about i's type, and thus j's behaviour at future rounds of the BCBG. However, the belief dynamics present in a multi-round Bayesian bargaining game *cannot* possibly be reflected in a static cooperative solution concept. Therefore, we make the simplifying assumption that agents' beliefs remain *fixed* throughout bargaining.

By making this restriction to BCBGs played under fixed beliefs[14], we can then prove that if the (strict or the weak) BC of the coalitional game is non-empty (and so are its subgames, which are), then there exists an SEFB profile of the bargaining game that produces an element of the Bayesian core (if we also assume that the BCs of the subgames have one element with some reasonable properties); and conversely, we can show that if there exists an SEFB bargaining equilibrium profile with certain properties (to be elaborated below), then it leads to a configuration that has to be in the (weak) BC of the coalitional game. Thus, we show that the existence of stable coalition structures in a coalition formation problem under uncertainty implies the existence of an equilibrium bargaining profile that leads to their formation; and also that, even under uncertainty, bargaining according to an equilibrium strategy profile leads to stable coalitions. We are thus able to describe some notion of equivalence between the cooperative and non-cooperative Bayesian coalition formation solution concepts, and provide a non-cooperative justification for the use of the Bayesian core as a coalitional stability concept under uncertainty.

We start by proving our first relevant proposition. Note that throughout our proofs, we will assume an infinite horizon for bargaining. We first define a subclass of bargaining games that we will be interested in in this proof.

Definition 26. *Let $\mathcal{C}$ be the class of N-player BCFPs with the following properties:*

[14]One could envisage dropping the fixed beliefs assumption, and trying to show that equilibrium play leads to stable BC outcomes with respect to settled beliefs of the agents "at the end of the game". However, we do not think that this could be possible, given that agents who form coalitions abandon the game, and thus are unable to update beliefs anymore: At which point (of the several possible ones) does a coalition actually abandon the game? Which is the set of beliefs that one could consider as "settled" in order to define a BC outcome? It does no seem possible to provide a clear answer to those questions.

1. *All subgames have a nonempty strict BC.*

2. *For every member $B_N = \langle CS_N, \boldsymbol{d}_N, \boldsymbol{a}_N \rangle$ of the strict BC, where CS_N is of the form $\{S_1, ..., S_k\}$, every subgame with set of players $T \subseteq N$ has an element in its strict BC in which the coalition structure is of the form $\{T \cap S_1, ..., T \cap S_k\}$ and also the demand vector is the projection of $\boldsymbol{d}_N$ to the corresponding coalition.*

In the above definition we ignore the empty sets that may arise if T does not intersect any of the S_j's. Note that by Observation 5, the properties of Definition 26 are already satisfied by subgames in which the set of players is a union of some of the coalitions of CS_N. Hence the definition simply imposes the same properties for other subsets of N as well.[15] We are now ready to prove the following:

Proposition 6. *Let $\mathcal{P} \in \mathcal{C}$ be an N-player BCFP. Then, for every member $B_N = \langle CS_N, \boldsymbol{d}_N, \boldsymbol{a}_N \rangle$ of the strict BC of $\mathcal{P}$, there exists an SEFB equilibrium $\sigma^* = \sigma^*(B_N)$ in pure strategies, of the corresponding BCBG G (under fixed beliefs) with N players and random proposers, such that, the coalition structure induced by σ^* is exactly $\langle CS_N, \boldsymbol{d}_N, \boldsymbol{a}_N \rangle$.*

Proof. Let $B_N = \langle CS_N, \boldsymbol{d}_N, \boldsymbol{a}_N \rangle$ be an arbitrary element of the BC of $\mathcal{P}$ and let BC_S represent the strict BC of the subgame where the set of agents is S. Let $CS_N = \{S_1, S_2, ..., S_k\}$ for some k, where $\cup S_i = N$ and $S_i \cap S_j = \emptyset$ for every i, j. For each S we choose an element $B_S = \langle CS_S, \boldsymbol{d}_S, \boldsymbol{a}_S \rangle \in BC_S$. In particular, we choose such an element according to Definition 26. For example if S is the union of some of the S_i's, i.e., if it consists precisely of some of the coalitions of CS_N, then we let B_S be the restriction of $\langle CS_N, \boldsymbol{d}_N, \boldsymbol{a}_N \rangle$ to S, which by Observation 5 lies in BC_S. Our way of choosing these core elements for each subgame ensures the following, easy to verify, fact:

Fact 1. *Let $T \subseteq N$ and suppose the coalition structure in B_T is $\{T_1, ..., T_l\}$. Then for the subgame where the set of players is $S = T \setminus T_1$, the structure in the corresponding core element B_S is $\{T_2, ..., T_l\}$ and the demand and action vectors are the same as in T.*

[15] We have to make this assumption about this property of the subgames' BC in order to guarantee subgame perfection for the equilibrium strategy of some agent in *any* subgame, regardless of any preceding history—even in the case of subgames where the equilibrium strategy profile was *not* followed in the preceding subgame, due to any random reason. However, this assumption is in essence a technicality, and is redundant in all the parts of the game tree where the actual equilibrium strategy profile is followed: the equilibrium strategy profile actions taken there are actually best responses, without the need of this assumption. However, subgame perfection has to be guaranteed for an equilibrium strategy even if some player's "hand was trembling" and the equilibrium was not followed in the preceding history (a solution concept that tries to capture and combat those issues is the *trembling hand equilibrium*—see, e.g., [MCWG95]).

Given a triplet $\langle C, \boldsymbol{d}_C, \alpha_C \rangle$, with $i \in C$, we will denote by $\bar{p}_i^i(C, \boldsymbol{d}_C, \alpha_C)$ the expected payoff of i from the formation of coalition C, according to i's beliefs. We will also use $\bar{p}_i^i(B_S)$ to denote $\bar{p}_i^i(C_i, \boldsymbol{d}_{C_i}, \alpha_{C_i})$, where C_i is the coalition in the coalition structure of B_S that i belongs to.

Consider now the following strategy σ_i^* for a player i:

(i) If i is the proposer in some round of the game, and the set of agents still present is S, let C denote i's coalition in the coalition structure of B_S, and let $\boldsymbol{d}_C$, α_C be its corresponding demand vector and action, i.e., we look at the projection of B_S to coalition C that contains i. Then, i proposes $\langle C, \boldsymbol{d}_C, \alpha_C \rangle$ to the rest of the agents in C[16].

(ii) If i is a responder, the subset of agents still present is S, and the standing proposal is $\langle T, \boldsymbol{d}_T, \alpha_T \rangle$ (with $i \in T$), then i accepts *iff* $\bar{p}_i^i(\langle T, \boldsymbol{d}_T, \alpha_T \rangle) \geq \bar{p}_i^i(B_S)$.

Let σ^* be the profile of the strategies of all players. It is clear that, no matter what the nature's choice of proposers in the game is, if σ^* is played then the outcome of the game is exactly B_N. To see this, suppose that the first random proposer at round 1, say i, belongs to S_1 (recall $B_N = \langle CS_N, \boldsymbol{d}_N, \boldsymbol{a}_N \rangle$ and $CS_N = \{S_1, S_2, ..., S_k\}$). Then i will propose $\langle S_1, \boldsymbol{d}_{S_1}, \alpha_{S_1} \rangle$, with $\boldsymbol{d}_{S_1}, \alpha_{S_1}$ being the projection of $\boldsymbol{d}_N, \boldsymbol{a}_N$ to S_1. By the definition of σ^* all other members of S_1 will accept and the game will go to round 2 with the remaining players $L = N \setminus S_1$. Suppose agent j is now the proposer and $j \in S_2$. By the way we defined B_L, we know that $CS_L = \{S_2, ..., S_k\}$, (because L consists of a collection of coalitions of CS_N) therefore j will propose $\langle S_2, \boldsymbol{d}_{S_2}, \alpha_{S_2} \rangle$ with $\boldsymbol{d}_{S_2}, \alpha_{S_2}$ being the projection of $\boldsymbol{d}_N, \boldsymbol{a}_N$ on S_2. The other members of S_2 will accept the proposal and the game will continue in the same manner. Hence after k rounds the game will end and the outcome will be B_N.

It remains to show that σ^* is an *SEFB* equilibrium.

Assume i is a proposer at some round of the game, the set of active players is S and all players apart from i will play according to σ^* from this point of the game onwards. Let $\langle C, \boldsymbol{d}_C, \alpha_C \rangle$ be the triplet of B_S with $i \in C$. We want to show that i cannot gain by deviating from σ^*. Suppose i deviates by proposing $\langle T, \boldsymbol{d}_T, \beta_T \rangle$, different from $\langle C, \boldsymbol{d}_C, \alpha_C \rangle$, where C is the coalition of B_S that i belongs to. Consider first the case that $\bar{p}_i^i(T, \boldsymbol{d}_T, \beta_T) > \bar{p}_i^i(B_S)$. Note that in this case T cannot be a singleton since that would contradict the fact that $B_S \in BC_S$. Hence $|T| \geq 2$. Then the proposal is accepted *only if* for *all* agents $j \in T$,$j \neq i$, it is the case that $\bar{p}_j^j(T, \boldsymbol{d}_T, \beta_T) \geq \bar{p}_j^j(B_S)$ (since they follow σ^*). However, if this is the case, then we have found a coalition, namely T, along with a demand vector and an action such that agent i

[16] We must assume the non-emptiness of BC_S in order to define this strategy at any bargaining round.

believes he is strictly better off and no other agent believes that he is worse off. This contradicts the fact that $B_S \in BC_S$, hence the proposal is never accepted and i cannot gain from such a deviation.

Consider now the case that $\bar{p}_i^i(T, \boldsymbol{d}_T, \beta_T) \leq \bar{p}_i^i(B_S)$. Agent i cannot gain from such a deviation either. If the proposal is accepted he does not gain more than his payoff under σ_i^*. If the proposal is rejected, then the game moves to the next round without any coalition forming. In the next round, if the proposer is some other member of C, then the proposal for i will be $\langle C, \boldsymbol{d}_C, \alpha_C \rangle$, which does not give him a better payoff than $\bar{p}_i^i(B_S)$. If i is chosen to be a proposer again, then we already know that he cannot propose a coalition that gives him better payoff. Now suppose the chosen proposer, say j, does not belong to C and let $C_j \subseteq S$ be the coalition of B_S that j belongs to. Since every player apart from i follows σ^*, j will propose C_j which will be accepted and the game will move to the next round where the set of players is $S \setminus C_j$. By Fact 1 the core element $B_{S \setminus C_j}$ for this subgame still contains $\langle C, \boldsymbol{d}_C, \alpha_C \rangle$. Hence by repeating the above arguments, i cannot gain more than $\bar{p}_i^i(B_S)$. Therefore, whenever nature i becomes a proposer, i cannot gain a better payoff than the payoff he obtains if he follows σ^*.

Assume now that i is a *responder* to some offer $\langle T, \boldsymbol{d}_T, \beta_T \rangle$ and the current set of active players is S. Suppose that all the agents in T that responded before i have already accepted the proposal and let U be the set of agents who are to decide after agent i.

Case 1: $\bar{p}_i^i(T, \boldsymbol{d}_T, \beta_T) \geq \bar{p}_i^i(B_S)$. Then according to σ^* agent i should accept the proposal. If i deviates from σ^* and rejects the proposal then there are two subcases to consider. If all agents in U are going to accept the proposal then i would receive a payoff of at least $\bar{p}_i^i(B_S)$ had he followed σ^*. Since he rejected the proposal, no coalition forms and the game goes to the next round. In the next round either he is the proposer, in which case we know by the previous arguments that he cannot gain more than $\bar{p}_i^i(B_S)$ or someone else is the proposer in which case again, using Fact 1, he cannot gain more than $\bar{p}_i^i(B_S)$ because all other agents follow σ^*. If on the other hand some agent in U will reject the proposal then it does not matter whether i accepts or rejects. The game moves to the next round and agent i cannot obtain a payoff better than the payoff under σ^*.

Case 2: $\bar{p}_i^i(T, \boldsymbol{d}_T, \beta_T) < \bar{p}_i^i(B_S)$. Then according to σ^* agent i should reject the proposal. If i deviates from σ^* and accepts the proposal then if all agents in U also accept, the coalition T forms and agent i receives a payoff which is less than $\bar{p}_i^i(B_S)$. However, if he had followed σ^*, the proposal would have been rejected and in the future he would have obtained $\bar{p}_i^i(B_S)$.[17] If

[17] Notice that this argument holds only for $\delta = 1$.

some agent in U will reject the proposal then as in Case 1, i cannot profit by the deviation from σ^*.

Overall, agent i cannot benefit by any deviation from σ_i^* and, thus, σ^* is a *SEFB* of the corresponding coalitional bargaining game G. □

Observation 7. *Proposition 6 remains true if we use the weak BC instead of the strict one.*

In that case, we only have to modify the strategy σ^* accordingly. Specifically, the strategy σ_i^* for a *responder* i has to become: "i accepts *iff*: either $\langle T, \boldsymbol{d}_T, \alpha_T \rangle \in B_S$ or $\bar{p}_i^i(\langle T, \boldsymbol{d}_T, \alpha_T \rangle) > \bar{p}_i^i(B_S)$". The proof then is completely similar with the one above (and thus we omit it).

Now, for the reverse direction (i.e., "can an *SEFB* give rise to a configuration that belongs to the Bayesian core?"), we cannot hope to always have a positive answer since the Bayesian core does not always exist. However we can provide a positive answer if the bargaining game possesses equilibria whose outcomes do not depend on the random choice of the proposers. The following definition is a generalization of the one given in [MW95].

Definition 27. *An SEFB equilibrium in pure strategies is* order independent *if, whenever it is played, it leads to the same outcome $\langle CS, \boldsymbol{d}, \boldsymbol{a} \rangle$ regardless of the choice of proposers at any round.*

Note that the equilibrium defined in the proof of Proposition 6 is also order independent. In the following result, we show that order independent equilibria lead to outcomes that belong to the weak BC. Intuitively, the order independence property is important for this result to hold because, if no order independent equilibrium exists then it is implied that a change in the order of proposers can lead to outcomes among which the payoff to any agent varies substantially—due to the difference in values of coalitions formed in each outcome, and due to discounting. Thus it cannot be guaranteed that there exists a specific outcome that makes all agents better off: indeed, in that case there might exist sets of agents that are better off in different equilibria outcomes, and the core will be empty. However, assuming order independence we can prove the following[18]:

Proposition 7. *Let σ^* be an order independent SEFB equilibrium strategy profile in pure strategies for a BCBG G with random proposers and discount factor δ. Then, the outcome of σ^*, $\langle CS, \boldsymbol{d}, \boldsymbol{a} \rangle$ must be in the* weak *Bayesian core of the corresponding BCFP.*

[18] We are not aware at the moment if this second result is true for the strict or the strong BC.

Proof: Let $\langle CS, \boldsymbol{d}, \boldsymbol{a}\rangle$ be the outcome of the game if the equilibrium σ^* is played. Assume, contrary to the proposition, that $\langle CS, \boldsymbol{d}, \boldsymbol{a}\rangle$ is *not* in the weak Bayesian core. Let $\bar{p}_i^i(\sigma^*; t=1)$ denote i's expected payoff under σ^* (i.e., if everybody follows σ^* right from the first round). Since $\langle CS, \boldsymbol{d}, \boldsymbol{a}\rangle$ derived by σ^* is *not* in the weak Bayesian core, there exists a coalition $S \subseteq N$, a demand vector $\boldsymbol{d}_S$ and an action α_S such that:

$$\bar{p}_j^j(S, \boldsymbol{d}_S, \alpha_S) > \bar{p}_j^j(\sigma^*; t=1) \; \forall j \in S \tag{5.17}$$

Consider now an agent $i \in S$, and consider the following strategy for i:
(i) If i is chosen by nature to be the *first* proposer, then i proposes $\langle S, \boldsymbol{d}_S, \alpha_S\rangle$.
(ii) In all other cases, i follows σ_i^*.

We will show that this deviation from σ^* benefits agent $i \in S$ in expectation when the other agents play according to σ^*, and therefore σ^* cannot be an *SEFB* equilibrium.

Assume that i was chosen by nature to be the first proposer. Then, i proposes $\langle S, \boldsymbol{d}_S, \alpha_S\rangle$ with the above property. Note that $|S| \geq 2$, otherwise we would already have a contradiction to σ^* being an equilibrium.

All other agents $j \in S$ follow their σ^*-equilibrium strategies. Consider a responder $j \in S$ and consider the subgame that starts at the node where j is to decide whether to accept or reject the proposal and assume every other agent in $S \setminus \{i, j\}$ has accepted. Note that from this point onwards every agent (including i) plays according to σ^*, which is an equilibrium for this subgame (since σ^* is a sequential equilibrium). We will show that *it is optimal* for j to accept.

If j rejects the proposal, then the game moves to round 2 where all agents are present and from then on they all play σ^*. Since σ^* is order independent, the configuration $\langle CS, \boldsymbol{d}, \boldsymbol{a}\rangle$ will form and therefore agent j can get a payoff of at most $\bar{p}_j^j(\sigma^*; t=1)$ (possibly discounted). On the other hand if j accepts the proposal he obtains a better payoff by (5.17). Hence rejecting the proposal cannot be optimal for j. By backward induction, the proposal of agent i must be accepted by all agents of S, therefore the coalition S will form and i will obtain a better payoff. This implies that σ^* is not an equilibrium, a contradiction. □

Remark 1. *Note that in Proposition 7 we allow the bargaining game to have an arbitrary discount factor $\delta \leq 1$, whereas in the game defined in the proof of Proposition 6 we did not allow any discounting ($\delta = 1$).*

A consequence of Proposition 7 is the following:

Corollary 1 (Sufficient condition for the existence of the weak Bayesian core). *If an order independent SEFB equilibrium strategy profile exists, then the* weak BC *cannot be empty.*

Thus, as a corollary to Proposition 7, we managed to provide *a condition for the existence of the weak Bayesian core*.

In summary, in this section we established strong connections between the Bayesian core, a cooperative solution concept for coalition formation under type uncertainty, and non-cooperative equilibrium solutions of the corresponding bargaining games. We proved that if the BC of a coalitional game is non-empty, then there exists an equilibrium of the corresponding bargaining game that induces an element of the BC; and we showed that if an order independent coalitional bargaining equilibrium exists, then it leads to a BC configuration. Those results imply that the use of the Bayesian core as a cooperative stability concept for coalition formation under uncertainty is further justified from a non-cooperative point of view.

5.6 Experimental Evaluation

To evaluate our approach, we conducted several experiments in three settings, two of them with 5 agents and one of them with 8 agents. Agents repeatedly engage in repeated coalition formation activities: that is, they repeatedly engage in *episodes* of coalitional bargaining, each episode consisting of a number of negotiation rounds. When an episode ends, the agents break away from their formed coalitions and the process is repeated: one experimental run consists of several episodes.

In a nutshell, the purpose of our experiments was to demonstrate that updating beliefs during bargaining, and reasoning about the potential strategies of others—as our algorithm does—is potentially beneficial to the agents. The iterative bargaining settings we use help demonstrate that our algorithm helps the agents take rewarding decisions, by facilitating learning of the partners' types during negotiations, and that it can be combined with belief updates after observing the results of coalitional actions (in reinforcement learning style). Below, we elaborate further on the conclusions drawn from our experiments.

During an episode, agents progressively build a coalition structure and agree on a payment allocation. The action executed by a coalition at the end of an episode results in one of three possible stochastic outcomes $o \in O = \{0, 1, 2\}$, each of differing value (or reward). The probability of each outcome occuring depends on the coalitional type vector $\boldsymbol{t}_C$ and the coalitional action α_C taken: $Pr(o|\alpha_C, \boldsymbol{t}_C)$, according to the Bayesian coalition formation model

suggested in Section 4.2. For simplicity, in our experiments each agent's type determines its "quality points", and the "quality" $q(\boldsymbol{t}_C)$ of a coalition with members of types $\boldsymbol{t}_C$ is dictated by the quality points of those members (we elaborate on the particulars of each setting below). Coalition quality then determines the odds of realizing a specific outcome (higher quality coalitions having higher probability of achieving more valuable outcomes).

As described in Section 4.2, the value of a coalition given member types is the expected value w.r.t. the distribution over outcomes (Eq. 4.1; and is estimated by each agent under uncertainty by Eq. 4.2). In some of our experiments, agents are allowed to observe the outcomes of the coalitional actions α_C taken by each formed coalition C, and then update their beliefs using Bayes rule, as follows: $\mu_i'(\boldsymbol{t}_C) = \mu_i(\boldsymbol{t}_C|o, \alpha_C) = zPr(o|\alpha, \boldsymbol{t}_C)\mu_i(\boldsymbol{t}_C)$; μ_i represent the beliefs of agent i regarding types t_C of agents in C, with z being a normalizing factor. However, since the focus here was not on choosing a coalitional action but on focusing on the bargaining dynamics, in these experiments we considered only one coalitional action per coalition (for simplicity—we note, however, that we do deal with the problem of choosing coalitional actions more explicitly in the next chapter of this thesis).

We compare our heuristic Bayesian bargaining equilibrium approximation method (*BE*) with *KST*, an algorithm inspired by a method presented by Kraus et al. in [KST04] (we briefly presented the method in Section 5.1). Though it is better tailored to other settings, focusing on social welfare maximization, it is a rare example of a successfully tested discounted coalitional bargaining method under some restricted form of uncertainty, which combines heuristics with principled game theoretic techniques. It essentially calculates an approximation of a *K*ernel-*ST*able allocation for coalitions that form in each negotiation round with agents intentionally compromising part of their payoff in order to form coalitions. Like [KST04], our *KST* uses a compromise factor of 0.8, but we assume no central authority, only one agent proposing per round, and coalition values estimated given type uncertainty.

As we will see, BE agents manage to do a better job in selecting partners than KST. Interestingly, though the BE focus is on individual rationality (i.e., on serving the goal of each agent to maximize its individual gains through bargaining), it does well in terms of team rationality also. BE agents form more rewarding coalitions than KST agents, and, in addition, the way they divide the payoffs better reflects the "power structure" within each setup: "stronger" agents do not tend to be outwitted by less powerful agents (*"a player's* power *is his ability to help or hurt any set of players by agreeing to cooperate with them or refusing to do so"* [Mye91]). Further, using RL-style updates of beliefs (following the observation of the results of coalitional actions), usually proves to be helpful to the agents (even if not in all settings, since the

stochasticity underlying the observed outcomes can sometimes be, understandably, misleading). Finally, we will demonstrate that BE agents tend to strike a better balance when dealing with the exploration-exploitation problem while bargaining: they sometimes choose to settle with suboptimal coalitions when they believe that it is too risky to try and form more rewarding ones. By doing so, they avoid being exploited, and overall manage to make decisions that are better (both for the individual and for the team) than those taken by KST agents—even though the latter *do* employ compromise explicitly in their deliberations.

5.6.1 Experiments with 5 Agents

Here, we present two experimental settings, each with 5 participating agents and 5 possible types per agent. The quality points of the agents' types is as shown in Table 5.1(a).The quality of the coalition is given as the sum of the quality points of its members.

In the first setting, singleton coalitions receive a penalty of -1 quality points. The parameters of this setting are as follows. Outcome state probabilities are:

- $P(o = 2|q(\boldsymbol{t}_C)) = 0.01q(\boldsymbol{t}_C)$;
- $P(o = 1|q(\boldsymbol{t}_C)) = 0.03q(\boldsymbol{t}_C)$; and
- $P(o = 0|q(\boldsymbol{t}_C)) = 1 - P(o = 2|q(\boldsymbol{t}_C)) - P(o = 1|q(\boldsymbol{t}_C))$.

The rewards corresponding to outcome states are: $R(o = 2) = 1000$; $R(o = 1) = 100$; $R(o = 0) = 1$. Note that agent 0 (of type 0) is detrimental to any coalition.

In our first setting, singleton coalitions receive a penalty of -1 quality points (e.g., $\langle 4 \rangle$ has a quality of 3). We compare BE and KST under various learning models by measuring average total reward garnered by all coalitions in 30 runs of 500 formation episodes each, with a limit of 10 bargaining rounds per episode and a bargaining discount factor of $\delta = 0.9$. We also compare average reward to the reward that can be attained using the optimal, fixed "kernel-stable" coalition structure $\{\langle 1, 2, 3, 4 \rangle, \langle 0 \rangle\}$.

We compared BE and KST with agents either updating their prior regarding types' of partners after observing the coalitional actions—thus learning by reinforcement (*RL* versions) after each episode—or not (*No RL*). In all cases, BE agents update their beliefs after observing the bargaining actions of others during each negotiation round. There are 388 proposals a BE agent considers when negotiating in a stage with all five agents present (fewer in other cases).

Table 5.1(b), showing performance when each agent has a uniform prior regarding the types of others, indicates that the BE algorithm consistently outperforms KST, even though KST is

Agent	Type	Qual
0	0	-1
1	1	1
2	2	2
3	3	3
4	4	4

(a) The 5 participants, their types and quality points

Method	Reward
"Optimal" CS	65800 (expected)
KST-Uni-NoRL	$32521.3(\mathbf{49.4}\%)$
KST-Uni-RL	$44274.4(\mathbf{67.3}\%)$
BE-Uni-NoRL SS=20, LA=3	$60037.7(\mathbf{91.2}\%)$
BE-Uni-RL SS=20, LA=3	$57775.4(\mathbf{87.8}\%)$
BE-Uni-NoRL SS=10, LA=2	$61444.3(\mathbf{93.4}\%)$
BE–Uni-RL SS=10, LA=2	$60086.7(\mathbf{91.3}\%)$
BE-Uni-NoRL SS=3, LA=2	$61269(\mathbf{93.1}\%)$
BE-Uni-RL SS=3, LA=2	$60301.1(\mathbf{91.6}\%)$

(b) Setting A

Method	Reward
"Optimal" CS	33890 (expected)
KST-Uni-NoRL	$20201.4(\mathbf{59.6}\%)$
KST-Uni-RL	$20157.7(\mathbf{59.5}\%)$
BE-Uni-NoRL	$31762.1(\mathbf{93.7}\%)$
BE-Uni-RL	$32275.9(\mathbf{95.2}\%)$
KST-Mis-NoRL	$20193.2(\mathbf{59.6}\%)$
KST-Mis-RL	$21642.5(\mathbf{63.9}\%)$
BE-Mis-NoRL	$31716.6(\mathbf{93.5}\%)$
BE-Mis-RL	$32293.7(\mathbf{95.3}\%)$
KST-Inf-NoRL	$22241.5(\mathbf{65.6}\%)$
KST-Inf-RL	$24748.1(\mathbf{73}\%)$
BE-Inf-NoRL	$31688.3(\mathbf{93.3}\%)$
BE-Inf-RL	$32401(\mathbf{95.6}\%)$

(c) Setting B; (BE uses SS=10, LA=2)

Table 5.1: Total accumulated reward (averaged over 30 runs). "SS":sample size used; "LA":lookahead; "Uni": uniform, "Mis": misinformed, "Inf": informed prior.

designed to promote social welfare (i.e., well aligned with total reward criterion) rather than individual rationality. With respect to individual rationality, KST agents without RL always converge to the coalition structure $\{\langle 4\rangle, \langle 3\rangle, \langle 2\rangle, \langle 0, 1\rangle\}$; this is due to the fact that they are discouraged from cooperating due to the lack of information about their counterparts. When KST agents learn from observed actions after each episode (KST-Uni-RL) they form the coalitions $\{\langle 2, 3, 4\rangle, \langle 0\rangle, \langle 1\rangle\}$ in the last episode of 16/30 runs. BE agents, in contrast, form coalitions based on their evolving beliefs about others, and do not form the optimal kernel-stable structure $\{\langle 1, 2, 3, 4\rangle, \langle 0\rangle\}$.[19] Rather they tend to form coalitions of 2 or 3 members which exclude agent 0 from being their partner. In addition, payoff division for BE agents is more aligned with individual rationality than it is with KST. The shares of (averaged) total payoff of KST-Uni-RL agents 0–4 are $0.8\%, 0.7\%, 28.8\%, 29.6\%, 40.1\%$, respectively, while for BE-Uni-RL (SS:10, LA:2) they are $1.3\%, 13.4\%, 18.8\%, 29.5\%, 37\%$; this more accurately reflects the power [Mye91] of the agents. Moreover, BE results are reasonably robust with changing sample size and lookahead value (at least in this environment with 3125 possible type vectors in a 5-agent coalition).

We attribute the poor performance of KST agents to the fact that they make their proposals without in any way taking into consideration the changing beliefs of others. With the beliefs of the agents varying, negotiations drag (up to the maximum of 10 rounds) due to refusals, resulting in reduced payoffs. BE agents do not suffer from this problem, since they keep track of all possible partners' updated beliefs, and use them during negotiation. Thus, they typically form a coalition structure within the first four rounds of an episode.

We also experimented with a second setting where the rewards and transition parameters are as in the first, but in which singleton coalitions receive a penalty of -2 quality points (rather than -1 above), and where $q(\boldsymbol{t}_C) = \sum_{t_i \in \boldsymbol{t}_C} q(t_i)/|C|$ (as coalitions get bigger they get penalized to reflect coordination difficulties). This setting makes the quality of coalitions a bit more difficult to distinguish. Here, a near-optimal (kernel-stable) configuration contains the structure $\{\langle 4, 3\rangle, \langle 2, 1\rangle, \langle 0\rangle\}$. We use three different priors: *uniform*, *misinformed* (agents have an initial belief of 0.8 that an agent with type t has type $t + 2$), and *informed* (belief 0.8 in the true type of each other agent).

The results (Table 5.1(c)) indicate that KST agents again do not do very well, engaging in long negotiations due to unaccounted for differences in beliefs among the various agents. KST-Uni-RL agents for example typically use all ten bargaining rounds available; in contrast,

[19] Nor should they, as this is not necessarily in their best interests, given the bargaining horizon and discount factor—the kernel and other stability concepts do not take the bargaining dynamics into account.

BE-Uni-RL usually form structures within 3 rounds. Even when KST uses informed priors (which resembles most the situations dealt with in [KST04]), the fact that the expected value of coalitions is not common knowledge (as was the assumption in [KST04]) takes its toll. BE agents, on the other hand, derive the true types of their partners with certainty in all experiments, and typically form profitable configurations with structures such as $\{\langle 4,3\rangle, \langle 2,1\rangle, \langle 0\rangle\}$ or $\{\langle 4,2\rangle, \langle 3,1\rangle, \langle 0\rangle\}$. We can also see in this setting that RL is, at least slightly, enhancing the performance of BE agents, helping them differentiate among the quality of the various partners: while it may sometimes be possible that RL blurs the picture a bit due to occassional unlikely payoffs received,[20] if it is relatively easier to distinguish the agents' types during bargaining (as was in the first setting), it is otherwise helpful, as demonstrated by Setting B results.

5.6.2 A Coalitional Climbing Game

We also developed a third experimental setting which helps us better demonstrate how BE agents are more qualified to deal with the exploration-exploitation tradeoff when bargaining under uncertainty. Specifically, we aimed at demonstrating that BE agents are rational, cautious negotiators while bargaining to form coalitions: if they are uncertain, or do not "believe" that they will have the time to become certain about others (i.e., if the bargaining discount factor is low and thus the agreements' value drops fast) they will opt to accept suboptimal agreements. In order to make the point, we developed an experimental setting that, as we will see, bears many resemblances to the *climbing game* examined in Section 3.5: the setting makes discovery of opponent types difficult, and thus rational agents should settle for suboptimal agreements (but, hopefully, they will be using knowledge gained to achieve better ones in the future). Though our agents often opt for suboptimal coalitional configurations, they still (once again) manage to outperform—both in terms of total coalitional accumulated reward (social welfare) and individual rationality—agents that do not update beliefs about others or do not take into account others' beliefs while negotiating.

[20] In bargaining, agents act (negotiate) exactly according to their type. Thus, belief updates based only on observed negotiation behaviour can work well. However, when belief updates follow the results of coalitional actions, the picture may in fact be clouded due to unlikely outcomes observed. Thus, RL is not always necessarily helpful when combined with belief updating during bargaining (or, at least, it may not always help make a difference).

The Setting Specifics

We now present the setting in detail. The general setup characteristics are as follows: There are 8 agents, with 2 types per agent (there are 255 possible coalitions with 8 agents present; from an agent's perspective, there are 128 expected possible type vectors in a 8-agent coalition—since agents do not know types of opponents or how many opponents are of each type); and there are 2841 bargaining actions available to proposer when all 8 agents present. Four of the agents are of type A, and four are of type B; therefore, the agents in the setting can form 24 possible coalitions of "distinct quality". Each experimental run consists of 250 episodes, with 8 bargaining rounds per episode. The lookahead value (LA) used was 2, and the sample size (SS) was 5. We experimented with uniform and informed priors, both when using RL style updates of beliefs or not (NoRL), and we also tried various values for the bargaining discount factor (0.95, 0.75 and 0.5).

We assumed 2 possible types per agent: type A, with "quality" $q_A = 2$ (the "strong" type), and type B with $q_B = 1$ (the "weak" type). The "coalitional quality" $q(\boldsymbol{t}_C)$ of any coalition C (with type vector $\boldsymbol{t}_C$) in this setting is, then, defined as described in Table 5.2. The rewards associated with the outcome states $o \in O = \{0, 1, 2\}$ were $R(o = 2) = 100$, $R(o = 1) = 10$ and $R(o = 0) = 1$, while the probabilities of reaching those states following the execution of the (single) coalitional action were $Pr(o = 2|q(\boldsymbol{t}_C)) = 0.01 * q(\boldsymbol{t}_C)$, $Pr(o = 1|q(\boldsymbol{t}_C)) = 0.02 * q(\boldsymbol{t}_C)$ and $Pr(o = 0|q(\boldsymbol{t}_C)) = 1 - Pr(o = 2|q(\boldsymbol{t}_C)) - Pr(o = 1|q(\boldsymbol{t}_C))$; *unless* $q(\boldsymbol{t}) = 30$, in which case $Pr(o = 2|q(\boldsymbol{t}_C) = 30) = 0.65$, $Pr(o = 1|q(\boldsymbol{t}_C) = 30) = 0.35$, and $Pr(o = 0|q(\boldsymbol{t}_C) = 30) = 0$.

Coalitions	$q(\boldsymbol{t}_C)$
Both types A and B present	$q(\boldsymbol{t}_C) = 0$ *unless* $C = \langle AABB \rangle$: then $q(\boldsymbol{t}_C) = 30$ *or* $C = \langle AAB \rangle$: then $q(\boldsymbol{t}_C) = 26$ *or* $C = \langle AB \rangle$: then $q(\boldsymbol{t}_C) = 5$
Only $X = A$ or $X = B$ present in non-singleton	$q(\boldsymbol{t}_C) = (\|C\| - 1) + \sum_{i \in C} q_X$
Singleton coalitions	$q(\boldsymbol{t}_C) = 0$

Table 5.2: Coalitional quality functions for the coalitional climbing game. Coalition $C = \langle AABB \rangle$ is the coalition with the maximum quality. The quality points for the rest of the coalitions are such that they can serve as "stepping stones" for the agents to progressively discover the better coalitions, and encourage cooperation of agents of different types.

The use of the above coalitional quality, reward, and outcome transition functions, has the following implications (after the necessary calculations):

- The expected value of a coalition C of type A agents is given by the function $V(C_{t:A}) = 3.51|C| - 0.17$; the expected value of an agent participating in such a coalition is given by the function $V(C_{t:A}) = 3.51 - (0.17/|C|)$.

- The expected value of a coalition C of type B agents is given by the function $V(C_{t:B}) = 2.34|C| - 0.17$; the expected value of an agent participating in such a coalition is given by the function $V(C_{t:B}) = 2.34 - (0.17/|C|)$.

- The singleton (reservation) value of an agent is 1.

- The expected value of coalition $\langle AAAA \rangle$ (with quality 11) is 13.87 and the expected value of each of its members is 3.4675.

- The expected value of coalition $\langle BBBB \rangle$ (with quality 7) is 9.19 and the expected value of each of its members is 2.2975.

- Thus, the *relative power* of agent types A/B is $3.4675/2.2975$, meaning that agent A is approximately 1.5 times stronger than agent B. (An A agent would expect to do 1.5 times better than a B agent, by cooperating only with A agents.)

- The expected value of $\langle AABB \rangle$ is 41.5. This is the most rewarding coalition in this setting. (Actually, assuming fully informed agents, $\langle 18.4525, 18.4525, 2.2975, 2.2975 \rangle$ is a *core* payoff allocation for this coalition: the core configuration is forming the coalition structure $\{\langle AABB \rangle, \langle AABB \rangle\}$, with the above payoff allocation for each coalition.)

- The expected value of coalition $\langle AAB \rangle$ is 31.42. This coalition is quite rewarding, but not as much as the $\langle AABB \rangle$. However, since it has 3 members instead of 4, $\langle AAB \rangle$ is easier for the agents to "discover". In $\langle AAB \rangle$, the allocation most prefered by the two type A agents—without them outweighing one another—is $\langle 14.5612, 14.5612, 2.2975 \rangle$.

- The expected value of coalition $\langle AB \rangle$ is 6.85.

Expected Behaviour of the Agents

The setup encourages agents to discover the types of others in order to form rewarding coalitions. However, rational agents ought to be cautious, since, in the presence of uncertainty, seeking the most beneficial agreement may result in receiving a payoff that is even less than the singleton reservation value. If agents are uncertain about others, then they should try to

form (suboptimal) coalitions that appear to be successful and stick to them in order to avoid the "implicit" penalties.

We expected BE agents to exhibit this "cautious" behaviour when negotiating, until they become more informed about others. This should be even more obvious as the bargaining discount factor drops. (If the bargaining discount factor is low, meaning that the value of coalitional agreements decreases rapidly, agents are forced to take decisions early.) BE agents should form suboptimal coalitions such as $\langle AB \rangle$, $\langle AAB \rangle$, $\langle AAAA \rangle$, $\langle BBBB \rangle$, or other "single-type" coalitions—if they come to have strong beliefs about the types of the corresponding agents. The coalition $\langle AAB \rangle$ is particularly likely to be formed, since it is easier for an agent to come to hold strong beliefs about two other agents (rather than three or more others)—this is a small and quite rewarding coalition to form.

In contrast, we expected KST agents to fall prey to the penalties, doing worse both in terms of social welfare and individual rationality. This is because KST agents would attempt to form the coalitions they consider rewarding, but since the agents' estimates for the values differ, negotiations are expected to drag. Even if "informed" priors are to be used, this will be posing a problem, and, since the best tool KST has to resolve disagreements in the face of uncertainty is "compromise", the individual earnings of the agents will suffer.[21]

The use of RL (i.e., belief updates after observing the outcomes of the actions of formed coalitions) was not expected to make a huge difference in this setting. This is because the expected values of the various coalitions, apart from the two most rewarding ones, are not so different from each other.

Results

The results for this setting are shown in Tables 5.3 and 5.4. We remind the reader that $NoRL$ means that no belief updates are performed after the execution of coalitional actions (but belief updates *are* performed after observing other agents taking bargaining actions during negotiations).

The metrics we used, and which we report on, were the following: first, the *total reward accumulated by coalitions* (averaged over all runs); second, the *frequency of appearance of an "optimal coalitional structure" or of other profitable structures*; third, the *expected value of Q of formation decisions*: this is a metric of *bargaining decision quality*, and also reflects the

[21] There would not be much point in trying to use KST *without* compromise, since in that case negotiations would be certain to last even longer, resulting to very low team and individual payoff.

Method	*Reward*	$Q = \sum_C f_C V(C)$	A/B
KST-NoRL-0.95	2960.59	2.15383	1.17
BE-NoRL-0.95	**4793.87**	**3.7698**	**1.71**
KST-NoRL-0.75	1654.86	2.52944	1.42
BE-NoRL-0.75	**2846.4**	**5.21792**	**2.07**
KST-NoRL-0.5	2808.56	6.88	4.26
BE-NoRL-0.5	**1450.88**	**8.4**	**1.5**
KST-RL-0.95	2999.89	2.20384	2.25
BE-RL-0.95	**5321.04**	**4.83322**	**1.7**
KST-RL-0.75	1776.29	2.46283	5.62
BE-RL-0.75	**2725.96**	**4.68843**	**2.01**
KST-RL-0.5	1467.72	2.96	3.15
BE-RL-0.5	**1916.95**	**9.96**	**1.43**

Table 5.3: Setting C results - Uniform Priors; BE uses SS=5, LA=2.

convergence of beliefs. It is given by

$$Q = \sum_C f_C V(C)$$

where f_C is the *average* frequency with which C forms in the experimental runs (or, more intuitively, its frequency of appearance in an "average run"), and $V(C)$ is the (expected, since the outcomes are stochastic) value of the coalition; and, finally, the *observed relative power* of type A over B (the actual payoff of A's over B's). We use this last metric as a further measure of individual rationality, as we explain here:

The *relative power* A/B is the expected payoff of A in coalitions excluding B, over the expected payoff of B in coalitions without A—as mentioned above, its value is approximately 1.5. Agents that enter negotiations *without* much knowledge about opponents should be observed to be doing close to that value, so that it is not the case that one type "exploits" the other. However, as agents gain knowledge, and manage to take coalition formation decisions which are closer to the the optimal ones, the significance of this specific measure as a measure of individual rationality is reduced: it may be well to the interest of informed players to form coalitions with others, even if this means that the payoffs of others will exceed theirs many times. Nevertheless, if the agents are equally uninformed, it shouldn't be the case that stronger types are able to outweigh weaker ones more than what their relative power indicates.

As mentioned, the setting makes discovery of opponent types difficult, and thus rational agents should settle for suboptimal coalitions (hopefully using them as stepping stones to form

Method	*Reward*	$Q = \sum_C f_C V(C)$	A/B
KST-NoRL-0.95	1353.54	1.01238	1.02
BE-NoRL-0.95	**5200.94**	**4.20369**	**1.63**
KST-NoRL-0.75	271.263	4.20369	1.34
BE-NoRL-0.75	**4123.24**	**7.85009**	**1.42**
KST-NoRL-0.5	1860.23	4.32893	3.02
BE-NoRL-0.5	**6189.38**	**20.0437**	**0.99**
KST-RL-0.95	2051.47	1.50318	1.87
BE-RL-0.95	**6322.38**	**5.4451**	**1.54**
KST-RL-0.75	1719.11	2.42433	5.94
BE-RL-0.75	**5554.52**	**7.95062**	**1.39**
KST-RL-0.5	1767.81	4.12974	3.05
BE-RL-0.5	**2793.71**	**11.5523**	**1.16**

Table 5.4: Setting C results - Informed Priors. BE uses SS=5, LA=2.

better ones later). Results in the *uniform priors* (Table 5.3) case show that BE agents outperform KST agents both in terms of social welfare and individual rationality and that RL updates are in general beneficial. Further, lowering the discount factor to 0.75 and to 0.5 forces the agents to form coalitions earlier, but also contributes to better decisions—in terms of Q; the actual reward coalitions get is, naturally, lower—because it enables the agents to discover the types of opponents with more accuracy, effectively reducing the number of possible opponent responses during bargaining (intuitively, given more time, both a "strong" and a "weak" type might refuse a proposal, while if time is pressing the "weak" might be the *only* one to accept).

The only time that KST agents seem to be doing relatively well in quality decisions is the *KST-NoRL-0.5* case. We believe that was the case because, in that scenario, the "pressure" put on the agents to make early decisions happened to fit well with the KST's "compromise" approach (we remind the reader that the KST agents are willing to settle for an 80% fraction of their best anticipated payoff when proposing or accepting agreements). Notice, however, that this relatively good performance with respect to Q is achieved to the expense of individual rationality, in the sense that type A agents managed to "exploit" the type B ones: A/B was 4.26, with the agents forming an $\langle AB \rangle$ coalition 37% of the time on an average run (the average percentage of forming $\langle AB \rangle$ in a run was 37%), while the $\langle BB \rangle$ or $\langle BBB \rangle$ coalitions (which, if the agents were informed would have guaranteed the same expected payoff with $\langle AB \rangle$ for B agents) were created only 8% and 4% of the time, respectively.

The results in the *informed priors* case (Table 5.4, where agents know the type of opponents with 85% accuracy) were similar to the ones in the uniform priors one. Once again, BE

agents are doing a lot better than KST ones. As expected, the quality of their decisions is improved as the discount factor falls (for much the same reasons as in the uniform priors case). Also, the general performance of informed BE agents is clearly better than their performance when uninformed. Notice however that, interestingly, the KST agents have not benefited from being informed, specially when the discount factor is high. Our interpretation is that, since they are informed, they tend to stick more on their ground, and this makes it harder for their "compromise" heuristic to work, at least in this scenario.[22]

Finally, it's worth mentioning the very good performance of BE agents in the (informed) *BE-NoRL-0.5* case. The average percentage of forming the optimal coalition $\langle AABB \rangle$ in a run was 41.9% in this scenario. Notice that type B agents were not at all weak in this setting: they have power, derived by their knowledge, and force the A agents to accept to grant them big shares to form the better coalitions. However, this does not mean that A agents are not doing well either: they do achieve a lot more payoff than they did in the uninformed *BE-NoRL-0.5*, and also in the informed *BE-NoRL-0.75* and *BE-NoRL-0.95* cases—so, the overall quality of decisions is better and rewarding for both types of agents, but their increased confidence in their beliefs forces them to compromise more instead of suffering discounted payoffs.[23] The RL version of this setup, *BE-RL-0.5*, also does well (takes better decisions than *BE-RL-(0.75 & 0.95)*, and *BE-NoRL-(0.75 & 0.95)* counterparts), but is not as successful as *BE-NoRL-0.5*. This is an example of a case where RL was not helpful: even though the agents did possess prior information that was close to being accurate, this information sometimes got "disturbed" due to the occassional unlikely payoffs received.

5.7 Conclusions

In this chapter we provided a Bayesian approach to discounted coalitional bargaining under (type) uncertainty. We defined Bayesian coalitional bargaining games (BCBGs), and described their PBE equilibrium solution. Since computing the PBE solution is intractable, we presented a heuristic, best-response bargaining algorithm, which uses belief monitoring of other agents

[22] We stress that this phenomenon is not due to some error in our implementation of the KST algorithm— when we tested a *fully* informed version of KST agents, they were indeed able to form the optimal coalition.

[23] One perhaps would expect that informed agents would be able to form "core" allocations, which would benefit the A agents; we remind the reader that this is not to be anticipated, as the bargaining dynamics (such as the horizon of negotiations and the discount factor) are *not* accounted for in stability concepts, so it is not realistic to expect that core allocations will be formed as the result of discounted coalitional bargaining, unless specific assumptions regarding the agent beliefs and other factors are in place.

and belief updates following every bargaining round. In repeated coalition formation settings, the algorithm can be combined with RL-style belief updates after the execution of coalitional actions at the end of each formation episode. We verified experimentally that Bayesian coalitional bargaining (using our heuristic approach) enables the agents to take sequentially rational decisions while bargaining, and to come up with agreements that are rewarding both with respect to self and team rationality.

We also contributed a new equilibrium solution concept for coalitional bargaining games, the sequential equilibrium under fixed beliefs (SEFB), which assumes that the agents have beliefs that remain fixed during bargaining. We then provided a non-cooperative justification of the Bayesian core by proving propositions that essentially show that SEFB equilibrium play can lead to outcomes that are in the BC—and, conversely, that if the BC is non-empty then there exists an SEFB equilibrium that produces a BC element. In addition, we established a sufficient condition for the existence of the (weak) Bayesian core.

To the best of our knowledge, the work presented in this chapter is the first to address the problem of discounted coalitional bargaining *under uncertainty*. The Bayesian model we proposed allows dealing with the problem under the generic assumption of type uncertainty, with the resulting benefits for learning explained in this and in the previous chapter. Further, to the best of our knowledge we were the first to relate a cooperative stability concept *under uncertainty* with non-cooperative equilibrium play in coalitional bargaining games, in the spirit of what many researchers in game theory have done in deterministic settings. By doing so, we provided further justification for the adoption of the Bayesian core as a coalitional stability concept under uncertainty.

Chapter 6

Bayesian RL for Coalition Formation under Uncertainty

In this chapter we provide a framework for agents to take sequentially rational decisions, balancing individual and team interests, while engaging in repeated coalition formation activities. Our adopted framework effectively integrates decision making during repeated coalition formation under uncertainty with Bayesian reinforcement learning.

It is often the case that a set of agents may have to engage in *repeated* coalition formation, having the opportunity to engage in a series of coalition formation episodes, each of which is followed by some collective action taken by each coalition formed. This suggests opportunities for agents to *learn* about each others' abilities through repeated interaction, refining how coalitions are formed over time; it also poses the question of how to make decisions that are *sequentially rational*, given the anticipated horizon of formation interactions and the evolution of agents' knowledge.

The agents should be able to make future use of information they gain in this process. Intuitively, the effects of *collective (coalitional)* actions provide information about the capabilities of partners, and agents can use this in order to make decisions and abandon formed coalitions for potentially more profitable ones if they have the chance—and, in most settings, this would be the case.

To account for such considerations, we propose a *reinforcement learning* (RL) model which enables the agents to improve their formation decisions and coalitional decisions through experience gained by repeated interaction with others, and the observation of the effects of coalitional actions. More specifically, we propose a *Bayesian RL* model in which agents maintain and update explicit beliefs about the types of others, and, through this process of pro-

gressive belief refinement, become increasingly able to make rewarding sequentially rational decisions—regarding *both* potential *coalition formation activities* on their part, and potential *choice of actions on behalf of their formed coalitions*. This is natural, since real world coalitions can in general have the opportunity to choose among several reward-producing activities or tasks to take up—for example, the construction agents in our toy example presented in Chapter 1 may have the choice to take up building a block of apartments in Toronto or a skyscraper in downtown Manhattan—and their choice of partners and choice of projects are interdependent.

We make use of *a POMDP formulation similar to the one used for MARL in stochastic games*. This formulation enables the agents *to assess the long-term value of coalition formation decisions*, including the value of potential collective actions. It thus also manages to implicitly trade off individual rationality (i.e., the goal of maximizing one's individual payoffs) with team rationality. The agents using our approach choose actions and coalitions not only for their *immediate* value, but also for their *value of information*, since an action has a value both because it provides the agents with immediate gains and also because it provides them with information about the types of others and the values of potential coalitions. Thus, our formulation enables us to deal with the problem of *optimal repeated coalition formation under uncertainty* (i.e., the problem of taking sequentially rational decisions in repeated coalition formation scenarios).

Further, we stress that our formulation (based as it is on the Bayesian model for coalition formation under type uncertainty that was presented in Chapter 4) enables us to simultaneously deal with uncertainty regarding both the types of others and the outcomes of coalitional actions. In addition, we describe how this RL framework can be combined with the dynamic formation processes presented in Chapter 4.

We demonstrate experimentally that our framework enables agents to make informed, rational decisions that are rewarding *both* in the short and in the long term. This is true, even if they do not converge to stable coalitions in the end of a series of repeated coalition formation episodes.[1] The agents make efficient use of the information at hand, taking decisions that are as informed as possible, and thus balance the need to explore in order to learn with the need to exploit current knowledge effectively.

The focus of the work described in this chapter, and the associated experiments, is mainly on the (online) behaviour of the agents while learning by observing the results of coalitional actions. We will not focus here on the negotiation processes themselves, or the strategic con-

[1] For example, stability cannot be reached unless the agents' beliefs themselves stabilize over time.

siderations of the agents during bargaining (as we do in Chapters 4 and 5 of this dissertation). However, we note that our Bayesian RL formulation is such that it allows for the incorporation of any potential bargaining process to be employed during formation.

The chapter's outline is as follows: We start by describing our Bayesian RL framework (and POMDP formulation) for optimal repeated coalition formation under uncertainty in Section 6.1. In Section 6.2 we present several Bayesian RL algorithms that we developed in order to approximate the solution to the problem's POMDP formulation. Section 6.3 then explains how our RL algorithms can be combined with negotiation processes for coalition formation (including the processes presented in Chapter 4 and Chapter 5). In Section 6.4 we detail the experiments we conducted to evaluate our algorithms, discuss their results, and provide intuitions for further experimentation. Following that, in Section 6.5 we provide a discussion comparing our approach with related work.[2] We argue there that our framework is much more general than existing approaches, as their specifics can be easily incorporated in it, if so desired. Finally, Section 6.6 recaps the main findings of this chapter.

Parts of the research described in this chapter appeared originally in [CB04].

6.1 A Bayesian RL Framework

In realistic settings, agents participating in coalition formation activities will have to face the forms of uncertainty (i.e., type uncertainty and uncertainty regarding coalitional actions and their results) described in Chapter 4. The possibility of repeated interaction, as explained earlier in this dissertation, provides the agents with the ability to *learn*, progressively updating their beliefs about the types of their potential partners. One may ask, of course, if agents are indeed 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 altogether, but furthermore that agents often have no incentive to engage in actions (or interactions) that would reduce this uncertainty.

Therefore, here we formulate a solution for what we call the problem of *optimal repeated coalition formation*: The participating agents are interested in eventually forming efficient, profitable coalitions, but they also want to gather as much reward as possible while doing so. To rephrase this, the problem of *optimal repeated coalition formation*, or *optimal coalitional learning*, is to maximize the lifetime performance of an agent that repeatedly engages

[2] This discussion has to be provided at the end of the chapter, as it assumes the reader's familiarity with the details of our approach.

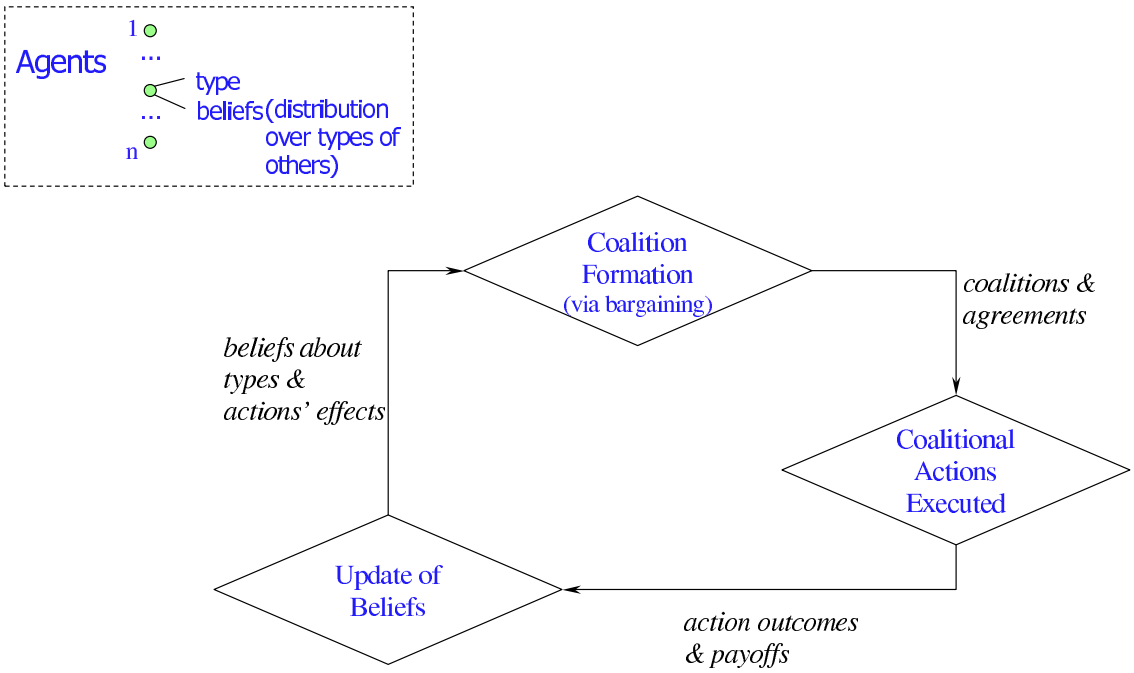


Figure 6.1: A Bayesian RL framework for repeated coalition formation under uncertainty.

in coalition formation activities and receives agreed-upon shares of payoffs arising following the execution of agreed-upon coalitional actions—as specified in agreements reached during the aforementioned coalitional activities (as illustrated in Figure 1.1, which we repeat here as Figure 6.1 for convenience).

So, in this section we describe an RL model in which agents repeatedly form coalitions and take coalitional actions—in accordance with the framework of Figure 6.1. This gives agents the opportunity, through observation of the outcome of coalitional actions, to update their beliefs about the types of their partners. 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*: 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*). For example, if agent i believes that some potential partner j is of lesser ability (and value to i) than its current partner, and if he believes that trying to form a partnership with j will probably not lead to some new and interesting information either, then i should not attempt to form the partnership with j at all. Of course, i may be wrong as j may in fact be an agent of great abilities and i would have been better off had he tried to learn about them; nevertheless, i's optimal course of action, given his "pessimistic" beliefs, is unquestionably not to abandon his current partner for j.

Optimal Repeated Coalition Formation (under Uncertainty)

To lay out the setting for optimal repeated coalition formation, let us first suppose the existence of N participating agents, each having *initial beliefs* B_i. The RL process proceeds in stages (as shown in Figures 6.1 and Fig. 6.2): 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 state s of that action. These are "local" outcome states, depending on the actions and type vectors of specific coalitions, and not on the whole coalition structure reached at the end of the formation process.[3] Recall from Chapter 4 that the model of the domain dynamics $Pr(s|\alpha, \boldsymbol{t}_C)$ is assumed

[3]The model can be extended by allowing the value of any coalitional action (equivalently, the probabilistic model for transitioning to the various outcomes) to depend on the current *state of the game*, such a state consisting of the coalitional configurations (i.e., the complete coalition structure and payoffs allocation) reached and the beliefs of the agents. This would allow for a sequential environment model (an underlying MDP) given a configuration, presumably allowing for the study of coordination games played among the various coalitions present in the coalition structure (without the agents regrouping). We don't consider this possibility here, instead focusing on the sequential nature of repeated coalition formation itself. We note however that a formulation like this may fit well in natural settings requiring the agents to dynamically pick a coordination problem to tackle, as is the case

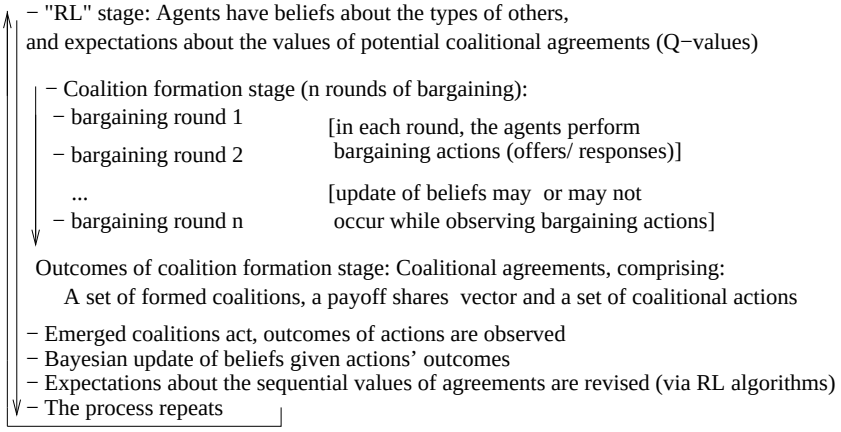


Figure 6.2: The "RL" and the "coalition formation" stages of the Bayesian RL framework for repeated coalition formation under uncertainty.

to be common knowledge, providing agents with the probability of occurence of outcome s given that coalition C with members type vector $\boldsymbol{t}_C$ takes coalitional action α. Each member of coalition C then updates its beliefs about its partners' types:

$$B_i^{t+1}(\boldsymbol{t}_C) = z \Pr(s|\alpha, \boldsymbol{t}_C) B_i^t(\boldsymbol{t}_C) \tag{6.1}$$

where z is a normalizing constant (we sometimes denote the updated belief state as $B_i^{s,\alpha}$). In order to make our model apply to realistic circumstances, and in order to be able to test the full potential of our RL algorithms, in this chapter we assume only *limited* observability of the realized outcomes: the agents only observe the outcome of their own coalition's action. The process then repeats. So, overall, our reinforcement learning process consists of coalition formation games being played (*"coalition formation stage"*) between the execution of coalitional actions, which causes the beliefs to be updated (*"RL stage"*)—as depicted in Figures 6.1 and 6.2. Notice that, even though our general framework allows for this (see for example 6.2), in this chapter we won't consider ourselves with the possibility of the agents updating beliefs during bargaining (i.e., within the coalition formation stage).

We adopt an approach to optimal repeated coalition formation that uses *Bayesian exploration*. As demonstrated in the approach to multiagent RL presented in Chapter 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: Bayesian

in the Robocup Rescue competition [KT01], where agents have to coordinate in order to combat the effects of a major natural disaster that has hit a city. We elaborate more on these ideas in Chapter 7 of this dissertation.

exploration outperforms in expectation any other method having the same prior knowledge.

We cast the problem of optimal coalitional learning as a partially observable MDP (POMDP), or a belief-state MDP. Assuming an infinite horizon problem, with discount factor γ (with $0 \leq \gamma < 1$), 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 $Q_i(C, \alpha, \boldsymbol{d}_C, B_i)$ denote the *long-term value* i places on being a member of coalition C that has agreed action α and demands $\boldsymbol{d}_C$, realizing that after this action is taken, the coalition formation process will repeat. This is accounted for using Bellman-like equations [Bel57] as follows:

$$Q_i(C, \alpha, \boldsymbol{d}_C, B_i) = \sum_s \Pr(s|C, \alpha, B_i)[r_i R(s) + \gamma V_i(B_i^{s,\alpha})] \tag{6.2}$$

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

$$V_i(B_i) = \sum_{C|i\in C, \boldsymbol{d}_C} \Pr(C, \alpha, \boldsymbol{d}_C|B_i) Q_i(C, \alpha, \boldsymbol{d}_C, B_i) \tag{6.3}$$

Recall that $R(s)$ is the reward paid to C for its action resulting to outcome state s, 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 $r_i R(s)$ describes i's reward share when coalitional action α results in s). $V_i(B_i)$ 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 with some Q-value (we elaborate below).

Thus, when the agents find themselves in a new belief state, following the formation of a coalition and the occurence of a coalitional action, they are able to estimate new values reflecting the worth of possible future decisions, and thus become capable of using these new values in future coalitional negotiations. The agents' uncertainty is effectively encapsulated in the belief-state MDP described by Equations 6.2 and 6.3; also, the expected *value of information* of a coalitional agreement is incorporated in those equation, since the agents *recognise—through examining the value of future belief states—the need to examine the potential effect that new information will have on their future decisions*. Specifically, the agents will enter in the negotiation process taking into account *the Q-values* of coalitional agreements, rather than using only immediate expected reward estimates: that is, they incorporate considerations of 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, making

formation moves and suggesting coalitional actions that are maximizers of Eq. 6.2.

Unlike typical Bellman-like equations, the value function V_i cannot be defined by maximizing Q-values. This is because the choice that dictates reward, namely, the coalition that is formed, is not in complete control of agent i. Instead, i must predict, based on it beliefs, the probability $\Pr(C, \alpha, \boldsymbol{d}_C|B_i)$ with which a specific coalition C (to which it belongs) and a corresponding action-demands pair $\langle \alpha, \boldsymbol{d}_C \rangle$ 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, 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.

Thus, we can now make our repeated RL process (depicted in Figures 6.1 and 6.2) more specific by providing the algorithm of Figure 6.3:

1. Each agent i with belief state B_i calculates the Q-value of any potential agreement $\langle C, \alpha, \boldsymbol{d}_C \rangle$ in which it can potentially participate, *by solving Equations 6.2 and 6.3*.
    - In solving Equations 6.2 and 6.3, agents may have to take into account the specifics of the coalition formation process to follow, so that they are able to estimate the probabilities $\Pr(C, \alpha, \boldsymbol{d}_C|B_i)$ in Eq. 6.3.
2. The agents engage in a coalition formation process, in which each one of them is using the Q-values calculated above in order to reflect the long term value of coalitional agreements. The process results in a coalition structure CS, a payoff allocation vector $\boldsymbol{d}$, and a vector of coalitional actions $\boldsymbol{\alpha}$, one for each coalition $C \in CS$.
3. Each agent i observes its own coalition $C \in CS$, and the corresponding $\boldsymbol{d}_C$, α that are the restrictions of $\boldsymbol{d}$, $\boldsymbol{\alpha}$ to this coalition.
4. Each agent i updates beliefs about partners $j \in C$, $j \neq i$, using Eq. 6.1.
5. The RL process repeats.

Figure 6.3: Optimal repeated coalition formation under uncertainty.

Now, returning to the issue of estimating the $\Pr(C, \alpha, \boldsymbol{d}_C|B_i)$ probabilities, we note that they can be approximated in a variety of ways, depending also on the coalition formation algorithm assumed to be in use during the formation stage. If a discounted coalitional bargaining model is assumed, for instance, where coalitions are expected to abandon the negotiations after forming, then the agents could (a) simulate the process of solving the game tree, possibly using a heuristic algorithm such as the one presented in Chapter 5, and then (b) assign probability 1 to ending up in the coalitional configuration emerging as a result of the bargaining tree solution

(or, if $X > 1$ such solutions exist, possibly corresponding to multiple bargaining equilibria, assign probability $1/X$ to each corresponding configuration).

If, however, the negotiations are assumed to be conducted according to a dynamic process that can be modeled by an underlying Markov chain (such as the BRE process introduced in Chapter 4), then these probabilities would be easily calculated if one was able to predict the probability with which the states of the Markov chain containing the $\langle C, \alpha, \boldsymbol{d}_C \rangle$ would arise at the end of the negotiation process. These probabilities correspond to the Markov chain's steady-state distribution probabilities. The steady-state distribution P^* is the solution of the system $P^* = P \cdot P^*$, with P being the Markov chain's transition matrix—which describes the probability of transitioning to each state ω' starting from any state ω (see Table 6.1).

		future states			
		ω_1	ω_2	$\ldots$	ω_n
current states	ω_1	p_{11}	p_{12}	$\ldots$	p_{1n}
	ω_2	p_{21}	p_{22}	$\ldots$	p_{2n}
	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
	ω_n	p_{n1}	p_{n2}	$\ldots$	p_{nn}

Table 6.1: A Markov chain transition matrix. For each row i, $\sum_j p_{ij} = 1$. For a dynamic coalition formation process such as *BRE*, states are of the form $\omega_i = (CS, \boldsymbol{d}, \boldsymbol{\alpha})$.

However, such an approach is inherently problematic in our setting. The Markov chain transition matrix and the steady-state distribution above can be determined by an agent only if the parameters affecting the state transitions are known (such as whether it is expected by someone else to make a specific proposal in the future); however, agent i does not have full knowledge of those parameters, since he is unaware not only of other agents' types, but also of other agents' beliefs. Even the use of a *common prior* to help approximate these beliefs is problematic, as we now explain.

Assume that at each RL time step there exists a common prior, shared by all agents, specifying the probability with which the agent type profiles are drawn; and that the agents use this common prior in order to estimate the probabilities an opponent of a specific revealed type assigns to type profiles of his opponents (i.e., they treat the common prior as *representing the beliefs* of their opponents; these beliefs can then be assumed to stay fixed throughout the coalition formation process). However, the use of such a *static* common prior to account for the beliefs of others at the initiation of a negotiation process is unrealistic and problematic. This is due to the fact that all agents' beliefs get updated after each RL trial: after k RL steps, the

agents will enter the formation process preceding RL step $k+1$ assuming that their opponents hold beliefs adhering to the common prior that was in use before the first RL step. This is clearly an unrealistic assumption, as all agents have updated their beliefs k times since then. Further, it is not possible for the agents to monitor the way the beliefs of others are changing, because, as we explained above, in this chapter we make the more realistic and challenging assumption that agents can observe the outcome of their *own* coalition's action *only*.

For those reasons (and others, which we relate later), we do not try to calculate any Markov chain transition matrices or steady-state distributions in this chapter; rather, when we need to approximate the $\Pr(C, \alpha, \boldsymbol{d}_C|B_i)$ probabilities corresponding to agreements that are outcomes of dynamic (e.g., BRE) processes, we do so in other ways (we elaborate in the next section). Furthermore, whenever we have to account for the beliefs of others in some way, we do not do so by using a simple static common prior assumption, but use a heuristic approach instead (again, we elaborate in the next section).

The Bayesian exploration formulation presented here is optimal with respect to the beliefs of the agents—assuming, of course, that the agents have the means to reasonably predict the outcomes of negotiations. The "optimality" of this approach is irrespective of the nature of the formation process that precedes the execution of a coalitional action. Whether the formation process will lead to a stable coalition configuration is not the major concern for the agents—even though our model allows for the use of (dynamic) formation processes with convergence guarantees, such as the BRE process presented earlier. The main concern of the agents is to maximize their long-term value without behaving naively (exploring perhaps aimlessly or with insufficient reason) while learning.

Next we examine specific Bayesian RL methods developed for use in the described framework. These methods are, in essence, computational approximations dealing with solving the optimal exploration POMDP described by Equations 6.2 and 6.3.

6.2 Computational Approximations

The calculation of an exact solution to the repeated coalition formation problem, using the Bayesian RL formulation of Equations 6.2 and 6.3 is infeasible, due to the bottlenecks discussed above. However, there exist ways to work around the problems, at least to some extent: here, we describe several algorithms that do so. These Bayesian RL algorithms can be combined with any underlying negotiation process. The agents can evaluate any potential coalitional agreement described by a triplet $\langle C, \alpha, \boldsymbol{d}_C \rangle$ that may arise as a result of a negotiation.

They will then use these valuations to enter negotiations that may be governed by any set of rules.[4] Thus, we use the algorithm of Figure 6.4 to provide an approximate solution to the problem of optimal repeated coalition formation under uncertainty.

1. Each agent i with belief state B_i calculates the Q-value of any potential agreement $\langle C, \alpha, \boldsymbol{d}_C \rangle$ in which it can potentially participate, *by approximating the solution to Equations 6.2 and 6.3*, using one of the following algorithms (described below): *OSLA*, *VPI*, *VPI-over-OSLA*, *Myopic*, *MAP*.
2. The agents engage in a coalition formation process, in which each one of them is using the Q-values calculated above in order to reflect the long term value of coalitional agreements. The process results in a coalition structure CS, a payoff allocation vector $\boldsymbol{d}$, and a vector of coalitional actions $\boldsymbol{\alpha}$, one for each coalition $C \in CS$.
3. Each agent i observes its own coalition $C \in CS$, and the corresponding $\boldsymbol{d}_C$, α that are the restrictions of $\boldsymbol{d}$, $\boldsymbol{\alpha}$ to this coalition.
4. Each agent i updates beliefs about partners $j \in C$, $j \neq i$, using Eq. 6.1.
5. The RL process repeats.

Figure 6.4: Approximating the optimal solution to the problem of repeated coalition formation under uncertainty.

[4] Of course, the agents may need to take the specific set of rules into account when evaluating the various agreements, as they may have to account for the agreements' probability of occurence, as discussed above.

One-Step Lookahead Algorithm

Here we present an 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 would be in general impossible (and it would not be made any easier by the fact that an agent is not in complete control of the choices that dictate reward), while it is possible to *approximately* calculate the value of the belief states that might follow the execution of a coalitional action (and the subsequent observation of outcome and update of beliefs) under the current agreement. When employing the OSLA method, $V_i(B_i^{s,\alpha})$ in equation 6.2, the value of a successor belief state will be calculated *myopically*.

Specifically, we define the *1-step lookahead* Q-value of a $\langle C, \alpha, \boldsymbol{d}_C \rangle$ agreement for i, under belief state B_i, to be given by

$$Q_i^1(C, \alpha, \boldsymbol{d}_C, B_i) = \sum_s \Pr(s|C, \alpha, B_i)[r_i R(s) + \gamma V_i^0(B_i^{s,\alpha})] \tag{6.4}$$
$$= \sum_{\boldsymbol{t}_C} B_i(\boldsymbol{t}_C) \sum_s \Pr(s|\alpha, \boldsymbol{t}_C)[r_i R(s) + \gamma V_i^0(B_i')]$$

(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 follows:

$$V_i^0(B_i') = \sum_{C', \beta \in A(C'), \boldsymbol{d}_{C'}|i \in C'} \Pr(C', \beta, \boldsymbol{d}_{C'}|B_i') Q_i^0(C', \beta, \boldsymbol{d}_{C'}, B_i') \tag{6.5}$$

$$Q_i^0(C', \beta, \boldsymbol{d}_{C'}, B_i') = r_i' \sum_{\boldsymbol{t}_{C'} \in T_{C'}} B_i'(\boldsymbol{t}_{C'}) \sum_{s'} \Pr(s'|\beta, \boldsymbol{t}_{C'}) R(s') \tag{6.6}$$

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'.

The $\Pr(C', \beta, \boldsymbol{d}_{C'}|B_i')$ above is the probability of negotiation ending with the agent in a specific coalition in a specific state (i.e., in a specific coalition structure under a specific agreement to do β while dividing the demands as prescribed by $\boldsymbol{d}_{C'}$) of the Markov chain that describes the dynamic coalition formation process that takes place after the execution of some coalitional action α.

One could envisage deriving the $\Pr(C', \beta, \boldsymbol{d}_{C'}|B_i')$ probabilities by calculating the Markov

chain steady-state distribution, assuming a common prior regarding types—even though, as we explained earlier, the assumption of a common prior (after each RL step) is inherently flawed in a limited observability setting such as ours. However, this approach can prove to be expensive, both computationally and also in terms of memory requirements, as it would require that each agent calculates, after each RL step, a big transition matrix for each negotiation process corresponding to a Markov chain defined given each possible successor B_i'. To provide some further intuitions regarding the complexity of the necessary calculations, if the use of BR or BRE is assumed, then all the agents have to estimate, for *every* potential state of *every* such Markov chain (corresponding to each possible successor belief state of the agent), the $d_i^{max}(C, \alpha)$ maximal realistic demands of any type of any other agent i regarding *any* $\langle C \cup \{j\}, \alpha \rangle$ proposal that could be made by some j in that state. Further, the agents would have to take into account the possibility of agents experimenting (if BRE is used).

Thus, we instead choose to adopt a different strategy to approximate the $\Pr(C', \beta, \boldsymbol{d}_{C'}|B_i')$ probabilities, regarding them as being the probabilities of ending in a state containing the specific agreement $\langle C', \beta, \boldsymbol{d}_{C'} \rangle$ after one negotiation step, and not after the whole negotiation process—this is in the spirit of employing a bound (or lookahead) for the size of the bargaining game tree. (We explain the choice of 1 as a size bound below.) Once negotiations are entered, the OSLA agents need to utilize Q^1 values in order to make decisions; in order to calculate the Q^1 values, the probabilities in question are calculated at each negotiation step, but *only* by the agents that are involved in negotiations at this step.

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, should have to calculate (as prescribed in 6.5), the $\Pr(C', \beta, \boldsymbol{d}_{C'}|B_i')$ probabilities for each B_i' reached after the assumed execution of α and assumed observation of s. However, we make the assumption that the agent cares only for $\langle C', \beta, \boldsymbol{d}_{C'} \rangle$ agreements that are reachable within one negotiation step after "fixing" his beliefs to B_i': Assuming the use of the BRE negotiations algorithm, agent i calculates the probability with which a certain agreement $\langle C', \beta, \boldsymbol{d}_{C'} \rangle$ will be reached within one negotiation step, by solving the game tree corresponding to the way the BRE process will evolve in one step, and appropriately summing up the probabilities of any calculated best response strategies that give rise to $\langle C', \beta, \boldsymbol{d}_{C'} \rangle$. Essentially, the deliberations of the agents are similar to the ones used when solving a BCBG under fixed beliefs with a lookahead value of 1, using a common prior assumption, but also accounting for the BRE process specifics (i.e., the agents can propose only to coalitions already in place, and may experiment when proposing or accepting).

We chose to use a lookahead value of 1 when solving the game tree in our experiments, for

computational efficiency (considering that such a game tree has to be solved by each agent i for each B_i' successor belief state—the B_i' beliefs corresponding to the agent's prior when solving a BCBG game tree such as the one shown in Fig. 5.1). However, this tree-size lookahead could take any value of $l \geq 1$, depending on the specific setting's requirements.

An additional issue interfering with the solution of the game tree is that, as mentioned before, the common prior assumption is not a valid assumption in our setting, firstly because any common prior assumed in the very beginning of the whole process cannot be assumed to remain unchanged ("static"), and secondly because it is not possible to monitor the way the other agents' beliefs change, due to limited observability of coalitional actions. As a partial remedy for this problem, we considered making an "optimistic" assumption that the beliefs of others would coincide with one's beliefs over time (apart from their part refering to their beliefs regarding that agent, of course)—in other words, each agent i can assume that his changing B_i beliefs are shared by others at every point in time, and use these to describe the common prior whenever solving a BCBG game tree. However, this assumption is itself quite unrealistic—and, when we experimented with it, it did not lead to good results. We thus tried to empirically identify the solution that would be most beneficial for our agents; we eventually settled for a "hybrid" approach, which uses a "static" common prior in the initial[5] stages of learning, but switches to using an "optimistic" prior later on.

Once the $\Pr(C', \beta, \boldsymbol{d}_{C'}|B_i')$ probabilities are calculated, the agent is able to estimate his 1-step Q-values regarding any $\langle C, \alpha, \boldsymbol{d}_C \rangle$ that he needs to consider, and use these Q-values to negotiate with others. However, two more computational difficulties arise when one tries to sum over all possible $\boldsymbol{t}_C$ in equations 6.4 and 6.6, and over all possible formation actions (choice of coalition, action and demands) in equation 6.5 above. Nevertheless, the use of 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 $\langle C, \alpha, \boldsymbol{d}_C \rangle$ in which it can potentially participate, using Equations 6.4, 6.5 and 6.6.

    - The $\Pr(C', \beta, \boldsymbol{d}_{C'}|B_i')$ probabilities in Eq. 6.5 are derived for each potential successor belief state B_i' by each agent solving a BCBG game tree describing the anticipated negotiations, assuming a tree size of l (i.e., a negotiation horizon of size

[5] Specifically, for the first 50 RL steps in the experiments described in Section 6.4.

l), and a common prior derived by the "hybrid" approach described above. (In our experiments, we used $l = 1$.)

2. The calculated Q_i^1 values are then used by i in the coalition formation process of the subsequent coalition formation stage. (Specifically, if this process is the BR or the BRE dynamic process described in Chapter 4, the Q_i^1 values are used to calculated the $\bar{p}_i^i$ and d_i^{max} values used by i in its deliberations.)

VPI Exploration Method

In Chapter 3 of this dissertation, we described a multiagent *VPI exploration method* that was based on the single-agent RL method of the same name (initially developed in [DFR98, DFA99]). Recasting the relevant ideas to the repeated coalition formation setting, we now 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.

Let us consider what can be gained by learning the true value of *some* coalitional agreement $\sigma = \langle C, \alpha, \boldsymbol{d}_C \rangle$. If σ is adopted and corresponding action α is executed, assume that it leads to specific *exact evidence* regarding the types of the agents in C. Thus, we assume that the real 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; let it be denoted as $q_\sigma^* = q_{\langle C,\alpha,\boldsymbol{d}_C\rangle}^* = Q_i(C, \alpha, \boldsymbol{d}_C|\boldsymbol{t}_C^*)$, with

$$Q_i(C, \alpha, \boldsymbol{d}_C|\boldsymbol{t}_C^*) = r_i \sum_s \Pr(s|\alpha, \boldsymbol{t}_C^*)R(s) \tag{6.7}$$

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 to a change of the agent's policy. This can happen in two cases: (a) when the new knowledge shows that a coalitional action that was previously regarded as inferior to others is now revealed to be the best choice, and (b) when the new knowledge indicates that the action previously regarded as best, is actually worse than originally predicted.

For case (a), 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 $q_1 = Q_i(C_1, \alpha_1, \boldsymbol{d}_{C1}|B_i) = E_{B_i}[q_{\langle C_1,\alpha_1,\boldsymbol{d}_{C1}\rangle}]$. Moreover, suppose

that the new knowledge indicates that σ is a better action; that is, $q^*_\sigma > q_1$. Thus, we expect i to gain $q^*_\sigma - q_1$ by virtue of performing σ instead of σ_1.

For case (b), suppose that the value of the second best action $\sigma_2 = \langle C_2, \alpha_2, \boldsymbol{d}_{C2}\rangle$ is $q_2 = Q_i(C_2, \alpha_2, \boldsymbol{d}_{C2}|B_i) = E_{B_i}[q_{\langle C_2,\alpha_2,\boldsymbol{d}_{C2}\rangle}]$. 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—that is, if $q^*_{\sigma_1} < q_2$—then the agent should perform σ_2 instead of σ_1 and we expect it to gain $q_2 - q^*_{\sigma_1}$.

Thus, the *gain* from learning the true value q^*_σ of the σ agreement is:

$$\textit{gain}_\sigma(q^*_\sigma|\boldsymbol{t}^*_C) = \begin{cases} q_2 - q^*_\sigma, & \text{if } \sigma = \sigma_1 \text{ and } q^*_\sigma < q_2 \\ q^*_\sigma - q_1, & \text{if } \sigma \neq \sigma_1 \text{ and } q^*_\sigma > q_1 \\ 0, & \text{otherwise} \end{cases} \tag{6.8}$$

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

$$VPI(\sigma|B_i) = \sum_{\boldsymbol{t}^*_C} \textit{gain}_\sigma(q^*_\sigma|\boldsymbol{t}^*_C) B_i(\boldsymbol{t}^*_C) \tag{6.9}$$

The value of perfect information gives an upper bound on the myopic value of information for exploring coalitional action σ. The expected *cost* for this exploration is given as the difference between the (expected) value of σ and the value of the action currently considered best, i.e., $q_1 - E_{B_i}[q_\sigma]$ (with $E_{B_i}[q_\sigma] = E_{B_i}[q_{\langle C,\alpha,\boldsymbol{d}_C\rangle}]$ calculated as $E_{B_i}[q_\sigma] = r_i \sum_{\boldsymbol{t}_C \in T_C} B_i(\boldsymbol{t}_C) \sum_s \Pr(s|\alpha, \boldsymbol{t}_C) R(s)$). Consequently, an agent should choose to perform the action that maximizes

$$VPI(\sigma|B_i) - (q_1 - E_{B_i}[q_\sigma]) \tag{6.10}$$

This strategy is equivalent to choosing the proposal that maximizes:

$$QV_i(\sigma|B_i) = E_{B_i}[q_\sigma] + VPI(\sigma|B_i) \tag{6.11}$$

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, then sampling has to be employed.

In summary, the VPI algorithm proceeds as follows:

1. The "true" Q-values of any potential agreement σ, with respect to each realization of the relevant type vector, are myopically calculated via Eq. 6.7.

2. The gain from reaching agreement σ is calculated via Eq. 6.8.

3. The VPI for σ is calculated via Eq. 6.9.

4. The Q-values QV_i for (any) σ are calculated through Eq. 6.11 (and are subsequently used in the coalition formation process).

VPI exploration is a non-myopic method, since it does reason about the value of future belief states (accounting as it does for the value of perfect information of future coalitional agreements and its impact on the agents' decisions). Notice, however, that the VPI algorithm 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, instead of making tedious attempts to estimate the specific value of anticipated coalitional actions, which is what lookahead methods explicitly try to do. Thus, unlike lookahead methods, 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 (in other words, it does not try to explicitly approximate the solution to the POMDP described in Eq. 6.2 and 6.3). The VPI exploration method is thus not tightly tied to the specific formation process used. As we will see later in this chapter, 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). In brief, VPI-over-OSLA proceeds as follows:

1. The "true" q-values of any potential agreement σ are calculated, assuming one-step lookahead and calculation of the V_i^0 and Q_i^0 values of the successor belief state (following the revelation of the true $\boldsymbol{t}_C^*$) through Eq. 6.5 and 6.6.

2. The gain from reaching agreement σ is calculated via Eq. 6.8, where the values q_1 and q_2 of the best and second-best actions are calculated through Eq. 6.4, 6.5 and 6.6.

3. The VPI for σ is calculated via Eq. 6.9.

4. The Q-values QV_i for (any) σ are calculated through Eq. 6.11 (and are subsequently used in the coalition formation process).

Myopic Bayesian RL Algorithm

A myopic Bayesian RL algorithm may be defined exactly as was earlier described: the agents do not reason about future belief states, but rather just assess myopically the value of various coalitional moves, apply an inner coalition formation process (such as the BRE process presented in Chapter 4), and repeat.

An agent i using myopic Bayesian RL calculates the value of agreements $\langle C, \alpha, \boldsymbol{d}_C \rangle$ under belief state B_i, as follows:

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

(where r_i is i's relative demand given $\boldsymbol{d}_C$).

Maximum A Posteriori Type Assignment RL Algorithm

A *maximum a posteriori type assignment (MAP)* algorithm can also be defined. This algorithm effectively reduces the problem of updating Q-values about agreements given beliefs about the types of opponents, into the problem of updating Q-values about agreements given a belief that opponents' true type is the one specified by our beliefs as being the most probable.

In other words, 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 will be able to calculate the value of agreements as follows

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

(where r_i is i's relative demand given $\boldsymbol{d}_C$).

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

6.3 On Combining the RL Algorithms with the Formation Process

It is easy to define variants of our Bayesian RL algorithms, in order to accomodate different environment requirements. What is more, we can partition the space of the possible variants of RL algorithms, by examining their combination with various coalition formation processes. For example, we can consider the following four classes of reinforcement learners, combining Q-value estimation with dynamic formation processes such as those identified in Chapter 4.

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 sequential reasoning (using the OSLA or the VPI RL method, for example), in their attempt to solve the POMDP described by equations 6.2 and 6.3.

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 coalition formation process (for example, BR or BRE), myopically 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 assumed to be chosen, and once a proposal has been made and accepted or rejected, no further negotiations are assumed to take place: the coalitional action is assumed to be executed after a *single* proposal. Finally, *non-myopic/one-step proposers (NM-OSP)* are, naturally, the obvious combination of NM-FN and M-OSP agents. Notice that the fact that OSP agents assume (from an RL perspective) that the negotiation process has only one round, does not necessarily mean that the actual negotiations will last for just one round. Specifically, an agent may deliberate about the value of various agreements by supposing one-step negotiation, to simplify its reasoning. This is possible even if the actual negotiation uses multiple rounds. Nevertheless, in the experiments of this chapter, all OSP agents simulations involve actual negotiations that last for one round.

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 is non-empty 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 (as was suggested in Chapter 5).[6]

However, FN-methods will face the problem that it is not possible for the agents to fully explore all coalition formation possibilities (if dynamic processes that lead to stable configurations are used): at the end of each stage, the agents will indeed have strong information on a sub-space of the coalition structure space, specifically the subspace that contains the stable coalition structure the agents have led themselves into; but the agents may not have many opportunities to explore coalition structures not reachable under their beliefs (since if they indeed reach a stable structure given their beliefs, they won't have "an interest" in exploring more). This is in contrast to OSP approaches, which may potentially provide the agents with more flexibility to investigate the whole space of structures.

6.4 Experimental Evaluation

In order to evaluate our methods experimentally, we conduct four sets of experiments. 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. This set of experiments also underscores the differences in behaviour and perfomance between the agents employing full negotiations during bargaining, and those not. In the second set of experiments, our agents act in a *dynamic* environment, being presented with a different problem after each RL step. This setting demonstrates that our approach allows for the *transfer of knowledge* between different tasks, allowing as it does the agents to progressively update beliefs regarding the partners' capabilities. Further, it helps demonstrate the benefits of using the VPI method in particular (we elaborate below). Our third experimental setting also helps demonstrate the transfer of knowledge benefits of our approach. However, unlike the second setting, here we allow the agents to have knowledge of the (different) formation problem that they will face in the next RL step. Finally, the fourth set of experiments attempts a comparison of our methods to the KST method (introduced in Chapter 5).

In all cases the experiments focus on assessing the performance of our methods while learning by observing the emerging coalitions' behaviour in repeated coalition formation settings—

[6] We do not explore this possibility in the experiments of this chapter, in order to focus more on the RL aspects of the repeated coalition formation problem, rather than those of bargaining.

unlike our experiments in Chapter 5 where the focus was on learning by monitoring the agents' interactions during bargaining. The main metric we use in all our experiments is discounted reward accumulated by the coalitions—this is a metric reflective of the sequential rationality of the agents' decisions. The formation process used during the coalition formation stages is the BRE dynamic process (thus providing us with the opportunity to make observations relating to the agents' BC-convergence behaviour). For sampling type vectors, we used the following approach: if $|T|^{|C|} \leq 1000$, 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. Overall, our experiments show that our Bayesian RL approach (and especially our VPI method) facilitates the agents' sequential decision making under uncertainty, and contributes to good online performance.

6.4.1 Learning while Repeatedly Facing a Specific Formation Problem

In our first set of experiments, we test our approach in two settings: the first has 5 agents, 10 types per agent, 3 actions per coalition and 3 outcome states per action; the second has 10 agents, 10 types per agent, 3 actions per coalition and 3 outcome states per action. The setting in our experiments is homogeneous (i.e., all agents in an experiment employ the same algorithm). Each experiment consists of 30 runs, and each run employs 500 RL steps. A discount factor of 0.985 was used in all of our experiments requiring discounting. Whenever a full negotiation (FN) approach is used, formation negotiations last for 50 rounds (per RL step). Agents can observe the results of the action taken by the coalition to which they belong, not those of any other coalition. Thus, they can only update their beliefs regarding their partners at any stage. However, the assumption of observability of the membership of all coalitions and all other agents' demands by any agent is in place. The coalition structure in place at the beginning of each RL step is the result of the preceding formation process.

The agents form companies to bid for software development projects. There are 3 "major" types, corresponding to project roles, each having 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 quality types correspond to quality "points" (starting with 0 points for "bad" types and increasing by 1), which, accumulated, characterize the overall quality of a coalition. The agents know the major type of their opponents, but not their quality types. The companies can bid for a large, an average-sized or a small project (actions), and they expect to make large, average or small profit (outcome), given their choices and their members' types.

The outcome (and subsequent coalitional reward) of a coalitional action depends on the

quality of the coalition and the action performed. In general, bidding for large projects is unlikely to be rewarding: a coalition will in general 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. A coalition with 2 (or more) members is "punished" if it does not have 2 (or, respectively, at least 3) members of different "major" types, by receiving only a fraction of the reward it is entitled to given the quality of its members (see Tables B.2, B.3, and B.4 in Appendix B). Also, the reward shares that the members of a size 2 coalition expect to receive are equal to their rewards for forming singletons, and less than these if the 2-member coalition is made up of members of the same "major" type. Thus, it is to be expected that agents using a Myopic method will find it hard to form size 2 coalitions (starting from a configuration structure of singletons), even if in fact these coalitions can serve as the "building blocks" for more promising ones. We omit further details about the rewards (and the outcomes staes' transition function) here—we refer to Appendix B for further details.

Agent	Type	quality points
0	expert interface designer	2
1	good programmer	2
2	expert systems engineer	2
3	bad programmer	0
4	bad systems engineer	0

Table 6.2: Participants in the five-agents experiments.

Agent	Type	quality points
0	expert interface designer	2
1	good programmer	2
2	expert systems engineer	2
3	bad programmer	0
4	bad systems engineer	0
5	bad interface designer	0
6	average interface designer	1
7	average programmer	1
8	average systems engineer	1
9	bad programmer	0

Table 6.3: Participants in the ten-agents experiments.

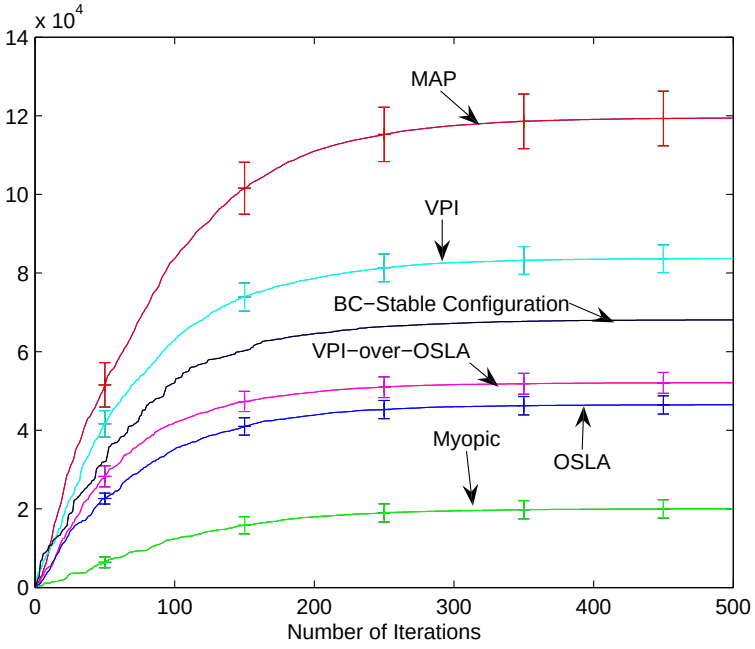
When 5 agents are present, the actual types of the agents are as in Table 6.2, while when 10 agents were present the participants' types are as in Table 6.3. The 5-agent environment is such that the (classic deterministic) core is not empty: the core contains the coalition structure $\{\langle a0, a1, a2\rangle, \langle a3\rangle, \langle a4\rangle\}$ with $\langle a0, a1, a2\rangle$ (which is a coalition of expert agents) bidding for large projects and $\langle a3\rangle$, $\langle a4\rangle$ for small. When 10 agents are present the (deterministic) core is empty.

Fully informed agents	*5 agents*	*10 agents*
Full negotiation	183713	258726
One-step proposals	139965	226490

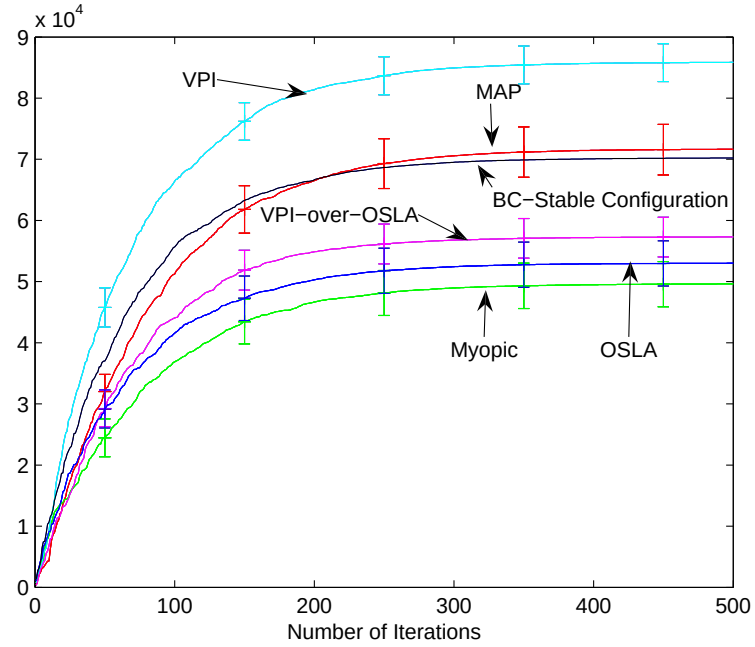
Table 6.4: Discounted average (over 30 runs) total accumulated payoffs after 500 RL steps for *fully informed* agents employing either FN or OSP formation algorithms.

Within each environment, 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 agents has a belief of 0.7 that each one of its opponents is of a quality type other than its real one. The results of the experiments are shown in Figures 6.5, 6.6, 6.7 and 6.8. The plots in these figures show how the agents in each homogeneous environment of Myopic, OSLA, VPI, MAP, or VPI-over-OSLA fared, comparing the average (over 30 runs) discounted payoff accumulated by coalitions in each one of these environments to each other. In order to have a comparison metric against some form of "optimal" behaviour of the agents in the 5-agent or 10-agent environments, 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 plots regarding fully informed agents (in order not to congest the figures), but we report their discounted average (over 30 runs) total accumulated payoff after 500 iterations, in Table 6.4. We also note that in the 5-agent case, the structure agreed upon by the fully informed agents is the $\{\langle a0, a1, a2\rangle, \langle a3\rangle, \langle a4\rangle\}$ structure, with $\langle a0, a1, a2\rangle$ bidding for a large project and $\langle a3\rangle, \langle a4\rangle$ for small ones, which, apart from being an optimal configuration (i.e., one resulting to maximum possible collective payoffs), is also a *core-stable* one. In addition, in Tables 6.5 and 6.6 we report on the average "per step" rewards accumulated in the final 50 RL steps of an average run, when the agents' beliefs and behaviour are expected to have stabilized.

Considering the experiments involving 5 agents (Figures 6.5 and 6.6) first, we can see that VPI is usually (cases of Fig. 6.6(a), 6.6(b) and 6.5(b)) doing at least slightly better than the other methods, in terms of discounted average total accumulated payoff (but in the cases of

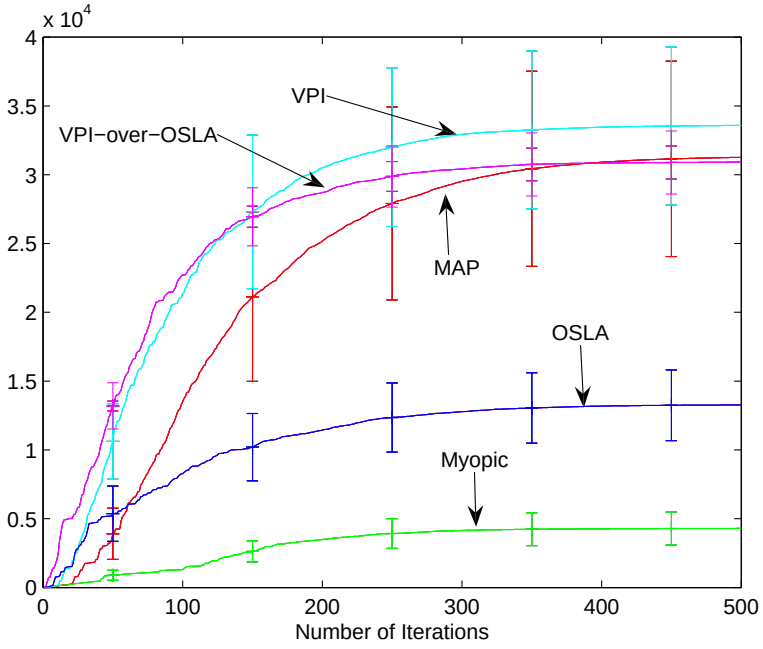


(a) Uniform priors

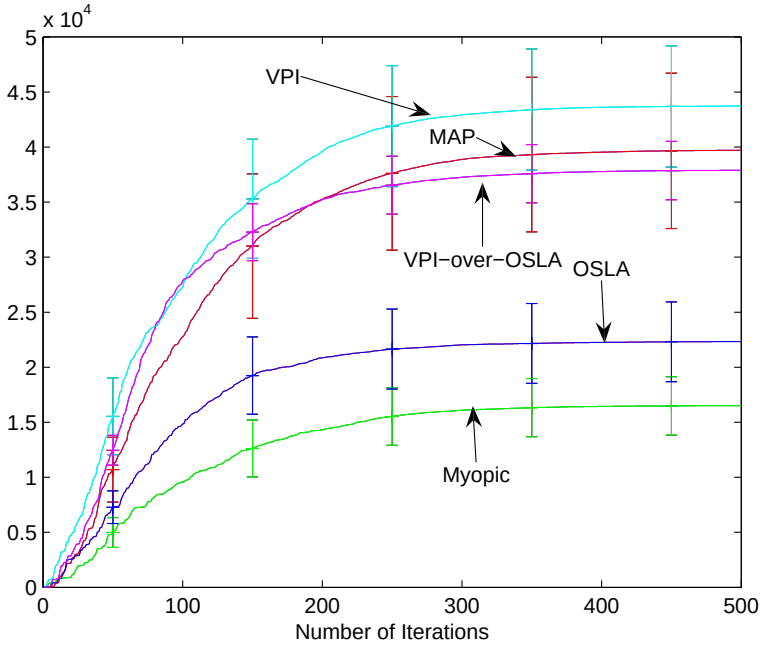


(b) Misinformed priors

Figure 6.5: Experiments with five agents, full negotiation. Discounted average total payoff accumulated by coalitions (in 30 runs). Error bars are 95% confidence intervals. The "BC-Stable configuration" is a non-optimal one, and involves no learning. The discounted averarage accumulated payoff for an optimal core-stable configuration at step 500 is as shown in Table 6.4 (i.e., $183,713$).



(a) Uniform priors



(b) Misinformed priors

Figure 6.6: Experiments with five agents, one-step proposals. Discounted average total payoff accumulated by coalitions (in 30 runs). Error bars are 95% confidence intervals. The discounted averarage accumulated payoff for an optimal core-stable configuration at step 500 is as shown in Table 6.4 (i.e., $139,965$)

Fig. 6.6 its "lead" is not significant, unlike as in Fig. 6.5(b)). In contrast, the method with the worst performance is Myopic.

Interestingly, the MAP method manages to do quite well in the five-agent experiments. Notably, it tops the other methods in the uniform priors-full negotiation case (Figure 6.5(a)). The MAP method effectively employs "crude" exploration, with agents behaving "greedily" (acting in an overly optimistic or pessimistic manner) 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 in this setting, 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). In this setting, the MAP agents are able to determine, after the initial stages, which partners have beneficial quality types—and then they stick to their choice. In fact, the MAP agents manage to achieve high rewards without ever reducing their type uncertainty regarding most of their potential partners. Defining $D(x, \tau_y) = 1 - B_x(t_y = \tau_y)$ as being the distance of x's beliefs regarding the true type τ_y of agent y from this true type, we observe a distance of approximately 0.75 or 0.66667 appearing regularly at the end of step 500, which coincides with the initial distance of an agent's prior beliefs from the true type of his partners.

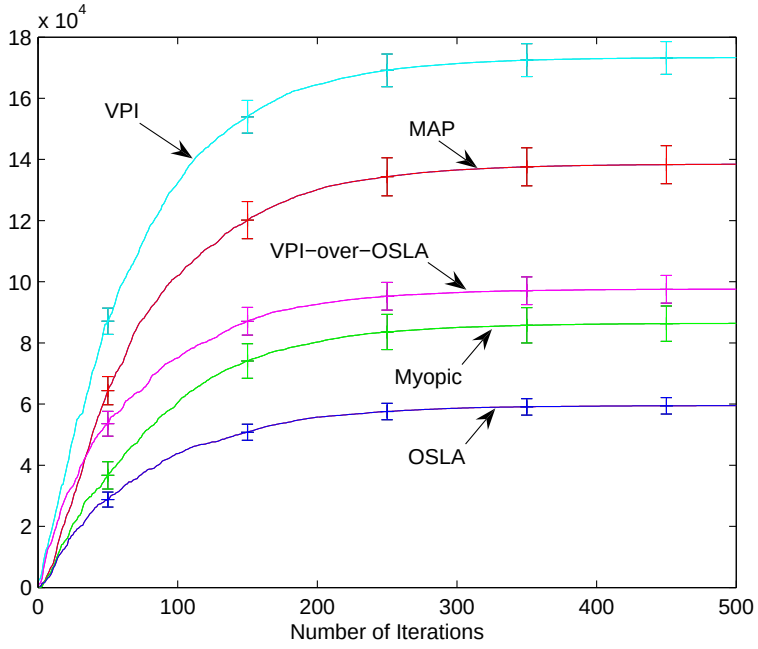
Notice that in the case of the five agents-full negotiation experiments we plot a "BC-Stable Configuration" curve corresponding to rewards accumulated by agents that are placed, at the beginning of each experiment, in a configuration that lies within the strong Bayesian core (according to their initial beliefs), and which is never left afterwards: *no renegotiations or learning is involved*. This BC-stable configuration contains, both in the uniform and in the misinformed case, the coalition structure $\{\langle a0, a3, a4\rangle, \langle a1\rangle, \langle a2\rangle\}$ with $\langle a0, a3, a4\rangle$ bidding for large projects and $\langle a1\rangle$, $\langle a2\rangle$ for small; this is in fact a quite rewarding configuration under the specific experimental setting (even though a "suboptimal" one), with $\langle a0, a3, a4\rangle$ having 10% chance to "make a large profit" (best outcome). We plot this curve to examine whether our learning methods have the potential, if used, to improve the performance of agents that have (perhaps luckily) found themselves initially in rewarding stable structures but do not care to expand their prior knowledge. In both the uniform and misinformed priors cases (Figure 6.5(a) and 6.5(b)), we can see that the agents employing the VPI method are performing substantially better than the agents lying in the rewarding BC-stable configuration. We do not plot the "BC-Stable Configuration" graph in the case of five agents-OSP experiments in order not to congest the corresponding figures. (Those graphs would have been identical to their full negotiation counterparts, since there is no learning or change of the coalition structure involved in the

"BC-Stable Configuration' case; plotting the stable configuration would have resulted in a big gap between its graph and the reward graphs generated by the various methods.)

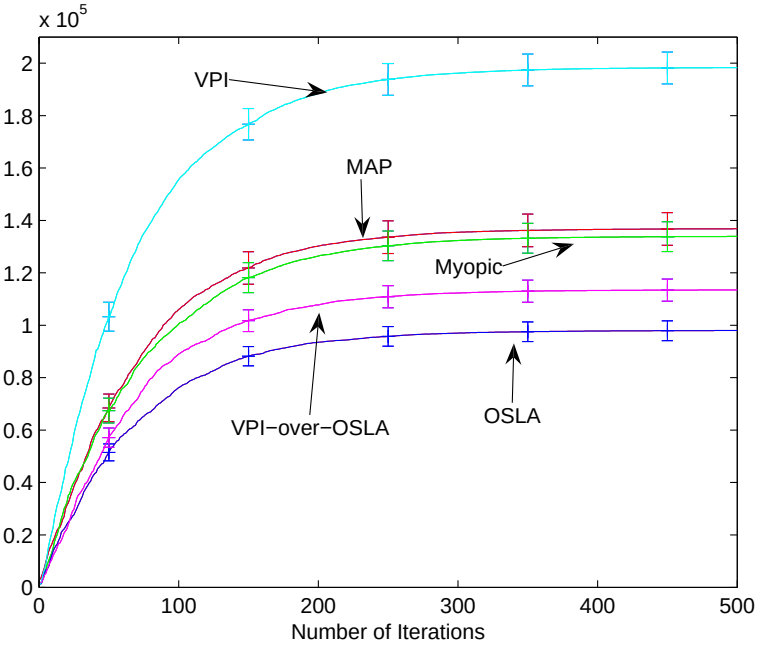
In all cases the results do indicate that agents using full negotiation are more successful than those using one-step proposals in terms of discounted accumulated reward. The magnitude of difference can be quite high, with FN agents regularly being three times as successful as their OSP counterparts (in terms of discounted accumulated reward). The fact that FN agents are taking the time to engage in lengthy dynamic formation processes between RL steps, enables them to reach more stable configurations (as these are to a greater extent the product of "collective consensus", and for that matter closer to the truly rewarding states). The more "exploratory" nature of OSP agents is, in any case, apparent when observing the error bars in the figures describing FN and OSP results. Nevertheless, when examining the results of Tables 6.5 and 6.6, we observe that OSP agents do manage to catch up with FN agents in the long run: in the final RL stages, they usually manage to gather per step reward that is comparable to that gathered by FN agents. Especially in the five-agent settings (Table 6.5), they consistently (with the only exception of MAP-Uni) achieve per step reward which is, as a percentage of that gathered by fully informed agents, higher than that gathered by their FN counterparts. Thus, the more exploratory nature of OSP agents pays out in the long run—but their performance suffers more along the way.

As the number of agents increases and the environment becomes more complicated, VPI establishes itself as the most successful of the methods, both in terms of discounted accumulated reward (Figures 6.7 and 6.8) and in terms of reward calculated when the agents beliefs have "converged" (Table 6.6). The method managed to accumulate 76.6% of the average discounted rewards accumulated by fully informed agents in the misinformed priors-full negotiation case (and 67% in the uniform priors-full negotiation case), with the rest of the 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 partners. As an example, let us discuss the reduction of uncertainty for the VPI agents in the experiments shown in Figure 6.7(b). Defining $D(x, \tau_y) = 1 - B_x(t_y = \tau_y)$ as being the distance of x's beliefs $B_x(t_y = \tau_y)$ regarding the true type τ_y of agent y from this true type, we observe that in most of the cases the $D(x, \tau_y)$ metric ranges from 0.5 to 0.96 at the end of step 500, after averaging its value over 30 experimental runs. (To be more exact, only 8 out of the 90 possible $D(x, \tau_y)$ quantities—since we are calculating the average $D(x, \tau_y)$ distance of each one of the 10 possible agents' beliefs regarding the real type of his 9 possible partners—

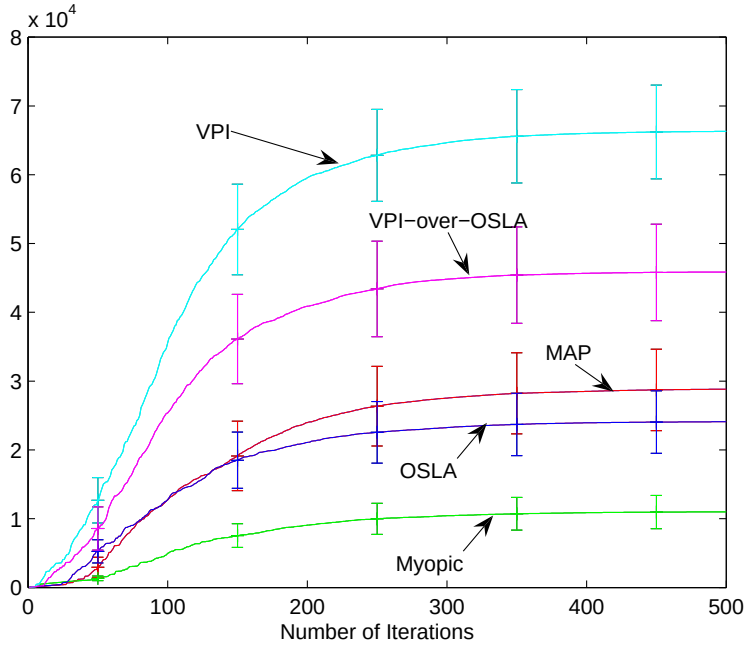


(a) Uniform priors

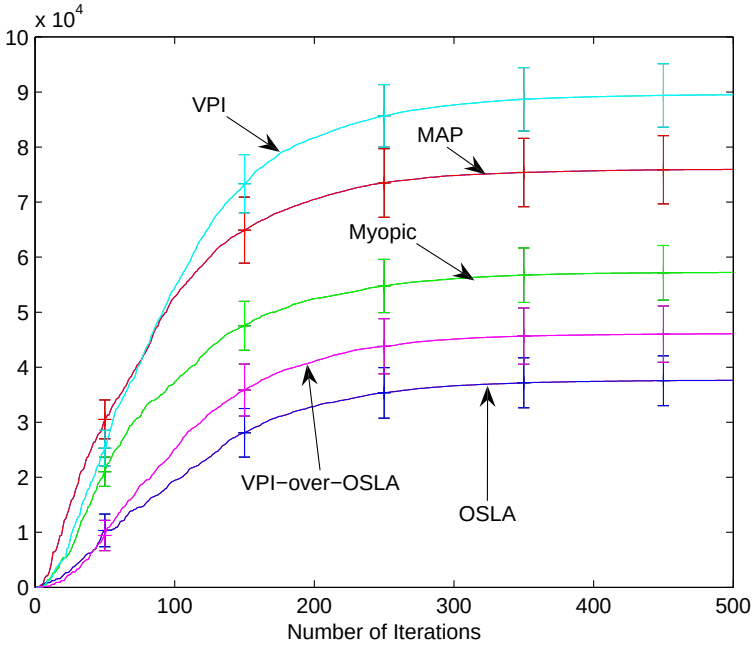


(b) Misinformed priors

Figure 6.7: Experiments with ten agents, full negotiation. Discounted average total payoff accumulated by coalitions (in 30 runs). Error bars are 95% confidence intervals.



(a) Uniform priors



(b) Misinformed priors

Figure 6.8: Experiments with ten agents, one-step proposals. Discounted average total payoff accumulated by coalitions (in 30 runs). Error bars are 95% confidence intervals.

had a value of less than 0.5.) Thus, the agents are usually not even coming close to being certain of the true type of most of their opponents—nevertheless, they manage to balance their uncertainty with their expectations in such a way that it is possible for them to make beneficial decisions. This observation reiterates the point that it is not always necessary for agents to seek to forcefully reduce uncertainty (e.g., by employing some kind of uninformed exploration) in order to achieve satisfactory performance.

The results of the experiments involving five agents, and especially those of experiments involving ten agents, indicate that 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 the common prior assumption that OSLA makes is not realistic, as explained in the previous section. It is notable, however, that the VPI-over-OSLA method consistently achieves better performance than OSLA (using VPI over OSLA seems thus to be helpful); also, in most cases, both VPI-over-OSLA and OSLA perform better than Myopic (actually, they always do better than Myopic in the 5-agent settings). Furthermore, the reward-gathering performance of OSLA and VPI-over-OSLA in the final stages of the experiments, as depicted in Tables 6.5 and 6.6, is in several cases comparable to (and in some cases superior to) the performance of methods that fare better in terms of discounted accumulated reward. Finally, Myopic usually exhibits poor performance; its approach is far too cautious, the agents not being very successful in progressively building profitable coalitions. Nevertheless, this was to be expected, since Myopic does not in any way incorporate an assessment of the value of information.

Further, it is interesting to observe when comparing the 6.5 and 6.6 tables with the discounted accumulated reward figures that, quite often, methods that are doing well in terms of accumulating reward at the end of the runs, do not necessarily fare well in terms of discounted accumulated reward. This is for example the case for MAP in the 10-agent setting with one-step proposals and uniform priors: even though it does very well in terms of rewards accumulated towards the end of the runs (Table 6.6(b)), it still ranks by far under both VPI and VPI-over-OSLA in terms of discounted accumulated reward (Figure 6.8(a)).

With regard to the stability of formed coalitions in the 5-agent setting, we observe that the VPI, OSLA, VPI-over-OSLA, and Myopic agents frequently find themselves in a BC configuration while learning (i.e., at the end of formation stages, before executing coalitional actions), even if they do not "converge" to one. "Convergence" was assumed if a least 50 consecutive RL trials before the algorithm run's termination resulted in a BC configuration. The conver-

Method	Reward
Fully informed agents	2992.82
VPI-Uni	1503.08(**50.23**%)
VPI-Mis	1387.28(**46.35**%)
VPI-over-OSLA-Uni	873.74(**29.19**%)
VPI-over-OSLA-Mis	783.44(**26.18**%)
OSLA-Uni	860.72(**28.76**%)
OSLA-Mis	807.18(**26.97**%)
MAP-Uni	2745.6(**91.73**%)
MAP-Mis	1218.24(**40.7**%)
Myopic-Uni	824.96(**27.56**%)
Myopic-Mis	1046.64(**34.97**%)

(a) Full negotiations

Method	Reward
Fully informed agents	2392.54
VPI-Uni	1611.4(**67.35**%)
VPI-Mis	1562(**65.29**%)
VPI-over-OSLA-Uni	1063.14(**44.44**%)
VPI-over-OSLA-Mis	973.92(**40.71**%)
OSLA-Uni	1562.86(**65.32**%)
OSLA-Mis	1253.8(**52.4**%)
MAP-Uni	1588.6(**66.4**%)
MAP-Mis	1459.4(**61**%)
Myopic-Uni	674.44(**28.2**%)
Myopic-Mis	723.76(**30.25**%)

(b) One-step proposals

Table 6.5: Experiments with 5 agents. Average "per step" reward accumulated within the final 50 RL steps of a run; "Uni": uniform, "Mis": misinformed prior.

Method	Reward
Fully informed agents	3884.77
VPI-Uni	2987.6(**76.9**%)
VPI-Mis	2893.6(**74.48**%)
VPI-over-OSLA-Uni	1622.5(**41.76**%)
VPI-over-OSLA-Mis	1768.86(**45.53**%)
OSLA-Uni	1564.6(**40.27**%)
OSLA-Mis	1669.4(**42.97**%)
MAP-Uni	2736(**70.42**%)
MAP-Mis	2144.34(**55.2**%)
Myopic-Uni	2419.4(**62.28**%)
Myopic-Mis	2235.2(**57.54**%)

(a) Full negotiations

Method	Reward
Fully informed agents	3881.7
VPI-Uni	2764.2(**71.21**%)
VPI-Mis	2736.4(**70.49**%)
VPI-over-OSLA-Uni	1642.88(**42.32**%)
VPI-over-OSLA-Mis	1710.96(**44.08**%)
OSLA-Uni	1542.4(**39.74**%)
OSLA-Mis	1541.8(**39.72**%)
MAP-Uni	2657.58(**68.46**%)
MAP-Mis	1660.7(**42.78**%)
Myopic-Uni	1078.68(**27.8**%)
Myopic-Mis	1462.8(**37.68**%)

(b) One-step proposals

Table 6.6: Experiments with 10 agents. Average "per step" reward accumulated within the final 50 RL steps of a run; "Uni": uniform, "Mis": misinformed prior.

	FN Unif.	FN Misinf.	OSP Unif.	OSP Misinf.
MAP	27/30	0/30	14/30	0/30
Myopic	0/30	0/30	1/30	2/30
VPI	0/30	0/30	1/30	3/30
OSLA	0/30	0/30	2/30	2/30
VPI-over-OSLA	0/30	0/30	0/30	0/30

Table 6.7: The convergence to BC results (converged/30 runs) for the algorithms (for 5 agents). "Convergence" is assumed if at least 50 consecutive RL trials before a run's termination result in a BC configuration.

gence results are shown in Table 6.7.[7] The MAP agents managed to converge to the rewarding stable configurations quite often (and this contributed to their good performance in the uniform priors-full negotiations case.) When 10 agents were present, none of the algorithms ever converged to a BC allocation, but this was to be expected since a "core" allocation *did not* actually exist.

To conclude this subsection, perhaps the most significant observation regarding these results is the consistently good performance exhibited by the VPI method. The method seems to be robust, ranking first in all but one experiment—in terms of both short-term (during the initial stages of learning) and long-term reward-gathering performance. Furthermore, VPI is a quite scalable method, whose worst case running time (for an *entire* run) is in the order of 700 sec (when 10 agents are present and full negotiation is used). By comparison, OSLA-FN can exhibit running time of the order of 6 hours/run, if 5 agents are used; or $>$25 hours /run, if 10 agents are present. (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.)

6.4.2 Learning while Facing Dynamic Tasks

The experiments we presented so far involved the agents facing the same coalition formation problem—involving the same transition to outcome states model—at each RL step. Maybe a better way to think of those scenarios is as 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 (which

[7] When we tried some runs for 10000 RL steps, the methods did seem to be able to converge to BC allocations more often.

change over time) [MCW04, KG02, SK98, SL04]. Thus, it would be interesting to see how our methods fare when they are presented with a different problem at each time step.

One of the challenges in dynamic situations, as before, is for the agents to discover the type of their opponents. The need to achieve this particular goal is more emphatic in this case, since they will have to put their beliefs to test facing different situations each time. Thus, this setting tests the abilities of our agents to achieve *transfer of knowledge* between tasks; this is one of the benefits deriving from assuming type uncertainty and using learning to tackle it: once agents learn about the abilities of partners, they can re-use this knowledge when encountering those partners again under different circumstances.

Further, in this setting we assume that the agents *do not* know in advance which task 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 is to make the environment truly dynamic. However, the POMDP assumptions do not now hold, due the non-stationarity of the environment—since the $Pr(s|\alpha, \boldsymbol{t}_C)$ domain dynamics keep changing at each time step, and the agents are not aware of the way this happens. Therefore, any method that tries to approximate the solution to the POMDP assuming stationary transition model dynamics is making an assumption that is flawed in this setting. Therefore, one would expect lookahead methods to do poorly here, since agents employing them would wrongfully anticipate to encounter a specific task (described by the same transition outcomes model currently in place) again in the near future—for example, in the next RL step, if one-step lookahead is employed. However, the performance of the myopic VPI method should not be negatively affected, since its main characteristic is employing the myopic value of perfect information regarding the types of partners, and *not* dealing 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).

To put these hypotheses to test, we used the following experimental setup: Five agents co-exist in a homogeneous environment, and form coalitions over a period of 500 RL steps. Between two RL steps, the BRE algorithm is used with 50 (full negotiation) steps. Each one of the agents is assigned with one of five different types (so that the agents are of different types). The agents share a uniform common prior regarding the types of opponents (but know their own types). At each RL step, the formed coalitions have 3 coalitional actions at their disposal, with 3 possible outcome states per action—however, the model for transitions to outcome states differ from one RL step to another, as we will describe shortly. As mentioned, the agents do not know beforehand which transition model they will encounter in the next RL step—and assume

that the current transition model will be again encountered. Finally, at the beginning of each RL step, the coalition structure is re-initialized to a coalition structure of singletons.

In fact, the experiment is designed to study the agents' behaviour while facing changing problems, but also so that the effectiveness of their type-learning ability is clearly exposed. The setup describes the problem 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 some experience-gathering phase, during which it is possible to coalesce with other villains ("El Viejo" and "Sancho Villa"), performing "petit crime" actions of lesser significance (such as "Rob Cornerstore" or "Rob Saloon") which may result to them getting to the "Some Change" or the "Some Decent Cash" outcome states—given the coalition qualities and underlying stochasticity. The setup is summarized in Figure 6.9.

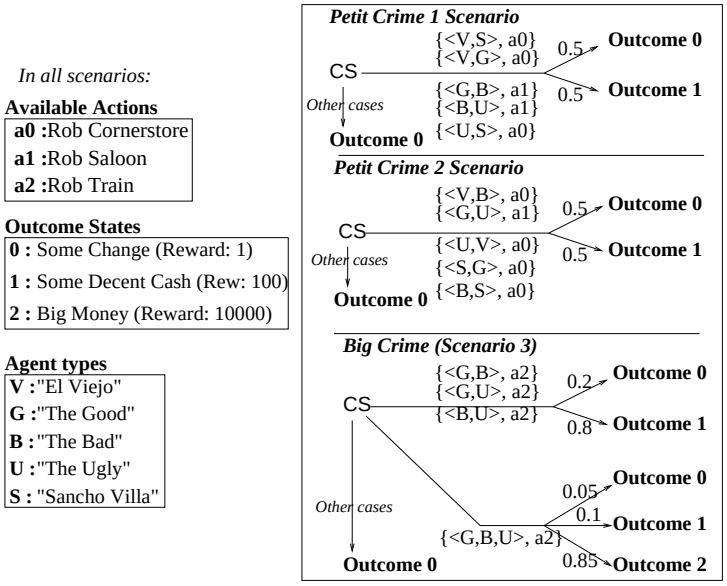


Figure 6.9: The Good, the Bad and the Ugly.

During the experience-gathering phase of the first 400 RL steps, the bandits are faced with problems 1 and 2 alternatively (with each problem employing its own, distinct outcome transition model) while they face problem 3 during the last 100 RL steps (this is the "Big Crime" phase of the experiment). 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 all of them form a coalition and decide to rob the train, they have 85% probability of making Big Money; if only two of them form a coalition, they can expect, with 80% probability, to make Some Decent Cash by taking that same action (Figure 6.9). 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. We can see in Figure 6.9, for example, that the transition model for the first "Petit Crime" scenario specifies that a coalition of the Good and the Bad have 50% chance of achieving a reward of 100 if they decide to rob the saloon, and so on.

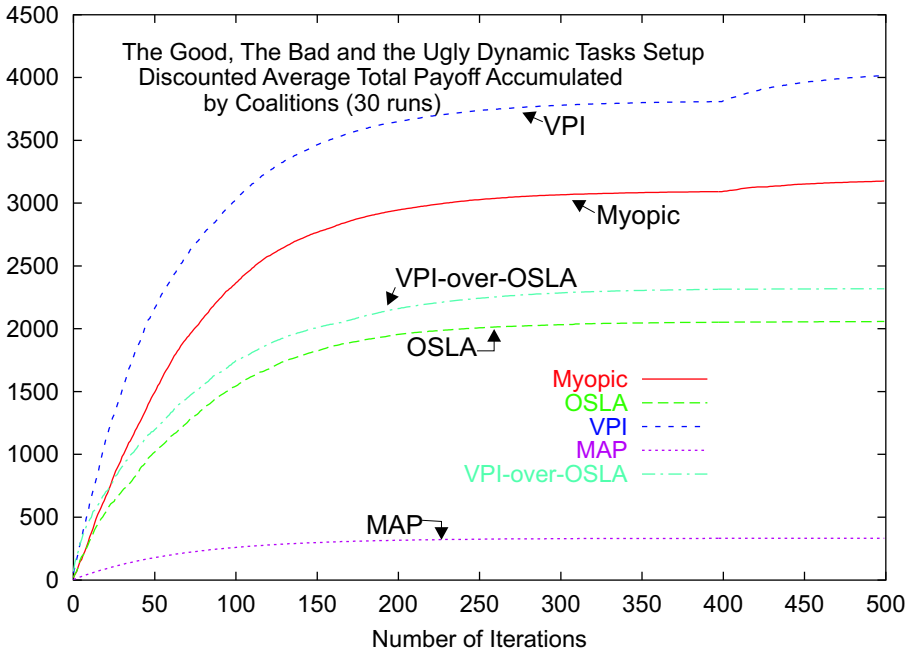


Figure 6.10: The Good, the Bad and the Ugly: Discounted accumulated reward results.

Results are presented in Figure 6.10 and Figure 6.11. It is obvious from these results that VPI dominates the other methods both in terms of discounted accumulated rewards (i.e., behaviour during the "experience-gathering" phase), and also in terms of accumulated rewards during the final stage of the experiment. In contrast, the lookahead methods' behaviour was much poorer, as expected. Nevertheless, the OSLA and VPI-over-OSLA agents do manage to collect, in the last phase of the experiment, approximately 10 and 6 times respectively more reward than the MAP agents, who appear to have been utterly confused by the setup.

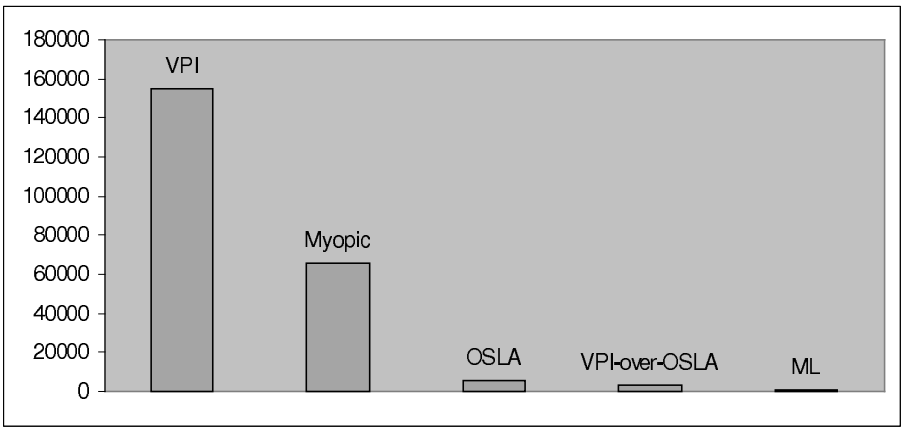


Figure 6.11: The Good, the Bad and the Ugly: Rewards gathered during the "Big Crime" phase (averaged over 30 runs).

6.4.3 More on Transfering Knowledge among Tasks

We repeated the experiment above, using again "The Good, the Bad and the Ugly" setup of Figure 6.9, with the difference that now the agents did have prior knowledge of the order with which the tasks are arriving. In other words, the agents know at every point in time the correct outcomes' transition model for the current and the following RL stages. Thus, the OSLA and VPI-over-OSLA agents are now again able to evaluate the 1-step Q-values without having false beliefs regarding the coalition formation problem to be faced after their successor belief states are reached. Again, however, the setup allows for the demonstration of the way our agents "transfer knowledge" between tasks, capable as they are of updating beliefs regarding the types of others.

The results we got are presented in Figures 6.12 and 6.13. We can see there that VPI-over-OSLA's and OSLA's performance has improved substantially in comparison to the previous setting. VPI-over-OSLA, in particular, was even able to surpass the performance of Myopic, both in terms of discounted accumulated rewards and in terms of rewards received during the final "Big Crime" phase of the experiment. Nevertheless, VPI still dominates all others in every respect.

6.4.4 Comparison to a Kernel-Based Coalition Formation Approach

To the best of our knowledge, there does not exist so far any other work that combines dynamic coalition formation with learning under "type uncertainty". However, as mentioned in Chapter 5, Kraus et al. have dealt in [KST04] with coalition formation under some restricted form

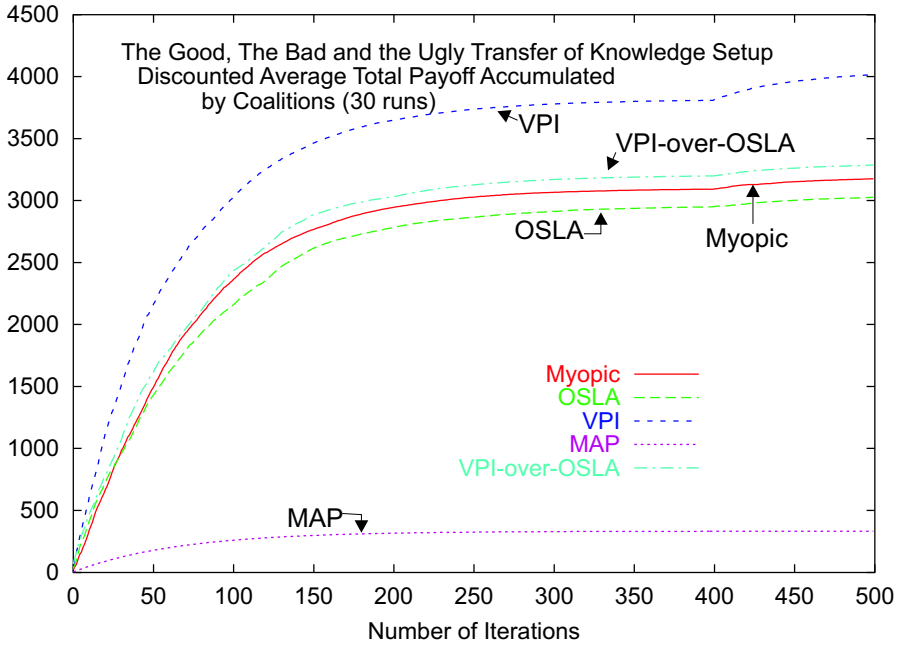


Figure 6.12: Transfer of knowledge setup: Discounted accumulated reward results.

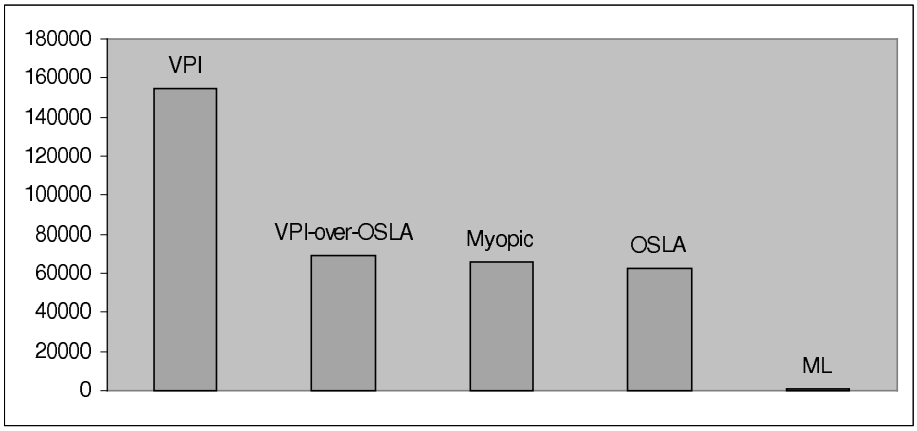


Figure 6.13: Transfer of knowledge setup: rewards gathered during the "Big Crime" phase (averaged over 30 runs).

of uncertainty (regarding coalitional values) in the "Request For Proposal" domain.

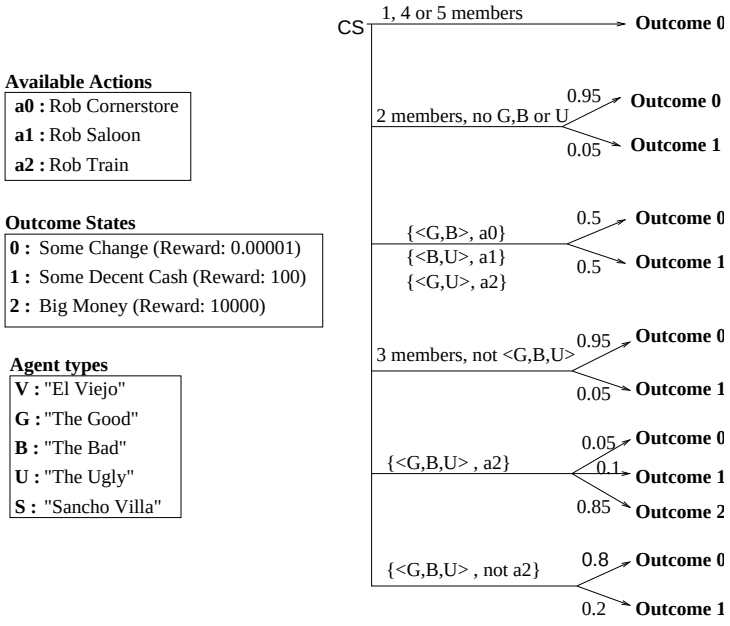


Figure 6.14: Setup for the fourth set of experiments (comparison to the KST method).

Of course, a direct comparison of our techniques with the [KST04] would not be fair to that method, since it does not use any learning and was not designed to be used in our adopted settings (for example, there are some heuristic assumptions involved which we consider inappropriate, such as computing the kernel for the coalition with the greatest coalitional value, even though this might not at all be the coalition ensuring the highest payoff to the agent).[8] Nevertheless, we chose to treat it as a benchmark (adapting it in our setting so that agents use beliefs about opponents types), in order to exhibit the benefits of learning versus non-learning approaches. Also, for interest, we combined it with our own Myopic RL algorithm (treating it as its dynamic coalition formation component), in an attempt to assess whether there exist any clear benefits between using a core-based or a kernel-based formation approach in our RL setting. As in Chapter 5, we will refer to our adapted version of [KST04] as the "KST" algorithm. Essentially, our KST is a myopic RL method, using the kernel-stable payoff allocation method and the "compromise" assumption used by [KST04].

We used an experimental setup with five agents, five possible types each, with each agent being of a different type. There exist 3 possible actions per coalition. The setup is shown in Figure 6.14. We compare KST (with and without using learning) against our VPI method

[8] [KST04] does not assume an initial demand vector or any renegotiation of agreements, so agents have to assume that the coalition with the greatest estimated value is the one to prefer.

(since this was the method that performed best in our previous experiments) and also against our Myopic algorithm (since KST is essentially a myopic method).

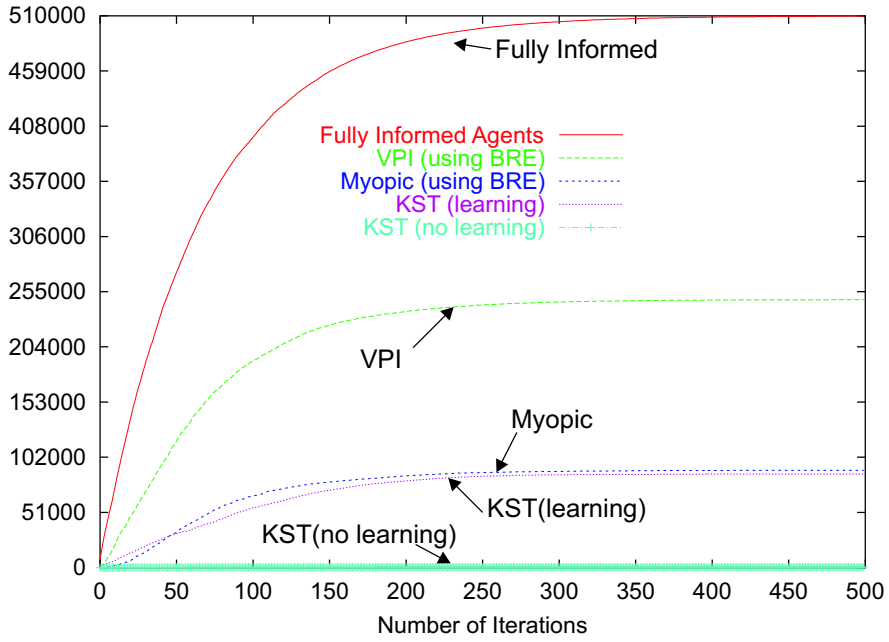


Figure 6.15: Comparison with the (adapted) KST coalition formation approach. "KST(learning)" is Myopic RL having KST as its coalition formation component. The y axis shows discounted average accumulated reward gathered in 30 runs.

Each coalition formation process consists of 50 negotiation steps, and there are 1000 RL steps in each one of the 30 experimental runs. The sampling size used was 100 (when sampling type vectors). The agents hold uniform prior beliefs regarding partners' types. At the beginning of each RL step, agents are assumed to lie in singleton coalitions. Singleton, four membered or five membered coalitions get a reward that is close to zero; two-membered coalitions have a 5% chance to "Some Decent Cash", if one of the bandits is not "Good", "Bad" or "Ugly" (both members are better off in expectation than being singletons). However, two-membered coalitions of "Good", "Bad" or "Ugly" agents have greater (50%) chances to "Some Decent Cash". Finally, forming a 3-member coalition will in general not be as rewarding to agents as forming 2-member coalitions, unless they form the $\langle Good, Bad, Ugly \rangle$ coalition and choose to "Rob the Train", in which case they have an 85% chance to get "Big Money".

The results of the experiments are shown in Figure 6.15. Clearly, VPI (using BRE) is the best method in this domain, achieving reward close to 50% of the maximum possible. When no learning is involved, the agents do poorly: "KST (no learning)" achieves accumulated discounted reward that is not greater than 350 after 500 RL steps. However, the Myopic RL

approach seems to work approximately equally well whether a core-based (BRE) or kernel-based formation approach is used (with the core-based approach doing a bit better).

6.4.5 Discussion

Preliminary as they may be, the experiments demonstrate the effectiveness of our approach, verifying that our algorithms improve the learning and decision making capabilities of the agents in coalition formation domains under type uncertainty and payoff stochasticity, even when having to deal with dynamic tasks. They also point, as we have already hinted and as we will further discuss shortly, to VPI being the most reliable of our methods—exhibiting as it does competitive behaviour in all of the settings tested and usually ranking higher than all other methods.

However, the experiments presented here do not constitute an exhaustive and systematic evaluation of our methods. Thus, in this section we outline several key factors (i.e., environment characteristics or parameters) that are expected to affect the performance of our methods, provide further interpretation of our results with respect to these factors, and discuss the scalability of our methods and their expected behaviour in more realistic scenarios.

A first factor expected to have an impact on the performance of any automated coalition formation method is the *problem size*, determined by the *number of participating agents* and the *number of possible agent types*. As these numbers grow, the number of computations required by the agents increases,[9] and so does the importance of any approximations or sampling methods used. An increased problem size is expected to put an increased computational strain particularly to any (one-step or multi-step) lookahead method used, since lookahead methods explicitly attempt the evaluation of future belief states, the number of which grows exponentially with the problem size. (We will return to this issue later in this section.) In contrast, VPI, Myopic and MAP are far less affected by this factor, since they do not attempt to evaluate successor belief states. Our results point to VPI as being the most robust of the methods against problems incurred by increased problem size: this can be deduced by the results regarding the 10 agents/10 types environments, presented in Figures 6.7 and 6.8, and Table 6.6. We can see there that VPI tops all other methods in terms of both discounted and undiscounted accumulated rewards.

However, we believe that the dominance of VPI in the 10-agent settings was basically due

[9] In addition, an increased number of agents may lead to more potential costs and bottlenecks, derived by the increased communication needs of the participating agents.

to another factor (which, as we will explain in a while, can also be related to the problem size) affecting the agents' reasoning: the *discriminative power of the transition—and, thus, the reward—model*. Simply put, a discriminative (or "information-rich") reward model allows an agent to easily differentiate between different types participating in a coalition by observing the results of coalitional actions. This factor is set to have a significant impact on the performance of methods not using sequential reasoning, such as MAP and Myopic. This is because these methods make assumptions that are myopic—and, in the case of MAP, the method's success relies on the agent making the correct "guess" about the actual type of the potential partners (given his beliefs, of course). For example, MAP agents did particularly well in the 5-agent settings (Figures 6.5 and 6.6), when the reward model was quite informative regarding the value of coalitions and the types of their members (i.e., it was not very difficult to infer which specific types were present in a coalition given the reward received following an action). When this was not anymore the case so much (Figures 6.7 and 6.8) or at all (Figures 6.10 and 6.12), MAP agents did not do as well, or failed completely. Naturally, it is not very likely that the transition and reward model of realistic environments will be highly informative; we will discuss this issue in a greater length shortly.

A further indication of the importance of the transition and reward model used is the fact that the agents can do well over time if discriminative (i.e. information-rich) transition models and reward signals are used in conjunction even with *misinformed* priors. Examples of this are shown in several occassions in our figures depicting discounted reward and in the tables describing average per step reward, with several "misinformed" agents doing better than the ones using uniform priors. We attribute this to the fact that the transition model and reward signal used in the experiments was quite informative regarding the quality of coalitions, enabling agents in many cases to discover the unrewarding parts of the coalition structure space earlier in the "misinformed" case: the information one gets by encountering a highly unexpected outcome (given the discriminative rewards model) is stronger than the one acquired when operating under uniform beliefs (in which case more exploration is required in order to resolve one's doubts).[10] Thus, an information-rich reward model can help an agent overcome the difficulties posed by a misinformed prior.

Another factor that is expected to have an impact on the performance of our methods is

[10]To make this point clearer, if agent A (mistakenly) believes that agent B is a great carpenter, he will try to interact with him early on. If the interaction is unrewarding, then agent A's initial beliefs will be promptly invalidated (given the informative reward model), and agent A will (early on) decide that he should refrain from interacting with B in the future.

the underlying negotiation process used, and more specifically *whether full negotiations or one-step proposals are assumed*. As we saw in our experiments (noticing the magnitude of difference in the amount of reward gathered by FN and OSP agents in Figures 6.5, 6.6, 6.7 and 6.8), FN tends to have a positive impact on the performance since it allows the agents the time to reach more stable configurations that are to a greater extent the product of "collective consensus", and for this reason closer to the truly rewarding states. (As we noted while presenting our experiments, the more exploratory nature of OSP agents is made apparent when observing the wider error bars appearing in figures showing OSP agent results.)

The *validity of the common prior assumption*, as well as the *accuracy of the information carried by the agents' priors* are additional factors that affect performance. Clearly, informative priors can help alleviate the drawbacks of using a myopic approach, rather than one that takes sequential behaviour into account.

One other factor that could play a role in determining the success of some of our methods is *the amount of computational power at hand*. As already discussed, any lookahead method, such as OSLA and VPI-over-OSLA, has increased computational needs. It is not very likely that these needs can be satisfied in a realistic automated coalition formation environment requiring the timely computation of coalitional decisions. In contrast to lookahead methods, the VPI method is not a computationally intensive one, since it does not require the evaluation of successor belief states and the calculation of the relevant probabilities for reaching future agreements. The MAP and Myopic methods are not computationally intensive either, but, as we will see in some detail shortly, they are themselves not very appropriate for realistic environments. Nevertheless, if enough computational resources are at hand, then using a *multi-step* lookahead method is most probably the best choice for an agent, as such a method is expected to maximize the accuracy of the sequential value calculations (subject to appropriate sampling size, which is in turn subject to computational power available).[11]

On a related note, one way to improve the performance of our lookahead algorithms is improving the accuracy of the computation of the $\Pr(C', \beta, \boldsymbol{d}_{C'}|B'_i)$ probabilities for reaching agreements in future belief states. This could, for example, be done by calculating the steady-state distribution of the Markov chain describing the underlying negotiation process. However, as noted earlier in this chapter, several concessions should be made (in the form of assumptions) for this calculation to be feasible: common knowledge of priors should be assumed, along with

[11] Naturally, the more computational power at hand, the more lookahead steps can be used, and the more samples of successor belief states can be evaluated. Thus, experimenting with different sample sizes for successor belief states could inform us of the scaling potential of a given k-step lookahead method.

complete observability of the actions of all coalitions and their resulting stochastic outcomes.

However, we envision our approach and algorithms to be used in realistic and complex task allocation environments, such as environments requiring the formation of coalitions to provide services in the computational grid [PTJ+05], most probably under specific time constraints not allowing for intensive computation [SK98]. Clearly, lookahead approaches with intense computational needs are not well-suited for such environments; and assumptions such as the ones mentioned above cannot be realistically expected to hold in open and distributed environments.

Further, realistic systems with great numbers of agents of many possible types (such as virtual entities probably located in remote regions of the computational grid) make it unlikely that the agents possess prior information that is accurate and informative; it is also unlikely in such environments that the transition and reward model provided is highly discriminative: Firstly, it is natural that in any such environment, populated by many different agents possessing a variety of capabilities and resources, there exists a multitude of coalitions that can have the same potential (i.e., similar coalitional values) while at the same time having widely varying member type vectors. Secondly, in environments like these it is conceivable that high value (or any value at all) is derived by participation in large coalitions (since complex tasks require a variety of skills), but the bigger a coalition, the more difficult it is to associate a particular member with a particular skill (and, in the likely absense of valuable small coalitions that can be used as "stepping stones" to derive knowledge, agents will find it hard to identify the types of potential partners with some certainty). As discussed above, and as seen in our experiments, the Myopic or MAP agents' reasoning would be negatively affected in such situations.[12] In contrast, VPI is a method that is expected to be far more robust in such situations, balancing as it does value of information with current expectations, without rushing to myopic conclusions (such as assigning to an agent the type currently most probable, as MAP does) that will most probably prove to be unjustified in such environments. Therefore, it is expected that VPI will have an advantage to other methods in most realistic environments.

We will now discuss existing approaches related to our work, demonstrating further the flexibility and generality of our framework and algorithms in the process.

[12] One additional factor expected to blur the picture in realistic environments, is the fact that it may not be able to guarantee that all agents will in fact allocate all of their available resources (or, even those promised to allocate) to a problem, and this would definitely affect the coalitional reward (since the resources used should in fact be, at least partly, determining the type of the agents). Such behaviour could emerge for a variety of reasons: agents may be trying to cheat, may not be rational, or, simply, they may be experiencing technical problems or facing a sudden need to allocate resources elsewhere.

6.5 Related Work

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. This is because involving all available agents in a detailed coordination/negotiation process (as, for example, is done in [SL04] and elsewhere), can seriously limit the scalability of the system. It is preferable to first form a coalition that has enough capabilities and resources to undertake the given common problem. Incidentally, notice that the concept of "resources" can be readily captured by our model, since resources can be thought of as determining (at least partly) the capabilities of the agents (and thus determining their type and coalitional values).

In this section we review some related articles, originating from both the AI and the game theory communities. Where appropriate, we highlight their differences to our work, and we provide a brief discussion to further compare our approach to others.

Shehory and Kraus [SK98] 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 initial estimation of others' capabilities. In addition, [SK98] does not deal with payoff allocation issues. The same authors, however, do address payoff allocation issues in [SK99]. There, they present two coalition formation algorithms to be used by self-interested agents in non-superadditive environments. The algorithms deal with expected payoff allocation, and the coalition formation mechanisms used are based on the kernel stability concept. However, information about the capabilities and resources of others is again obtained via communication. This is the case also for the approach of Shehory, Sycara and Zheng presented in [SSJ97], which utilizes coalition formation algorithms in order to achieve collaboration of agents within the RETSINA framework, so that tasks of common interest are executed successfully. This work focuses on serving the needs of the team (i.e., it deals with the social welfare question), and does not deal with payoff allocation issues.

As discussed in the previous chapter, and also in the experiments section above, Kraus, Shehory and Taase [KST03, KST04] proposed a heuristic method to deal with coalition formation in the "Request for Proposal" domain, under a restricted form 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 iterative coalition formation: a formed coalition "walks away" from negotiations and it cannot be decomposed.

Campos and Willmott on the other hand, do bring iterative coalition formation into the picture in [MCW04]. They attempt to tackle "iterative coalition games" that may involve up

to 100 agents that may possess different abilities that will collectively enable coalitions to fullfill a task which does not change over time. The agents are initially assigned to coalitions randomly, and they do not concern themselves with the payoff allocation problem; instead, they use several *pre-described* strategies for choosing their coalition formation moves, based essentially on whether their current coalition is a winning coalition or not over several rounds of play. Those limitations make the approach basically static, and there is no attempt to employ learning to facilitate coalition formation.

In contrast, Abdallah and Lesser [AL04] utilize reinforcement learning in their approach to "organization-based coalition formation". They assume an underlying organization to guide the coalition formation process, and Q-learning is employed in order to optimize the decisions of coalition managers, who try to assess communication or action-processing costs. However, the agents involved in this setting are assumed to be cooperative, and there is no attempt to solve the allocation problem. Furthermore, the managers are assumed to possess full knowledge of their "children" agents' capabilities.

Another piece of work worth mentioning here is a paper by Banerjee and Sen [BS00], which does deal with uncertainty regarding members' payoffs deriving from entering a coalition, even though the authors do not concern themselves with the process of coalition formation or payoffs' allocation, but rather just with the problem of "coalition selection": an agent has some imperfect summary information on anticipated payoff from joining a coalition, and has to choose one coalition over another after a fixed number of allowed interactions with them. This "summary information" is provided by a payoff-structure encoding in the form of a multinomial distribution over possible payoffs for joining the coalition. The proposed decision making mechanism for choosing a coalition makes use of this distribution, and also employs an arbitration mechanism from voting theory in order to resolve ties. In the case of limited allowed interactions, the proposed mechanism notably outperforms the *maximization of expected utility* mechanism in terms of selecting the most beneficial coalition. If the interactions allowed are infinite, however, the former mechanism reduces to the latter.

Finally, Blankenburg et al. [BDR$^+$05] have recently implemented a coalition formation process that allows the agents to progressively update *trust* values regarding others, by communicating to each other their private estimates regarding task costs and coalition valuations. They use encryption-based techniques and developed a payment protocol that ensures that the agents have the incentive to be truthful when reporting their valuations. However, the proposed mechanism involves extensive inter-agent communication, and its effectiveness seems to rely on computing the optimal coalition structures and kernel stable solutions—which involves

exponential complexities. In any case, our approach can easily incorporate the mechanism proposed in this work as the internal coalition formation "stage" of the larger RL process.

In distinction to some of the methods mentioned here (e.g., [KST03, KST04, MCW04]), the agents in our framework—given our proposed Bayesian model of Chapter 4—have the ability not only to dynamically choose the tasks they wish to deal with, but also to choose the proper way (action) to deal with them. The incorporation of task execution in our model can be readily achieved by simply *viewing the tasks as entailing the use of specific action sets*: Tasks can be thought of as entailing (triggering the existence of) certain action sets (which define actions that need to be executed in order for the tasks to be accomplished). Therefore, we can implicitly abstract away tasks, considering them as being equivalent to action sets. In other words, the choice of an action can be thought of as being made in order to serve a task, and the outcome state implicitly corresponds to the resulting quality of the attempt to serve a task.

Finally, we note that 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. Instead, they can implicitly rank the possible choices and solutions (coalition partners' choices and coalition action choices) for dealing with the specific problem immediately when it appears (given their past experiences and their beliefs about other agents capabilities).

6.6 Conclusions

In this chapter we proposed a Bayesian MARL framework for (repeated) coalition formation under uncertainty. The framework enables the agents to improve their decision-making abilities through experience gained by repeated interaction with others, and the observation of the effects of coalitional actions. The agents in our model maintain and update beliefs about the types of others, and become increasingly able to make sequentially rational decisions that reflect their interests—regarding *both* potential coalition formation activities on their part, and potential choice of actions on behalf of their formed coalitions. We made use of a POMDP formulation, which enables the agents to assess the long-term value of coalition formation decisions, including the value of potential collective actions. Our formulation enables the agents to deal with uncertainty regarding both the types of others and the outcomes of coalitional actions, and to choose actions and coalitions not only for their immediate value, but also for their

value of information.

Our RL framework is a flexible one—being able to accomodate any underlying negotiation process, being able to incorporate prior beliefs, and being independent of the requirement of convergence to a specific stability concept. It is a generic framework, that allows the agents to dynamically form coalitions, serve tasks and transfer knowledge among them. Critically, our framework enables the agents to weigh their need to explore the abilities of their potential partners with their need to exploit knowledge acquired so far.

Our experiments verify the effectiveness of our approach, and show that our Bayesian coalitional VPI method, in particular, is the most successful of our methods. It consistently ranks high in all experimental settings, being successful in facilitating the agents' sequential decision making, and improving their ability to transfer knowledge among different tasks. In addition, VPI is expected to be a method well-suited for environments more realistic than the ones examined in our experiments.

Chapter 7

Conclusions

Sequential decision making under uncertainty is always a challenge for rational autonomous agents populating a multiagent environment. Any such agent inevitably faces the task to find the right balance between exploring in order to learn and acquire useful information, and exploiting current information regarding the environment and other agents present. Moreover, when it comes to forming teams or coalitions to tackle an underlying problem, agents may be tempted to abandon formed coalitions in search for more rewarding ones; this raises the interesting question of finding coalition structures that are stable, and ways to converge to such structures (if so desired). Last but not least, having the ability to bargain effectively (i.e., taking profitable sequential bargaining decisions) under uncertainty is an issue of utmost importance to any rational agent participating in negotiation scenarios of any sort.

In the work presented in this dissertation, we have combined (multiagent) reinforcement learning and game theoretic ideas to tackle the issues above. We adopted a principled, Bayesian framework in order to deal with the agents' uncertainty regarding the environment and the capabilities or the strategies of others. Our work resulted in several contributions, both theoretical and algorithmic/practical, described in detail in the chapters of this dissertation. We note in particular that, in many cases, our work was the first to address the problem of uncertainty in coalition formation, and to the best of our knowledge there exists no other approach to date enabling agents to take into account beliefs about the types (or capabilities) of others in coalition formation scenarios.

In Section 7.1 we provide a critical summary of our dissertation. We then proceed in Section 7.2 to identify problems of interest that could form the basis for future work, in relation to the work described in this thesis.

7.1 Summary

Throughout the main chapters (Chapters 3, 4, 5 and 6) of this thesis, we (a) formally described our approach and defined relevant concepts and algorithms, (b) presented our theoretical results and proved any relevant propositions (in chapters where this was required), and (c) provided experimental evaluation of our approach and the performance of our algorithms as appropriate.

In summary, in Chapter 3 we dealt with the generalized exploration-exploitation problem in MARL in the context of stochastic games. We described a generic Bayesian approach to MARL, allowing agents to explicitly reason about their uncertainty regarding both the underlying domain and the strategies of their counterparts. We provided a formulation that provides for an optimal solution to the multiagent exploration-exploitation problem; however, the computational intractability of the solution forces one to make several modeling assumptions to help approximate this solution. Thus, we developed two heuristic algorithms for Bayesian exploration in MARL; though these incorporate several assumptions (such as an assumption of independence of the model parameters) they are well-founded on the Bayesian optimal solution of the problem. Our algorithms incorporate and update fictitious play beliefs to model opponents; this is a simple opponent modeling technique, but adequate for the class of repeated games we studied in that chapter. Nevertheless, more work and experimentation has to be done in order to determine the degree of the burden that our assumptions impose on the accuracy of the solutions provided by our algorithms—especially in environments more complex than the ones tested. Also, though one can never fully do away with approximations and modeling assumptions—indeed, possessing and maintaining opponent and environment models is an integral part of the Bayesian solution—there is always room for the development of better related models: models that will be both more accurate and also more computationally efficient. We outline some thoughts on how to achieve this in the next section.

In Chapter 4, we provided a Bayesian cooperative approach to coalition formation under uncertainty, dealing mainly with the question of coalitional stability under uncertainty. We presented a Bayesian coalition formation model that enables the agents to tackle *type* uncertainty (and thus resulting in gains in terms of reusability of knowledge regarding partners), and accomodates action-related uncertainty as well. We introduced the concept of the Bayesian core (BC), presenting three different variants of it, and discussed its properties, presenting relevant theoretical results and algorithms (constituting dynamic coalition formation processes) with some convergence to BC properties. We believe that the Bayesian core is a natural stability concept for coalition formation under uncertainty, taking into account as it does the beliefs of

the agents regarding their potential partners. Further, by establishing a process, the BRE, that is guaranteed to converge to the (strong) Bayesian core, we provided an automated way to form stable coalitions. Of course, as was discussed in Chapter 4, one could investigate ways to relax some of the assumptions used in BRE to ensure convergence. Also, extending the concept of the Bayesian core in various ways is in our immediate interests—we discuss this issue more in the next section.

In Chapter 5, we dealt with non-cooperative aspects of coalition formation, focusing on the study of the problem of discounted coalitional bargaining. We defined Bayesian coalitional bargaining games (BCBGs), described their PBE equilibrium solution, and presented a heuristic algorithm that (a) is empirically shown to resemble optimal sequential bargaining behaviour (but without any bounds' guarantees), and (b) can be combined with belief updating following the execution of coalitional actions, in RL fashion. Admittedly, the computational complexity of the heuristic can be hard to tackle; however, in the next section we outline ways to do so. We also defined the sequential equilibrium under fixed beliefs (SEFB), and we were the first to relate a non-cooperative coalition formation solution concept (the SEFB) with a cooperative one (the BC)—even though this was done under specific assumptions (i.e., the use of fixed beliefs, and the assumption of order independent equilibria). Nevertheless, as we discussed in Chapter 5 and as we further discuss in the next section, these assumptions are difficult to do away with, partly because it is not very plausible to devise a coalitional stability concept that takes the dynamic aspects of the formation process into account.

Finally, in Chapter 6 we presented a generic Bayesian MARL framework for optimal repeated coalition formation under uncertainty. Our approach enables coalition participants to make informed, sequentially rational decisions (regarding both the bargaining and the coalitional actions to take, and taking into account the value of information of the various actions), balancing exploration of actions with exploitation of knowledge in repeated coalition formation scenarios. Our framework can in principle accomodate any underlying negotiation process, and enables the agents to dynamically form coalitions and serve tasks, also allowing them to transfer knowledge gained in the past to different problems to be faced in the future.

The flexibility of the framework is even more apparent when considering settings where certain structural properties are not (well, or at all) defined. There are several such properties in all aspects of the RL and coalitional model, including the type set from which the agent types are drawn, the transition and reward model, the availability of resources, and so on. The Bayesian approach still enables the agents to alleviate the problems through the use of priors: for example, even if the types of agents are not known in advance, Bayesian agents could them-

selves assume the existence of some uniform prior over some arbitrary type set, associate this with an unknown transition model, and then attempt to learn both simultaneously by observing rewards. Though understandably hard, this is not infeasible. We did not consider this possibility in this thesis since in real-world coalition formation settings we can in many cases expect to have some degree of information regarding the set of types and transition/reward models. However, dealing with uncertainty regarding such structural properties is a challenging but interesting problem.[1]

We presented and evaluated several RL algorithms for use in this framework, and demonstrated how these algorithms can be combined with the coalition formation processes we developed in Chapter 4. Our experiments demonstrate the effectiveness of our approach in improving the agents' learning and decision making capabilities in coalition formation domains under type uncertainty and payoff stochasticity—and which may even require the transfer of knowledge among different, dynamic tasks. We provided a comparative evaluation of the behaviour and the discussed the properties of our algorithms—including their scalability potential—given their performance in our experimental settings. In broad lines, we reached the conclusion that our Bayesian VPI technique is a quite robust method, expected to exhibit competitive performance even when computational resourses are scarce, the stochasticity is high, or the initial information of the agents is poor or misleading.[2] Of course, it is clear that more extensive experimentation is required for a full understanding of the impact that potentially increased computational requirements, in conjunction with the inherent algorithmic properties, may have on performance. To this end, in the next section we propose ways to extend our framework and test our approach in more complicated and challenging environments.

To conclude, we believe that the work described in this thesis can find application in domains where sequential decision making and team or coalition formation under uncertainty are employed, such as multiagent coordination and planning, formation of robotic teams, e-commerce and e-marketplaces, wireless and/or sensor networks, and the computational grid. We now describe possible extensions of our work, along with open problems we have identified

[1] On a somewhat related note, and also in relation with opponent modeling in MARL, the agents could try—by making appropriate observations and assumptions—to assess the possibility of taking advantage of suboptimal bargaining behaviour of potential partners during the coalition formation stage. For instance, an agent that calculates a bargaining equilibrium (or maintains a distribution over equilibria, assuming uncertainty regarding the reward model), while at the same time having—through any means—strong beliefs regarding the type or availability of resources of other agents, may be able to observe irrational, suboptimal or static (e.g., "always ask for 10%") behaviour of others and exploit them during negotiations.

[2] This is also true for our Bayesian VPI technique developed for use in stochastic games' environments (Chapter 3): BVPI is expected to scale better than BOL when computational resources are scarce.

as naturally presenting themselves in some of the application domains mentioned above.

7.2 Future Work and Open Problems

Firstly, it would be interesting to apply our Chapter 3 Bayesian MARL algorithms to more complex environments, containing more than 2, and possibly heterogeneous agents, as it would be interesting to test our algorithms in scenarios with antagonistic rather than cooperating agents. For example, we could try our model in a zero-sum game of two agents playing soccer, as in [BV01b, Lit94]; we could also try our approach with $N > 2$ soccer-playing agents.

Further, as explained in Section 3.5.3, the use of more efficient sampling techniques is required for our methods to scale. More work on computational approximations to solving the belief state MDP would also be desirable. The use of function approximation methods [KLM96, Duf02] in order to deal with large state spaces could be explored. When dealing with very large problems, focusing the attention of the agents to smaller regions (to subsets of opponents or successor states of interest) could be important. Related techniques and ideas found in the literature of DEC-MDPs [BZI00], could be of value, as could perhaps be exploring the use of MDPs with a first-order structure [BRP01, SB06].

In relation to dealing with more realistic, larger scale problems, another interesting direction could be to extend our Bayesian RL model to include reasoning about the "cost of computation" [Hor90, RW91] as part of the inferential process. An agent using an approximate method to do inference could be willing to specify, in the same units as the reward/utility are measured, the cost of improving that approximation by using more computation. The computation thus would have a value, arising as the expected gains or losses inflicted by its use. It would be interesting to incorporate this value in our framework.

Moreover, it would be interesting to explore the possibility to combine Bayesian learning with more sophisticated opponent modeling techniques, perhaps focusing on exploiting opponents with limited capabilities. Apart from using simple, fictitious play opponent models, one could try more elaborate opponent modeling techniques, such as FSMs for the effective detection of patterns of play [CM96]—and use computationally tractable means of representing and reasoning with distributions over specific classes of FSMs representing strategy models. Specifically, the detection of patterns of play or specific events would lead to the update of distributions over the potential strategies of the opponents.[3]

[3]This is essentially a "conditional Bayesian learning" approach [FL98].

To give a simple example, consider FSMs representing opponent strategies as being made up from combinations of specific patterns of play. Assuming a specific number of such potential patterns of play, an agent could use Dirichlet priors over sequences of actions (corresponding to the patterns of play triggering opponent actions) in order to update beliefs regarding the FSM class used by the opponent. In some more detail, a set of parameters θ could be used to describe the agent's prior over FSMs, and construct a Bayesian network describing the dependence of actions on the FSM used by an opponent (and the dependence of observing action a^{t+1} at time $t+1$ on the action b^t having been executed at time t etc.).[4] Then, Dirichlet updates could be used to compute the posterior $Pr(\theta|a^t)$ given the observation of action a^t (corresponding to an FSM state) at time t. This posterior belief state could then be used by the agent to derive the probability of observing an opponent performing any action a^t at any time step, and then use this probability in his sequential value calculations.

Furthermore, on a somewhat related note, we would be interested in applying our Bayesian MARL ideas in the context of computational *trust* in e-marketplaces. There has recently been interesting work resulting in the suggestion of principled, Bayesian ways by which consumer agents update their degree of trust towards service-providing agents, the reputation of which is disseminated to consumers by existing reputation sources [RRRJ07, TJJL06, RPC06]. However, existing work has not dealt with the question of the buyers trying to make sequentially rational decisions on whom (which service provider and which reputation source) to interact with over a specified horizon of interactions. Given that buyers in such a setting can also act as reputation sources, this sequential aspect makes this problem a non-trivial multiagent one, since interacting agents—that model others, realizing that they are being modeled at the same time—should adopt non-myopic strategies that take the strategies of opponents into account. We believe that our Bayesian MARL formulation and algorithms can be useful in this setting, as it could allow agents to make decisions to balance exploitation of current knowledge with exploration of the space of potential providers and reputation sources.

There is a multitude of interesting questions that can be raised in such a setting: *What is the most proper way for one to update his degree of trust towards the service providers? How does one update his degree of trust towards the reputation sources? How should one combine one's own valuations and information with information coming from other sources, in the first place? When does the bias created by existing information, coming from various sources, begin to have an impact on the decisions of agents?* Thus, the space for research in this setting is

[4] One could also try to capture correlations between the actions of different opponents by using a similar network.

rich—but a proper formulation of the problem is needed, along with a proper choice of belief priors to use and independence assumptions to make, so as to guarantee the easy update of distributions. This can clearly be a challenging exercise.

Regarding coalition formation, there exists a plethora of research questions that are worth pursuing. First, it would be interesting to find ways to perhaps extend the Bayesian core concept. One idea could be to try to extend the BC to allow meta-reasoning regarding one's beliefs about the beliefs of others regarding itself. Further, a related topic could be extending the BC concept to allow for the dynamics of formation negotiations to be incorporated in the stability concept. Notice, however, that such extensions could lead to a core concept that would increasingly look like a PBE. If that is the case, then why not define stability as the PBE itself (or perhaps an order-independent instance of the PBE) in the first place? We do not as yet have clear answers to these questions.

One related research direction that is certainly worth following is extending the concept of the Bayesian core, or of the deterministic core even, to cover the possibility of *overlapping coalition formation*—i.e., settings in which the agents may simultaneously belong in more than one coalitions. Overlapping coalition formation is of much value in situations when agents can allocate different parts of their resources to serve different tasks as members of different coalitions simultaneously. To date, there is not much work on overlapping coalition formation, and, to the best of our knowledge, there exists no attempt to extend the traditional stability concepts in such settings. Such an extension is not a trivial exercise: unlike traditional coalition formation settings, where stability depends on the agents having the incentive to abandon coalitions, in an overlapping coalition formation setting stability has to depend on the agents having incentives to participate in different coalition *structures* (without necessarily abandoning their coalitions). In this context, the agents have to worry about the payoff they can achieve in the coalition structure as a whole, rather than the payoff they can achieve in only one coalition. Nevertheless, we believe there is much room for interesting theoretical research in this setting: one can define appropriate stability concepts, investigate their properties, and identify algorithms that can converge to stable structures of overlapping coalitions—or at least guarantee payoffs that are close to the ones achieved in these stable structures.

Even though the extension of the core in overlapping coalitional environments may initially look conceptually straightforward, it requires taking several careful technical steps, including proper redefinitions of concepts in this domain. Actually, as a first step towards this research direction, we believe that we can show that—having defined, in an appropriate non-trivial manner, an overlapping coalition formation core (OCF-Core) along with the concepts of super-

additivity and convexity[5] for overlapping coalition formation environments—if an overlapping coalition formation game is convex, then the OCF-Core is non-empty. We are also interested in defining balancedness for overlapping coalitional games, and relate it to the non-emptiness of the core.

We also believe that there is much space for work regarding discounted coalitional bargaining. The work described in Chapter 5 of this dissertation refers to bargaining environments with a discrete bargaining actions' space; it would be interesting to investigate the problem assuming a continuous bargaining actions' space. Further, we believe that it is possible (as is, of course, important) to improve the scalability of our heuristic bargaining algorithm, and compare the resulting variants to each other. Such less demanding variants could, for example, be constructed by having the agents track the belief updates of only a selected few of their opponents' types; and there is always room for experimentation with sample complexity when sampling type vectors.

As mentioned in Chapter 6, it is important to conduct an extensive, principled experimental evaluation of the Bayesian RL algorithms presented there, examining in more detail the behaviour of our algorithms when carefully varying the factors listed in Section 6.4.5 (e.g., the number of agents, the number of types, the descriptive power of the reward model etc.), and assessing the algorithms' scalability. This could be done in complex, realistic multiagent domains, with agents facing demanding task and resource allocation problems—perhaps for the provision of services by virtual organizations [PTJ+05], which would further require dealing with the problems at hand in a timely manner [SK98]. Also, as noted in Section 6.4.5, we can work on improving the approximations inherent in our lookahead algorithms when computing the probabilities of reaching future agreements. For the accuracy of these calculations to be improved, however, we would most probably have to make more concessions in terms of assumptions used, such as: a priori knowledge of the negotiations process used, observability of all coalitions' actions, common knowledge of priors and observability of all bargaining actions.

For interest, and in order to demonstrate the flexibility of our Bayesian MARL framework for coalition formation, we would also like to examine the effects of combining the Bayesian RL algorithms (which were combined in the experiments of Chapter 6 with one of our own formation algorithms, BRE) with other coalition formation algorithms found in the literature. Furthermore, we believe that it is easy to extend our framework to accomodate overlapping coalition formation, since it allows the formation process to be decoupled from the RL process.

[5] Briefly, a coalitional game is said to be convex if for any set of agents in the game it is more profitable to join a larger rather than a smaller coalition. All convex coalitional games are trivially superadditive.

It would also be interesting to extend our framework to accomodate uncertainty regarding any structural properties (e.g., the set of types), along the lines described in Section 7.1.

Moreover, as we observed in Chapter 6, our Bayesian coalition formation model can be extended by allowing the probabilistic model for transitioning to the various outcomes to depend on the current *state of the game*, such a state consisting of the configuration reached—i.e., the complete coalition structure and payoffs allocation reached, and the beliefs of the agents. This would allow for a sequential environment model (an underlying MDP) given a configuration, allowing for the study of coordination games played among the various coalitions present in the coalition structure (without the agents regrouping). Essentially, we propose that stages of coalition formation alternate with stages of coordination games played (and subsequent update of agents' beliefs after observing the strategies in the coordination games). Further, the coordination games played among coalitions could themselves depend on the coalition structure reached as the result of negotiations: if structure CS_1 is reached, then a specific coordination game G_1 follows (with the game payoffs depending on the coalitions' member type vectors); if structure CS_2 is reached, then game G_2 is played, and so on. Though this is a conceptually obvious extension to our model, it is not a trivial one, since now any agent's sequential decision making problem should take into account the possibility of different games being played given different structures reached, along with the possibility of coordination or miscoordination in any such game—given the capabilities and expected strategies of other players.

We note that a setting like this bears clear resemblances to natural scenarios requiring the continuous regrouping of teams of agents in need to coordinate in a variety of ways within ever-changing dynamic environments. This is the case in the Robocup Rescue competition [KT01], for example, where agents have to coordinate in order to combat the effects of major natural disasters; over time, the agents (autonomous rescue vehicles with different capabilities, moving and operating in a city hit by a disaster) have to regroup and tackle a multitude of novel coordination problems, in the face of new information. Agents operating in such a setting should be able to (a) decide which partners to cooperate with, (b) decide which coordination problem to tackle[6] and (c) choose a strategy to help them address a specific coordination problem.

Finally, we believe that there is room for interesting work in the intersection of coalition formation and mechanism design.[7] Coalition formation could be viewed as a *decentralized*

[6]For example, the agents have to decide which of the several victims trapped in buildings collapsed by an earthquake to try and save, or which parts of the city to protect from the spread of fires.

[7]Mechanism design is a subfield of game theory and microeconomics that studies the design of protocols (mechanisms) for non-cooperative environments, that attempt to implement the "rules of a game" in such a way that a system-wide goal (described by a *social choice function* that selects the optimal outcome given agent types)

mechanism design tool: via coalition formation, agents could be allowed to join forces (i.e., attempt to "collude") as they find appropriate given their beliefs, when participating in auctions or reverse auctions for the allocation of goods and services. One could perhaps apply this tool in the context of routing in networks or in the computational grid [MQ05, PTJ+05]. We believe that several interesting questions exist in this context: Would the behaviour of Bayesian agents in coalition formation scenarios resemble behaviour that could perhaps be enforced through centralized mechanisms (which can possess any of the usual desirable properties, such as incentive compatibility[8] [DJP03])? If not, how close to that behaviour would it be? Does it come close to resembling the play of any kind of equilibria? What is the difference between the earnings of the agents when using a decentralized coalition formation approach and when not? Similarly, what is the impact of using a decentralized approach to the system and the utility of the designer? What are the costs (or benefits) arising from collusion *under uncertainty*?[9] *In brief, what is the price of anarchy [KP99] in this case?* We believe that these are only some of the many related research questions worthy to be asked, and, hopefully, to be answered.

is satisfied. A *mechanism*, thus, defines the set of strategies available to the agents and the method used to select the final outcome—in order to satisfy the system-wide goal. An auction is an example of a mechanism.

[8]Briefly, a mechanism is incentive compatible if the equilibrium strategy profile has every agent reporting its true preferences to the mechanism.

[9]We are aware of only one piece of work that examines similar issues—but without assuming (or examining the properties of) a coalition formation model per se: Leyton-Brown *et al.*'s work on "Bidding clubs in first-price auctions"[LBST02].

Bibliography

[Aga97] M. Agastya. Adaptive Play in Multiplayer Bargaining Situations. *Review of Economic Studies*, 64:411–426, 1997.

[AH92] R.J. Aumann and S. Hart. Preface. In R.J. Aumann and S. Hart, editors, *Handbook of Game Theory with Economic Applications*, volume 1. Elsevier Science Publishers, 1992.

[AL04] S. Abdallah and V. Lesser. Organization-Based Coalition Formation. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'04)*, 2004.

[AS00] S. Arai and K. Sycara. Multiagent Reinforcement Learning for Planning and Conflict Resolution in a Dynamic Domain. In *Proceedings of the Fourth International Conference on Autonomous Agents*, pages 104–105, Barcelona, Spain, 2000.

[AS01] S. Arai and K. Sycara. Credit Assignment Method for Learning Effective Stochastic Policies in Uncertain Domains. In *Proceedings of Genetic and Evolutionary Computation Conference (GECCO-2001)*, San Francisco, CA, 2001.

[Ast65] K.J. Astrom. Optimal control of Markov decision processes with incomplete state estimation. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965.

[Axe84] R. Axelrod. *The Evolution of Cooperation*. Basic Books, NY, 1984.

[Bal97] T. Balch. Learning Roles: Behavioral Diversity in Robot Teams. In *1997 AAAI Workshop on Multiagent Learning*, 1997.

[BBS95] A.G. Barto, S.J. Bradtke, and S.P. Singh. Learning to Act using Real-Time Dynamic Programming. *Artificial Intelligence*, 72(1–2):81–138, 1995.

[BDH99] C. Boutilier, T. Dean, and S. Hanks. Decision Theoretic Planning: Structural Assumptions and Computational Leverage. *JAIR*, 11:1–94, 1999.

[BDR+05] B. Blankenburg, R.K. Dash, S.D. Ramchurn, M. Klusch, and N.R. Jennings. Trusted Kernel-Based Coalition Formation. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'05)*, 2005.

[Bel57] R.E. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[Bel61] R.E. Bellman. *Adaptive control processes: A guided tour*. Princeton University Press, 1961.

[Ber87] D.P. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, Englewood Cliffs, NJ, 1987.

[BF85] D.A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London, 1985.

[BKS03] B. Blankenburg, M. Klusch, and O. Shehory. Fuzzy Kernel-Stable Coalitions Between Rational Agents. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'03)*, 2003.

[Bon63] O.N. Bondereva. Some Applications of Linear Programming Methods to the Theory of Cooperative Games (in Russian). *Problemy Kibernetiki*, 10:119–139, 1963.

[Bou96a] C. Boutilier. Learning Conventions in Multiagent Stochastic Domains using Likelihood Estimates. In *UAI-96*, Portland, OR, 1996.

[Bou96b] C. Boutilier. Planning, Learning and Coordination in Multiagent Decision Processes. In *Theoretical Aspects of Rationality and Knowledge*, pages 195–201, 1996.

[Bou99] C. Boutilier. Sequential Optimality and Coordination in Multiagent Systems. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 478–485. Morgan Kaufmann, 1999.

[BPR96] S. Basu, R. Pollack, and M.-F. Roy. On the Combinatorial and Algebraic Complexity of Quantifier Elimination. *Journal of the ACM*, 43(6):1002–1045, 1996.

[BPR03] S. Basu, R. Pollack, and M.-F. Roy. *Algorithms in Real Algebraic Geometry*. Springer-Verlag, 2003.

[Bro51] G.W. Brown. Iterative solution of games by fictitious play. In T.C. Koopmans, editor, *Activity Analysis of Production and Allocation*, Wiley, New York, 1951.

[BRP01] C. Boutilier, R. Reiter, and B. Price. Symbolic Dynamic Programming for First-order MDPs. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01)*, 2001.

[BS00] B. Banerjee and S. Sen. Selecting Partners. In *Proceedings of the Fourth International Conference on Autonomous Agents*, pages 261–262, Barcelona, Catalonia, Spain, 2000.

[BSW97] C. Boutilier, Y. Shoham, and M.P. Wellman. AIJ Editorial: Economic Principles of Multi-Agent Systems. *Artificial Intelligence Journal*, 94(1):1–6, 1997.

[BV00] M. Bowling and M. Veloso. An Analysis of Stochastic Game Theory for Multiagent Reinforcement Learning. In *Technical Report CMU-CS-00-165*. Computer Science Department, Carnegie Mellon University, 2000.

[BV01a] M. Bowling and M. Veloso. Convergence of Gradient Dynamics with a Variable Learning Rate. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 27–34, June 2001.

[BV01b] M. Bowling and M. Veloso. Rational and Convergent Learning in Stochastic Games. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, Seattle, WA, August 2001.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[BZI00] D.S. Bernstein, S. Zilberstein, and N. Immerman. The Complexity of Decentralized Control of Markov Decision Processes. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI-2000)*, pages 32–37. Morgan Kaufman, 2000.

[CB98] C. Claus and C. Boutilier. The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, pages 746–752, 1998.

[CB03] G. Chalkiadakis and C. Boutilier. Coordination in Multiagent Reinforcement Learning: A Bayesian Approach. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'03)*, 2003.

[CB04] G. Chalkiadakis and C. Boutilier. Bayesian Reinforcement Learning for Coalition Formation Under Uncertainty. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'04)*, 2004.

[CB07] G. Chalkiadakis and C. Boutilier. Coalitional Bargaining with Agent Type Uncertainty. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*, 2007.

[CDS93] K. Chatterjee, B. Dutta, and K. Sengupta. A Noncooperative Theory of Coalitional Bargaining. *Review of Economic Studies*, 60:463–477, 1993.

[CGZM65] J.M. Cozzolino, R. Gonzales-Zubieta, and R.L. Miller. Markov decision processes with uncertain transition probabilities, 1965. Technical Report No. 11. Research in the Control of Complex Systems. Operations Research Center, MIT.

[Chv78] V. Chvatal. Rational behavior and computational complexity, 1978. Technical Report SOCS-78.9, School of Computer Science, McGill University, Montreal.

[CM96] D. Carmel and S. Markovich. Learning Models of Intelligent Agents. In H. Shrobe and T. Senator, editors, *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, pages 62–67, Menlo Park, California, 1996. AAAI Press.

[CMB07] G. Chalkiadakis, E. Markakis, and C. Boutilier. Coalition Formation under Uncertainty: Bargaining Equilibria and the Bayesian Core Stability Concept. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'07)*, 2007.

[CS03] V. Conitzer and T. Sandholm. Complexity of Determining Non-Emptiness of the Core. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, 2003.

[DDRJ06] V.D. Dang, R.K. Dash, A. Rogers, and N.R. Jennings. Overlapping coalition formation for efficient data fusion in multi-sensor networks. In *Proceedings of the 21st National Conference on AI (AAAI-06)*, pages 635–640, 2006.

[DeG70] M.H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

[DFA99] 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.

[DFR98] R. Dearden, N. Friedman, and S. Russell. Bayesian Q-Learning. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, pages 761–768, 1998.

[DJ06] V.D. Dang and N.R. Jennings. Coalition structure generation in task-based settings. In *Proceedings of the 17th European Conference on AI (ECAI-06)*, 2006.

[DJP03] R.K. Dash, N.R. Jennings, and D.C. Parkes. Computational-Mechanism Design: A Call to Arms. *IEEE Intelligent Systems & Their Applications*, 18:40–47, 2003.

[DM65] M. Davis and M. Maschler. The Kernel of a Cooperative Game. *Naval Research Logistics Quarterly*, 12:223–259, 1965.

[DP94] X. Deng and C. Papadimitriou. On the complexity of cooperative solution concepts. *Mathematics of Operation Research*, 19:257–266, 1994.

[DS98] 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.

[Duf02] M.O. Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, Department of Computer Science, University of Massachusetts, Amherst, 2002.

[Eva97] R. Evans. Coalitional Bargaining with Competition to Make Offers. *Games and Economic Behavior*, 19:211–220, 1997.

[FL98] D. Fudenberg and D. Levine. *The Theory of Learning in Games*. MIT Press, 1998.

[Fri91] J.W. Friedman. *Game Theory with Applications to Economics*. Oxford University Press, second edition, 1991.

[FV97a] J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer-Verlag, 1997.

[FV97b] D. Foster and R. Vohra. Regret in the online decision problem. *Games and Economic Behavior*, 21:40–55, 1997.

[FY01] D.P. Foster and H.Peyton Young. On the Impossibility of Predicting the Behavior of Rational Agents. *PNAS*, 98(22):12848–12853, 2001.

[GH99] J.K. Goeree and C.A. Holt. Stochastic Game Theory: For Playing Games, Not Just For Doing Theory. *Proceedings of the National Academy of Sciences*, 96(19):10564–10567, September 1999.

[GH03] A. Greenwald and K. Hall. Correlated-Q Learning. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 242–249, Washington, DC, 2003.

[Gil53] D.B. Gillies. *Some Theorems on n-Person Games*. PhD thesis, Department of Mathematics, Princeton University, Princeton, 1953.

[Gin00] H. Gintis. *Game Theory Evolving*. Princeton University Press, 2000.

[GJ03] A. Greenwald and A. Jafari. A General Class of No-Regret Algorithms and Game-Theoretic Equilibria. In *Proceedings of the 2003 Computational Learning Theory Conference*, pages 1–11, 2003.

[GLP02] C. Guestrin, M. Lagoudakis, and R. Parr. Coordinated Reinforcement Learning. In *Proceedings of the 2002 AAAI Spring Symposium Series: Collaborative Learning Agents*, Stanford, CA, March 2002.

[Gre88] J. Grefenstette. Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms. *Machine Learning*, 3:225–245, 1988.

[Gro02] I.E. Grossmann. Review of Nonlinear Mixed-Integer and Disjunctive Programming Techniques. *Optimization and Engineering*, 3:227–252, 2002.

[HMC96] S. Hart and A. Mas-Colell. Bargaining and Value. *Econometrica*, 64(2):357–380, March 1996.

[Hor90] E. Horvitz. *Computation and Action Under Bounded Resources*. PhD thesis, Stanford University, 1990.

[How60] R.A. Howard. *Dynamic programming and Markov processes*. MIT Press, 1960.

[HS88] J.C. Harsanyi and R. Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, 1988.

[HS03] P. Huang and K. Sycara. Multi-agent Learning in Extensive Games with Complete Information. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'03)*, 2003.

[HW98] J. Hu and M.P. Wellman. Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 242–250, San Francisco, 1998. Morgan Kaufman.

[HW99] J. Hu and M. Wellman. Multiagent Reinforcement Learning in Stochastic Games, 1999. Submitted for publication.

[JG00] K.-C. Jim and C.L. Giles. Talking Helps: Evolving Communicating Agents for the Predator-Prey Pursuit Problem. *Artificial Life*, 6(3):237–254, 2000.

[JJS94] T. Jaakkola, M. Jordan, and S. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6:1185–1201, 1994.

[JKP72] S.L.S. Jacoby, J.S. Kowalik, and J.T. Pizzo. *Iterative Methods for Nonlinear Optimization Problems*. Prentice-Hall, 1972.

[JS01] P. Jehiel and D. Samet. Learning to Play Games in Extensive Form by Valuation. *NAJ Economics*, 3, December 2001. http://www.najecon.org/naj/v3.htm.

[Kae93] L.P. Kaelbling. *Learning in Embedded Systems*. MIT Press, Cambridge, MA, 1993.

[KG02] M. Klusch and A. Gerber. Dynamic Coalition Formation among Rational Agents. *IEEE Intelligent Systems*, 17(3):42–47, 2002.

[KK02] S. Kapetanakis and D. Kudenko. Reinforcement Learning of Coordination in Cooperative Multi-agent Systems. In *Proceedings of AAAI-02/IAAI-02*, pages 326–331, Edmonton, Alberta, 2002.

[KL93] E. Kalai and E. Lehrer. Rational Learning Leads to Nash Equilibrium. *Econometrica*, 61(5):1019–1045, September 1993.

[KLC98] L.P. Kaelbling, M.L.. Littman, and A.R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[KLM96] L.P. Kaelbling, M.L. Littman, and A.W. Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[KMS00] M. Kearns, Y. Mansour, and S. Singh. Fast Planning in Stochastic Games. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI-2000)*, pages 309–316. Morgan Kaufman, 2000.

[KP99] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science*, pages 403–413, 1999.

[KR84] J.P. Kahan and A. Rapoport. *Theories of Coalition Formation*. Lawrence Erlbaum Associates, 1984.

[KR02] H. Konishi and D. Ray. Coalition Formation as a Dynamic Process, 2002. Boston College Working Papers in Economics 478.

[KS76] J.G. Kemeny and J.L. Snell. *Finite Markov Chains*. Springer, Berlin, 1976.

[KST03] S. Kraus, O. Shehory, and G. Taase. Coalition Formation with Uncertain Heterogeneous Information. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'03)*, 2003.

[KST04] S. Kraus, O. Shehory, and G. Taase. The Advantages of Compromising in Coalition Formation with Incomplete Information. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'04)*, 2004.

[KT01] H. Kitano and S. Tadokoro. RoboCup Rescue: A Grand Challenge for Multiagent and Intelligent Systems. *AI Magazine*, 22(1):39–52, 2001.

[KWZ95] S. Kraus, J. Wilkenfeld, and G. Zlotkin. Multiagent Negotiation under Time Constraints. *Artificial Intelligence*, 75(2):297–345, 1995.

[LBST02] K. Leyton-Brown, Y. Shoham, and M. Tennenholtz. Bidding Clubs in First-Price Auctions. In *Proceedings of AAAI-02/IAAI-02*, pages 373–378, Edmonton, Alberta, 2002.

[LCRS03] C. Li, S. Chawla, U. Rajan, and K. Sycara. Mechanisms for Coalition Formation and Cost Sharing in an Electronic Marketplace. In *Proceedings of the 5th International Conference on Electronic Commerce*, pages 68–77, 2003.

[LDK95] M.L. Littman, T.L. Dean, and L.P. Kaelbling. On the complexity of solving Markov decision problems. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI–95)*, pages 394–402, Montreal, Québec, Canada, 1995.

[LGM01] C. Lusena, J. Goldsmith, and M. Mundhenk. Nonapproximability Results for Partially Observable Markov Decision Processes. *JAIR*, 14:83–103, 2001.

[Lit94] M.L. Littman. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Proceedings of the 11th International Conference on Machine Learning (ML-94)*, pages 157–163, New Brunswick, NJ, 1994. Morgan Kaufmann.

[Lit01] M.L. Littman. Friend-or-Foe Q-learning in General-Sum Games. In *Proceedings of the 18th International Conference on Machine Learning (ICML-01)*, pages 322–328, Williams College, 2001. Morgan Kaufmann.

[LR57] R.D. Luce and H. Raiffa. *Games and Decisions*. John Wiley & Sons, New York, New York, 1957.

[LR00] M. Lauer and M. Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 535–542, Stanford, CA, 2000.

[LS02] C. Li and K. Sycara. Algorithms for Combinatorial Coalition Formation and Payoff Division in an e-Marketplace. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'02)*, 2002.

[MA93] A.W. Moore and C.G. Atkeson. Prioritized Sweeping: Reinforcement Learning With Less Data and Less Time. *Machine Learning*, 13:103–130, 1993.

[Mar67] J. Martin. *Bayesian decision problems and Markov chains*. Wiley, 1967.

[Mat94a] M.J. Mataric. Learning to Behave Socially. In D. Cliff, P. Husbands, J-A. Meyer, and S. Wilson, editors, *From Animals to Animats 3, Third International Conference on Simulation of Adaptive Behavior (SAB-94)*, pages 453–462. MIT Press, 1994.

[Mat94b] M.J. Mataric. Reward Functions for Accelerated Learning. In W. Cohen and H. Hirsh, editors, *Proceedings of the Eleventh International Conference in Machine Learning*, pages 181–189, San Francisco, CA, 1994. Morgan Kaufmann Publishers.

[MB99] N. Meuleau and P. Bourgine. Exploration of multi-state environments: Local measure and back-propagation of uncertainty. *Machine Learning*, 35(2):117–154, 1999.

[MCW04] C. Merida-Campos and S. Willmott. Modelling Coalition Formation Over Time for Iterative Coalition Games. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'04)*, 2004.

[MCWG95] A. Mas-Colell, M. Whinston, and J.R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

[MGLA00] M. Mundhenk, J. Goldsmith, C. Lusena, and E. Allender. Complexity of Finite-Horizon Markov Decision Process Problems. *Journal of the ACM*, 47(4):681–720, 2000.

[Mon82] G. Monahan. A survey of partially observable Markov decision processes: Theory, models and algorithms. *Management Science*, 28(1):1–16, 1982.

[MQ05] P. Marbach and Y. Qiu. Cooperation in wireless ad hoc networks: a market-based approach. *IEEE/ACM Transactions on Networking*, 13(6):1325–1338, 2005.

[MS96] D. Monderer and L.S. Shapley. Fictitious Play Property for Games with Identical Interests. *Journal of Economic Theory*, 68(1):258–265, January 1996.

[MW95] B. Moldovanu and E. Winter. Order Independent Equilibria. *Games and Economic Behavior*, 9:21–34, 1995.

[Mye91] R.B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1991.

[Nac97] J.H. Nachbar. Prediction, Optimization and Learning in Repeated Games. *Econometrica*, 65(2):275–309, March 1997.

[Nac01] J.H. Nachbar. Bayesian Learning in Repeated Games of Incomplete Information. *Social Choice and Welfare*, 18:303–326, 2001.

[Nas51] J.F. Nash. Noncooperative Games. *Annals of Mathematics*, 54(2):286–295, 1951.

[Nas53] J.F. Nash. Two-Person Cooperative Games. *Econometrica*, 21:128–140, 1953.

[NPC+04] T.J. Norman, A. Preece, S. Chalmers, N.R. Jennings, M. Luck, V.D. Dang, T.D. Nguyen, V. Deora, J. Shao, A. Gray, and N. Fiddian. Agent-based formation of virtual organisations. *Knowledge Based Systems*, 17:103–111, 2004.

[Oka96] A. Okada. A Noncooperative Coalitional Bargaining Game With Random Proposers. *Games and Economic Behavior*, 16:97–108, 1996.

[OR94] M.J. Osborne and A. Rubinstein. *A course in game theory*. MIT Press, 1994.

[PB99] B. Price and C. Boutilier. Implicit Imitation in Multi-agent Reinforcement Learning. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML'99)*, 1999.

[PB03] B. Price and C. Boutilier. A Bayesian Approach to Imitation in Reinforcement Learning. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, 2003.

[Pou05] P. Poupart. *Exploiting Structure to Efficiently Solve Large Scale Partially Observable Markov Decision Processes*. PhD thesis, Department of Computer Science, University of Toronto, Toronto, 2005.

[PR94] M. Perry and P.J. Reny. A Noncooperative View of Coalition Formation and the Core. *Econometrica*, 62(4):795–817, July 1994.

[Pri03] B. Price. *Accelerating Reinforcement Learning with Imitation*. PhD thesis, University of British Columbia, 2003.

[PTJ+05] J. Patel, W.T.L. Teacy, N.R. Jennings, M. Luck, S. Chalmers, N. Oren, T.J. Norman, A. Preece, P.M.D. Gray, G. Shercliff, P.J. Stockreisser, J. Shao, W.A. Gray, N.J. Fiddian, and S. Thompson. Agent-based virtual organisations for the Grid. *Multiagent and Grid Systems*, 1(4):237–249, 2005.

[Put94] M.L. Puterman. *Markov Decision Processes*. Wiley, 1994.

[Rap70] A. Rapoport. *N-Person Game Theory*. University of Michigan Press, 1970.

[RN95] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 1995.

[RPC06] K. Regan, P. Poupart, and R. Cohen. Bayesian Reputation Modeling in E-Marketplaces Sensitive to Subjectivity, Deception and Change. In *Proceedings of AAAI-06*, Boston, MA, 2006.

[RRRJ07] S. Reece, S. Roberts, A. Rogers, and N.R. Jennings. Rumours and Reputation: Evaluating Multi-Dimensional Trust within a Decentralised Reputation System. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'07)*, 2007.

[Rub82] A. Rubinstein. Perfect Equilibrium in a Bargaining Model. *Econometrica*, 50(1):97–110, January 1982.

[RW91] S. Russell and E. Wefald. *Do the Right Thing: Studies in Limited Rationality*. The MIT Press, Cambridge, MA, 1991.

[SB98] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[SB99] J. Suijs and P. Borm. Stochastic cooperative games: superadditivity, convexity and certainty equivalents. *Journal of Games and Economic Behavior*, 27:331–345, 1999.

[SB06] S. Sanner and C. Boutilier. Practical linear value-approximation techniques for first-order MDPs. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI-2006)*, 2006.

[SBWT99] J. Suijs, P. Borm, A. De Wagenaere, and S. Tijs. Cooperative games with stochastic payoffs. *European Journal of Operational Research*, 113:193–205, 1999.

[SDP+96] K. Sycara, K. Decker, A. Pannu, M. Williamson, and D. Zeng. Distribured Intelligent Agents. *IEEE Expert-Intelligent Systems and Their Applications*, 11(6):36–45, 1996.

[Sha53] L.S. Shapley. A Value for n-Person Games. In H. Kuhn and A.W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton, 1953.

[Sha67] L.S. Shapley. On Balanced Sets and Cores. *Naval Research Logistics Quarterly*, 14:453–460, 1967.

[SJLS00] S. Singh, T. Jaakkola, M.L. Littman, and C. Szepesvari. Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms. *Machine Learning Journal*, 38(3):287–308, 2000.

[SK98] O. Shehory and S. Kraus. Methods for Task Allocation via Agent Coalition Formation. *Artificial Intelligence*, 101(1–2):165–200, 1998.

[SK99] O. Shehory and S. Kraus. Feasible Formation of Coalitions among Autonomous Agents in Nonsuperadditive Environments. *Computational Intelligence*, 15:218–251, 1999.

[SKM00] S. Singh, M. Kearns, and Y. Mansour. Nash Convergence of Gradient Dynamics in General-Sum Games. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI-2000)*, pages 541–548. Morgan Kaufman, 2000.

[SL73] J.K. Satia and R.E. Lave. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21:728–740, 1973.

[SL97] T. Sandholm and V.R. Lesser. Coalitions Among Computationally Bounded Agents. *Artificial Intelligence*, 94(1):99 – 137, 1997.

[SL04] L.-K. Soh and X. Li. Adaptive, Confidence-Based Multiagent Negotiation Strategy. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'04)*, 2004.

[SLA+99] 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.

[Son78] E.J. Sondik. The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Operations Research*, 26:282–304, 1978.

[SPG04] Y. Shoham, R. Powers, and T. Grenager. On the Agenda(s) of Research on Multi-Agent Learning. In *AAAI 2004 Fall Symposium on Artificial Multi-Agent Learning*, 2004.

[SS73] R.D. Smallwood and E.J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21:1071–1088, 1973.

[SSH94] S. Sen, M. Sekaran, and J. Hale. Learning to Coordinate Without Sharing Information. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 426–431, Seattle, Washington, July 1994.

[SSJ97] O. Shehory, K. Sycara, and S. Jha. Multiagent Coordination through Coalition Formation. In *Agent Theories, Architectures and Languages*, 1997.

[ST92] Y. Shoham and M. Tennenholtz. On the Synthesis of Useful Social Laws for Artificial Agent Societies. In *Proceedings of AAAI-92*, pages 276–281, San Jose, 1992.

[Ste68] R.E. Stearns. Convergent Transfer Schemes for n-Person Games. *Transactions of the American Mathemetical Society*, 134:449–459, 1968.

[Ste74] G. Stengle. A nullstellensatz and a positivstellensatz in semialgebraic geometry. *Mathematische Annalen*, 207:87–97, 1974.

[Stu02] B. Sturmfels. *Solving Systems of Polynomial Equations*. American Mathematical Society, 2002.

[Sut88] R.S. Sutton. Learning to Predict by the Method of Temporal Differences. *Machine Learning*, 3:9–44, 1988.

[SV97] R. Serrano and R. Vohra. Non-cooperative implementation of the core. *Social Choice and Welfare*, 14:513–525, 1997.

[SZ96] K. Sycara and D. Zeng. Coordination of multiple intelligent software agents. *International Journal of Intelligent and Cooperative Information Systems*, 5(2 & 3):181–211, 1996.

[Tan91] A. Tanin. On the core of network synthesis games. *Mathematical Programming*, 50:123–135, 1991.

[Tan93] M. Tan. Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 330–337, June 1993.

[Thr92] S.B. Thrun. The Role of Exploration in Learning Control. In D.A. White and D.A. Sofge, editors, *Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches*, Florence, Kentucky, 1992. Van Nostrand Reinhold.

[TJJL06] W.T.L. Teacy, J.Patel, N.R. Jennings, and M. Luck. TRAVOS: Trust and reputation in the context of inaccurate information sources. *Journal of Autonomous Agents and Multiagent Systems*, 12(2):183–198, 2006.

[vNM44] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1944.

[WB93] R. Williams and L. Baird. Tight Performance Bounds on Greedy Policies Based on Imperfect Value Functions, 1993. Northeastern University Technical Report NU-CCS-93-14.

[WD92] C.J.C.H. Watkins and P. Dayan. Q-Learning. *Machine Learning*, 3:279–292, 1992.

[Wie99] M. Wiering. *Explorations in efficient reinforcement learning*. PhD thesis, University of Amsterdam, 1999.

[WLBS05] T. Wang, D. Lizotte, M. Bowling, and D. Schuurmans. Bayesian Sparse Sampling for On-line Reward Optimization. In *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

[Woo99] M.H. Wooders. Multijurisdictional economies, the Tiebout Hypothesis, and sorting. *Proceedings of the National Academy of Sciences of the United States of America*, 96:10585–10587, 1999.

[WS02] X. Wang and T. Sandholm. Reinforcement Learning to Play An Optimal Nash Equilibrium in Team Markov Games. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, Vancouver, BC, 2002.

[WW95] A. Walker and M. Wooldridge. Understanding the Emergence of Conventions in Multi-Agent Systems. In *Proceedings of the First International Conference on Multi–Agent Systems*, pages 384–389, San Francisco, CA, 1995.

[Wya01] J. Wyatt. Exploration Control in Reinforcement Learning Using Optimistic Model Selection. In A. Danyluk and C. Brodley, editors, *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[Yan03] H. Yan. Noncooperative selection of the core. *International Journal of Game Theory*, 31(4):527–540, September 2003.

[YCS+05] M. Yokoo, V. Conitzer, T. Sandholm, N. Ohta, and A. Iwasaki. Coalitional Games in Open Anonymous Environments. In *Proceedings of the 20th National Conference on AI (AAAI-05)*, 2005.

[YS93] H. Yanco and L. Stein. An Adaptive Communication Protocol for Cooperating Mobile Robots. In *From Animals to Animats: International Conference on Simulation of Adaptive Behavior*, pages 478–485, 1993.

[YS01] J. Yamamoto and K. Sycara. A stable and efficient buyer coalition formation scheme for e-marketplaces. In *Proceedings of the Fifth International Conference on Autonomous Agents*, pages 576–583, Montreal, Canada, 2001. ACM Press.

[ZR94] G. Zlotkin and J.S. Rosenschein. Coalition, Cryptography, and Stability: Mechanisms for Coalition Formation in Task Oriented Domains. In *National Conference on Artificial Intelligence*, pages 432–437, 1994.

Appendix A

Non-convexity of the PBE-calculating program

A *convex optimization problem* is a constrained optimization problem of the form

$$\textit{minimize } M(\vec{x})$$

$$\textit{subject to: } \quad g_j(\vec{x}) \geq 0, \quad j = 1, \ldots, m$$

where $M(\vec{x})$ and $-g_j(\vec{x})$ are convex functions of $\vec{x}$, with $\vec{x} = (x_1, \ldots, x_n)^T$ representing the problem's n variables [BV04, JKP72].

Therefore, to prove that the constraint satisfaction program describing the PBE solution of our problem is not convex, it suffices to show that one of the constraints in the program is a non-convex function.

Proposition 8. *The constraint satisfaction program describing the PBE solution for a coalitional bargaining game is non-convex.*

Proof: We will prove the theorem for the 2-agents, 2 types per agent, bargaining case. Assume agents A and B, with possible types t_A^1, t_A^2 and t_B^1, t_B^2 respectively. For simplicity, assume also that there exists only one possible coalitional action per coalition.

Consider the simple, last-round, responder-related constraint (for responder, say, B, of type t_B):

$$\begin{aligned}\sigma_{t_B}^{\langle y_A, y_B\rangle|\mu_B}(y) \cdot (y_B(\mu_B(t_A^1|\langle y_A, y_B\rangle)V\{t_A^1, t_B\} + \mu_B(t_A^2|\langle y_A, y_B\rangle)V\{t_A^2, t_B\})) \\ +(1 - \sigma_{t_B}^{\langle y_A, y_B\rangle|\mu_B}(y)) \cdot V\{t_B\} \\ \geq V\{t_B\}\end{aligned}$$

where $\sigma_{t_B}^{\langle y_A, y_B\rangle|\mu_B}(y)$ is a variable denoting the probability that B says *yes* to proposal $\langle y_A, y_B\rangle$ giving him share y_B and giving A share y_A, when his beliefs regarding the type of A (given that A proposed $\langle y_A, y_B\rangle$) are $\mu_B(t_A^1|\langle y_A, y_B\rangle$ and $\mu_B(t_A^2|\langle y_A, y_B\rangle)$.

This is a constraint of the form:

$$g(\vec{u}) = x(\lambda(yv_1 + zv_2)) + (1-x)v_3 - v_3 \geq 0$$

where $\vec{u} = (x, y, z)^T$ and v_1, v_2, v_3, λ are all constants. To show that the program is not convex, it suffices to show that $f(\vec{u}) = -g(\vec{u})$ is not convex.

In other words, it suffices to show that

$$f(\vec{u}) = -(x(\lambda(yv_1 + zv_2)) + (1-x)v_3 - v_3)$$

$$= -(xy\lambda v_1 + xz\lambda v_2 - xv_3)$$

$$= -xyV_1 - xzV_2 + xV_3$$

where $V_1 = \lambda v_1$, $V_2 = \lambda v_2$ and $V_3 = v_3$, is not convex.

To show this, it suffices to show that *not all* the principal minors[1] of $f(\vec{u})$'s Hessian matrix, $H(\vec{u})$, are *non-negative* (as this will imply that $H(\vec{u})$ is not a "positive semidefinite" symmetric matrix— it is known that $f(\vec{u})$ is convex if and only if $H(\vec{u})$ is positive semidefinite).

The Hessian $H(\vec{u})$ of $f(\vec{u}) = -xyV_1 - xzV_2 + xV_3$ is calculated to be:

$$H(\vec{u}) = \begin{pmatrix} 0 & -V_1 & -V_2 \\ -V_1 & 0 & 0 \\ -V_2 & 0 & 0 \end{pmatrix}$$

Calculating the principal minors of $H(\vec{u})$, we observe that they are not all non-negative. For example, the second-order leading principal minor of the Hessian is strictly negative:

$$\begin{vmatrix} 0 & -V_1 \\ -V_1 & 0 \end{vmatrix} = -V_1^2 < 0$$

Therefore, the Hessian is not positive semidefinite, which means that $f(\vec{u})$ is not convex, and therefore the program cannot be written as a convex optimization problem for the 2-agent

[1]The k-th order principal minors of an $n \times n$ symmetric matrix A are the determinants of the $k \times k$ matrices obtained by deleting $n-k$ rows and the corresponding $n-k$ columns of A (where $k = 1, \ldots, n$).

bargaining case.

The proof for any "$N > 2$ agents–more than 2 types per agent" case is similar. □

Appendix B

Experiments' setup tables

	Agent a	**Agent b**	**Agent c**
Q({a}, α)	120	150	100
Q({a}, β)	110	140	90
Q({a}, γ)	100	130	80
Q({b}, α)	210	200	180
Q({b}, β)	200	190	170
Q({b}, γ)	190	180	160
Q({c}, α)	380	440	500
Q({c}, β)	370	430	490
Q({c}, γ)	360	420	480
Q({a, b}, α)	1000	980	880
Q({a, b}, β)	990	970	870
Q({a, b}, γ)	980	960	860
Q({a, c}, α)	600	650	700
Q({a, c}, β)	590	640	690
Q({a, c}, γ)	580	630	680
Q({b, c}, α)	220	280	290
Q({b, c}, β)	210	270	280
Q({b, c}, γ)	200	260	270
Q({a, b, c}, α)	650	560	450
Q({a, b, c}, β)	640	550	440
Q({a, b, c}, γ)	630	540	430

Table B.1: Agents' beliefs regarding Q-values (for experiment mentioned in section 4.6); α, β, γ denote actions.

a: "action"; s: "state"; q : quality points; $*$: any;
N: number of coalition members
$penalty = N * 0.1$: penalty to discourage employing "cheap" workers
N_{MT}: number of different "major" types present in coalition
SP: small profit state;
AP: average profit state;
LP: large profit state
BFS: bid for small project action;
BFA: bid for average project action;
BFL: bid for large project action

Table B.2: Symbols used in tables describing transition functions (for the first experimental setting in Chapter 6).

1-member coal.		$Pr(LP\|a=*,q)=0$
		$Pr(AP\|a=BFS,q)=q*0.02$
		$Pr(AP\|a=BFA,q)=q*0.01$
		$Pr(AP\|a=BFL,q)=0$
		$Pr(SP\|a,q)=1-Pr(AP\|a,q)$
2-member coal.		if $N_{MT}<2$ then $q=q/2$
		$Pr(LP\|a=*,q)=0$
		$Pr(AP\|a=BFS,q)=q*0.04$
		$Pr(AP\|a=BFA,q)=q*0.02$
		$Pr(AP\|a=BFL,q)=0$
		$Pr(SP\|a,q)=1-Pr(AP\|a,q)$
3-member coal.	if $N_{MT}<3$ then :	if $N_{MT}=1$ then $q=q/3$
		if $N_{MT}=2$ then $q=q/2$
		$Pr(LP\|a=*,q)=0$
		$Pr(AP\|a=BFS,q)=q*0.06$
		$Pr(AP\|a=BFA,q)=q*0.02$
		$Pr(AP\|a=BFL,q)=q*0.01$
		$Pr(SP\|a,q)=1-Pr(AP\|a,q)$
	if $N_{MT}=3$ then :	$Pr(LP\|a=BFS,q)=q*0.01$
		$Pr(LP\|a=BFA,q)=q*0.04$
		$Pr(LP\|a=BFL,q)=q*0.05$
		$Pr(SP\|a,q)=(1-Pr(LP\|a,q))/(q+1)$
		$Pr(AP\|a,q)=1-Pr(LP\|a,q)-Pr(SP\|a,q)$
4 or 5-member coal.	if $N_{MT}<3$ then :	if $N_{MT}=1$ then $q=q/3$
		if $N_{MT}=2$ then $q=q/2$
		$Pr(LP\|a=*,q)=0$
		$Pr(AP\|a=BFS,q)=q*0.03$
		$Pr(AP\|a=BFA,q)=q*0.05$
		$Pr(AP\|a=BFL,q)=q*0.03$
		$Pr(SP\|a,q)=1-Pr(AP\|a,q)$
	if $N_{MT}=3$ then :	$Pr(LP\|a=BFS,q)=q*0.01$
		$Pr(LP\|a=BFA,q)=q*0.04$
		$Pr(LP\|a=BFL,q)=q*0.05$
		$Pr(SP\|a,q)=(1-Pr(LP\|a,q))/(q+1)$
		$Pr(AP\|a,q)=1-Pr(LP\|a,q)-Pr(SP\|a,q)$

Table B.3: Outcome states' transition function for 5-agents environments (for the first experimental setting in Chapter 6). In all cases, $Pr(SP|a,q)$, $Pr(AP|a,q)$ and $Pr(LP|a,q)$ are eventually normalized in order to sum to one.

1 or 2-member coal.		As in a 5-agents environment
3-member coal.	if $N_{MT} < 3$ then :	if $N_{MT} = 1$ then $q = q/3$
		if $N_{MT} = 2$ then $q = q/2$
		$Pr(LP\|a = *, q) = 0$
		$Pr(AP\|a = BFS, q) = q * 0.06$
		$Pr(AP\|a = BFA, q) = q * 0.02$
		$Pr(AP\|a = BFL, q) = q * 0.01$
		$Pr(SP\|a, q) = 1 - Pr(AP\|a, q)$
	if $N_{MT} = 3$ then :	$Pr(LP\|a = BFS, q) = q * 0.01$
		$Pr(LP\|a = BFA, q) = q * 0.04$
		$Pr(LP\|a = BFL, q) = q * 0.05$
		$Pr(SP\|a, q) = (1 - Pr(LP\|a, q))/(q + 1) + penalty$
		$Pr(AP\|a, q) = 1 - Pr(LP\|a, q) - Pr(SP\|a, q)$
4,5,6 or 7-member coal.	if $N_{MT} < 3$ then :	if $N_{MT} = 1$ then $q = q/3$
		if $N_{MT} = 2$ then $q = q/2$
		$Pr(LP\|a = *, q) = 0$
		$Pr(AP\|a = BFS, q) = q * 0.03$
		$Pr(AP\|a = BFA, q) = q * 0.05$
		$Pr(AP\|a = BFL, q) = q * 0.03$
		$Pr(SP\|a, q) = 1 - Pr(AP\|a, q)$
	if $N_{MT} = 3$ then :	$Pr(LP\|a = BFS, q) = q * 0.01$
		$Pr(LP\|a = BFA, q) = q * 0.04$
		$Pr(LP\|a = BFL, q) = q * 0.05$
		$Pr(SP\|a, q) = (1 - Pr(LP\|a, q))/(q + 1) + penalty$
		$Pr(AP\|a, q) = 1 - Pr(LP\|a, q) - Pr(SP\|a, q)$
8,9 or 10-member coal.	if $N_{MT} < 3$ then :	if $N_{MT} = 1$ then $q = q/3$
		if $N_{MT} = 2$ then $q = q/2$
		$Pr(LP\|a = *, q) = 0$
		$Pr(AP\|a = BFS, q) = q * 0.035$
		$Pr(AP\|a = BFA, q) = q * 0.05$
		$Pr(AP\|a = BFL, q) = q * 0.04$
		$Pr(SP\|a, q) = 1 - Pr(AP\|a, q)$
	if $N_{MT} = 3$ then :	$Pr(LP\|a = BFS, q) = q * 0.01$
		$Pr(LP\|a = BFA, q) = q * 0.04$
		$Pr(LP\|a = BFL, q) = q * 0.05$
		$Pr(SP\|a, q) = (1 - Pr(LP\|a, q))/(q + 1) + penalty$
		$Pr(AP\|a, q) = 1 - Pr(LP\|a, q) - Pr(SP\|a, q)$

Table B.4: Outcome states' transition function for 10-agents environments (for the first experimental setting in Chapter 6). In all cases, $Pr(SP|a, q)$, $Pr(AP|a, q)$ and $Pr(LP|a, q)$ are eventually normalized in order to sum to one.