Bayesian Reinforcement Learning for Multiagent Systems with State Uncertainty

Christopher Amato
CSAIL
MIT
Cambridge, MA 02139
camato@csail.mit.edu

Frans A. Oliehoek
Department of Knowledge Engineering
Maastricht University
Maastricht, The Netherlands
frans.oliehoek@maastrichtuniversity.nl

ABSTRACT

Bayesian methods for reinforcement learning are promising because they allow model uncertainty to be considered explicitly and offer a principled way of dealing with the exploration/exploitation tradeoff. However, for multiagent systems there have been few such approaches, and none of them apply to problems with state uncertainty. In this paper we fill this gap by proposing two frameworks for Bayesian RL for multiagent systems with state uncertainty. This includes a multiagent POMDP model where a team of agents operates in a centralized fashion, but has uncertainty about the model of the environment. We also consider a best response model in which each agent also has uncertainty over the policies of the other agents. In each case, we seek to learn the appropriate models while acting in an online fashion. We transform the resulting problem into a planning problem and prove bounds on the solution quality in different situations. We demonstrate our methods using sample-based planning in several domains with varying levels of uncertainty about the model and the other agents' policies. Experimental results show that overall, the approach is able to significantly decrease uncertainty and increase value when compared to initial models and policies.

1. INTRODUCTION

In recent years Bayesian reinforcement learning (RL) techniques have received increased attention. Bayesian methods are promising in that, in principle, they give an optimal exploration/exploitation trade-off with respect to the prior belief. In many real-world situations, the true model may not be known, but a prior can be expressed over a class of possible models. The uncertainty over models can be explicitly considered to choose actions that will maximize expected value over models, reducing uncertainty as needed to improve performance. In practice, experience and domain knowledge can be used to construct an informed prior which can be improved online while acting in the environment.

Unfortunately, in multiagent systems, only a few Bayesian RL methods have been considered. For example, the Bayesian RL framework has been used in stochastic games [7] and factored Markov decision processes (MDPs) [28]. While either model is intractable to solve optimally, both approaches generate approximate solutions (based on the value of perfect information) which perform well in practice. Both approaches also assume the state of the problem is fully observable (or can be decomposed into fully observable components). This is a common assumption to make, but many real-world problems have partial observability due to noisy or inadequate sensors as well as a lack of communication with the other agents. To the best of our knowledge, no approaches have been proposed that can model and solve problems with partial observability. In fact, while planning in partially observable domains has had success [4, 5, 6, 13, 14, 16, 17, 19], very few multiagent RL approaches of any kind consider partially observable domains (notable exceptions, e.g., [1, 8, 20, 34]).

In this work, we propose two approaches for Bayesian learning in multiagent systems with state uncertainty: a centralized perspective using multiagent partially observable MDPs (MPOMDPs) and a decentralized perspective using best response models. In the centralized perspective, all agents share the same partially observable view of the world and can coordinate on their actions, but have uncertainty about the underlying environment. To model this problem, we extend the Bayes Adaptive POMDP (BA-POMDP) model [23, 24], which represents beliefs over possible model parameters using Dirichlet distributions. A BA-POMDP can be characterized and solved as a (possibly infinite state) POMDP. Because an MPOMDP can be mapped to a POMDP, our centralized approach can be converted into and solved as a BA-POMDP. This approach learns policies and the underlying model while acting, but makes strong assumptions about centralization of viewpoints and decisions.

As an alternative, we consider a decentralized perspective. We explore two different models that assume 1) all other agents' policies are known and fixed, 2) all other agent policies are fixed, but unknown. In both scenarios, we assume there is uncertainty about the underlying environment model. To represent and solve these models we show how the BA-POMDP approach can be extended using appropriate distributions over other agent policies. In both the centralized and decentralized perspectives, we propose a Bayesian approach to online learning which represents the initial model and the initial policies for the other agents using priors and updates probability distributions over these models as the agent acts in the real world. In many real-world scenarios, these approaches should be able to quickly learn a high-quality policy while acting online.

In Section 2, we describe the POMDP, MPOMDP and Dec-POMDP [6] models as well as summarize the BA-POMDP approach. We then introduce the BA-MPOMDP model in Section 3 and discuss the theoretical results that transfer from the BA-POMDP case, allowing solution quality to be

Appears in *The Eighth Annual Workshop on Multiagent Sequential Decision-Making Under Uncertainty* (*MSDM-2013*), held in conjunction with *AAMAS*, May 2013, St. Paul, Minnesota, USA.

bounded when solving a finite approximation. In Section 4, we describe the best-response models and extend theoretical results to include uncertainty over the other agents' policies. In Section 5, we present proof-of-concept experimental results showing that model uncertainty and solution quality can be improved over a small number of learning episodes when compared to priors over models and policies. We discuss related work in Section 6 and conclude in Section 7.

2. BACKGROUND

This section provides a concise description of the relevant frameworks for multiagent planning under uncertainty as well as the previous work on Bayesian RL for POMDPs.

2.1 POMDPs, MPOMDPs, & Dec-POMDPs

Dec-POMDPs form a framework for multiagent planning under uncertainty. We will reserve this name for the setting where there is no explicit communication (e.g., no sharing of observations) between agents. This means that each agent will act based only on its individual observations. Formally, a Dec-POMDP is a tuple $\langle I, S, \{A_i\}, T, R, \{Z_i\}, O, h\rangle$ with:

- I, a finite set of agents;
- S, a finite set of states with designated initial state distribution b_0;
- A_i, a finite set of actions for each agent, i;
- T, a set of transition probabilities: $T^{s\vec{a}s'} = \Pr(s'|s,\vec{a})$, the probability of transitioning from state s to s' when the set of actions $\vec{a}$ are taken by the agents;
- R, a reward function: $R(s,\vec{a})$, the immediate reward for being in state s and taking the set of actions $\vec{a}$;
- Z_i, a finite set of observations for each agent, i;
- O, a set of observation probabilities: $O^{\vec{a}s'\vec{z}} = \Pr(\vec{z}|\vec{a},s')$, the probability of seeing the set of observations $\vec{o}$ given the set of actions $\vec{a}$ was taken which results in state s';
- h, the horizon.

When agents are permitted to have different reward functions, this model becomes the partially observable stochastic game (POSG)[14]. Alternatively, it is possible to consider the multiagent setting where the agents are allowed to share their individual observations. In this case, we will restrict ourselves to the setting where such communication is free of noise, costs and delays and call the resulting model a multiagent POMDP (MPOMDP). Thus, an MPOMDP is a Dec-POMDP with the additional assumption of explicit communication.

A POMDP [15] can be seen as the special case of a Dec-POMDP with just one agent. Also, an MPOMDP can be reduced to a special type of POMDP in which there is a single centralized controller that takes joint actions and receives joint observations [22].

Most research concerning these models has considered the task of *planning*: given a full specification of the model, determine an optimal (joint) policy (e.g., [6, 14]). However, in many real-world applications, the model is not (perfectly) known in advance, which means that the agents have to learn about their environment during execution. This is the task considered in (multiagent) *reinforcement learning (RL)* [27].

2.2 Bayesian RL for POMDPs

A fundamental problem in RL is that it is difficult to decide whether to try new actions in order to learn about the environment, or to exploit the current knowledge about the rewards and effects of different actions. In recent years, Bayesian RL methods have become popular because they potentially can provide a principled solution to this exploration/exploitation trade-off [9, 11, 12, 21, 30].

In particular, we consider the framework of Bayes-Adaptive POMDPs [23, 24]. This framework utilizes Dirichlet distributions to model uncertainty over transitions, $T^{sas'}$, and observations, $O^{as'z}$ (typically assuming the reward function is chosen be the designer and thus known). In particular, if the agent could observe both states and observations, it could maintain vectors ϕ and ψ of counts for transitions and observations respectively. That is, $\phi^a_{ss'}$ is the transition count representing the number times state s' resulted from taking action a in state s and $\psi^a_{s'z}$ is the observation count representing the number of times observation z was seen after taking action a and transitioning to state s'.

While the agent cannot observe the states and has uncertainty about the actual count vectors, this uncertainty can be represented using the regular POMDP formalism. That is, the count vectors are included as part of the hidden state of a special POMDP, called BA-POMDP. Formally, a BA-POMDP is a tuple $\langle S_{BP}, A, T_{BP}, R_{BP}, Z, O_{BP}, h\rangle$ with

- S_{BP}, the set of states $S_{BP} = S \times \mathcal{T} \times \mathcal{O}$;
- A, a finite set of actions;
- T_{BP}, a set of state transition probabilities;
- R_{BP}, a reward function;
- Z, a finite set of observations;
- O_{BP}, a set of observation probabilities;
- h, the horizon.

We discuss these components in more detail below.

First, we point out that actions and observations remain the same as in case there was no uncertainty about the transition and observation function (i.e., the same as in the regular POMDP). However, as mentioned, the state of the BA-POMDP now includes the Dirichlet parameters: $s_{BP} = \langle s, \phi, \psi\rangle$ and the set of states $S_{BP} = S \times \mathcal{T} \times \mathcal{O}$ where $\mathcal{T} = \{\phi \in \mathbb{N}^{|S||A||S|}\}$ is the space of all possible transition parameters where each state action pair is visited at least once. Similarly $\mathcal{O} = \{\psi \in \mathbb{N}^{|S||A||Z|}\}$ is the space of all possible observation parameters.[1]

Given a pair of count vectors ϕ,ψ, we can define the expected transition and observation probabilities as:

$$T^{sas'}_{\phi} = \mathbf{E}[T^{sas'}|\phi] = \frac{\phi^a_{ss'}}{N^{sa}_{\phi}}, \quad O^{as'z}_{\psi} = \mathbf{E}[O^{as'z}|\psi] = \frac{\psi^a_{s''z}}{N^{as'}_{\psi}},$$

where $N^{sa}_{\phi} = \sum_{s''} \phi^a_{ss''}$, and $N^{as'}_{\psi} = \sum_{z'} \psi^a_{s'z'}$.

Remember that these count vectors are not observed by the agent, since that would require observations of the state. The agent can only maintain belief over these count vectors. Still, when interacting with the environment, *the ratio of the true—but unknown—count vectors will converge to coincide with the true transition and observation probabilities in expectation.* It is important to realize, however, that this convergence of count vector ratios does not directly imply learnability by the agent: even though the ratio of the count vectors specified by the true hidden state will converge, *the agent's belief over count vectors might not.*

The expected transition and observation probabilities can be used to define the transition and observation model of

[1]Note that at least one of the counts per Dirichlet parameter vector needs to be non-zero.

the BA-POMDP. In particular, the transition probabilities $P(\langle s', \phi', \psi' \rangle | \langle s, \phi, \psi \rangle, a)$ can be defined using a vector $\delta^a_{ss'}$ which is 1 at the index of a, s and s' and 0 otherwise. Similarly, for observations, we define $\delta^a_{s'z}$ to be a vector that has value 1 at the index a, s' and z and 0 otherwise. The resulting transition and observation models are

$$T_{BP}((s,\phi,\psi),a,(s',\phi',\psi')) = \begin{cases} T^{sas'}_{\phi} O^{as'z}_{\psi} & \text{if } \phi' = \phi + \delta^a_{ss'} \text{ and } \psi' = \psi + \delta^a_{s'z} \\ 0 & \text{otherwise} \end{cases} \quad (2.1)$$

$$O_{BP}((s,\phi,\psi),a,(s',\phi',\psi'),z) = \begin{cases} 1 & \text{if } \psi' = \psi + \delta^a_{s'z} \\ 0 & \text{otherwise} \end{cases} \quad (2.2)$$

Note that the observation model is now deterministic as the observation expectations are in the transition function: in T_{BP}, z is determined by $\psi' - \psi = \delta^a_{s'z}$.

The reward model remains the same (since it is assumed to be known), $R_{BP}((s,\phi,\psi),a) = R(s,a)$. An initial state distribution b_0 as well as initial count vectors ϕ_0 and ψ_0 are also assumed.

Notice that the BA-POMDP model described above has an infinite number of states if we allow ourselves to consider all possible count vectors for transitions and observations. This infinite state representation makes performing belief updates and solving the BA-POMDP impossible without sampling. Fortunately, Ross et al. prove that the solution quality can be bounded when considering a finite number of count vectors [24]:

THEOREM 1. *Given any BA-POMDP, $\epsilon > 0$ and horizon h, it is possible to construct a finite POMDP by removing states with count vectors ϕ,ψ that have $N^{s,a}_{\phi} > N^{\epsilon}_{S}$ or $N^{as'}_{\psi} > N^{\epsilon}_{Z}$ for suitable thresholds N^{ϵ}_{S}, N^{ϵ}_{Z}.*

However, even with finite count vectors, the BA-POMDP formulation remains very large, necessitating sample-based planning approaches to provide solutions. For example, the approach of Ross et al. used an online planning approach that determined the best action to take at a given belief by performing dynamic programming using a simulator for a small horizon, updating the belief (approximately) after taking that action and receiving an observation and continuing this process until the end of the of the problem is reached. Different methods for updating the belief were used such as Monte Carlo sampling and heuristics for limiting the number of belief states considered.

3. BA-MPOMDPS

In this section, we extend the BA-POMDP to the multiagent setting. As mentioned in Section 2 it is well-known that, under the assumption of instantaneous communication without noise or costs, a Dec-POMDP can be reduced to an MPOMDP, which can then be treated as a single agent model. In the same way, for a multiagent setting in which there are uncertainties about the model, we propose to treat the problem as a *BA-MPOMDP*. A BA-MPOMDP can be seen as a BA-POMDP where the actions are joint actions and the observations are joint observations. Due to this correspondence, the theoretical results related to BA-POMDPs also apply to the BA-MPOMDP model. The BA-MPOMDP model is in principle applicable in any multiagent RL setting where there is such instantaneous communication.

3.1 The Model

Formally, a BA-MPOMDP is a tuple $\langle I, S_{BM}, \{A_i\}, T_{BM}, R_{BM}, \{Z_i\}, O_{BM}, h \rangle$ with:

- I, a finite set of agents;
- S_{BM}, states $S \times \mathcal{T} \times \mathcal{O}$ with initial state distribution b_0 and initial counts ϕ_0 and ψ_0;
- A_i, a finite set of actions for each agent, i;
- T_{BM}, a set of state transition probabilities as defined below;
- R_{BM}, a reward function as defined below;
- Z_i, a finite set of observations for each agent, i;
- O_{BM}, a set of observation probabilities as defined below;
- h, the horizon.

The framework is very similar to the POMDP case, but actions, a, now become joint actions, $\vec{a}$, and observations, z's, become joint observations, $\vec{z}$. This means that a BA-MPOMDP is specified using count vectors $\phi^{\vec{a}}_{ss'}$ and $\psi^{\vec{a}}_{s'\vec{z}}$, from their respective spaces: $\mathcal{T} = \{\phi \in \mathbb{N}^{|S||A||S|}\}$ is the space of all possible transition counts and similarly $\mathcal{O}$ is the space of all possible observation parameters $\mathcal{O} = \{\psi \in \mathbb{N}^{|S||A||Z|}\}$.

Each pair of vectors has associated expected transition and observation probabilities $T^{s\vec{a}s'}_{\phi} = \phi^{\vec{a}}_{ss'}/N^{s\vec{a}}_{\phi}$, $O^{s'\vec{a}z}_{\psi} = \psi^{\vec{a}}_{s'z}/N^{\vec{a}s'}_{\psi}$. These in turn are used to specify the transition and observation model:

$$T_{BM}((s,\phi,\psi),\vec{a},(s',\phi',\psi')) = \begin{cases} T^{s\vec{a}s'}_{\phi} O^{s'\vec{a}z}_{\psi} & \text{if } \phi' = \phi + \delta^{\vec{a}}_{ss'} \text{and } \psi' = \psi + \delta^{\vec{a}}_{s'z} \\ 0 & \text{otherwise} \end{cases} \quad (3.1)$$

$$O_{BM}((s,\phi,\psi),\vec{a},(s',\phi',\psi'),z) = \begin{cases} 1 & \text{if } \phi' = \phi + \delta^{\vec{a}}_{ss'} \text{and } \psi' = \psi + \delta^{\vec{a}}_{s'z} \\ 0 & \text{otherwise} \end{cases} \quad (3.2)$$

The reward model is given as $R_{BM}((s,\phi,\psi),\vec{a}) = R(s,\vec{a})$

3.2 Solution Methods

BA-MPOMDP formalism also suffers from an infinite state space, since there can be infinitely many count vectors. However, also in the multiagent case, it is possible to create a finite approximate model

THEOREM 2. *Given any BA-MPOMDP, $\epsilon > 0$ and horizon h, it is possible to construct a finite POMDP by removing states with count vectors ϕ,ψ that have $N^{s,a}_{\phi} > N^{\epsilon}_{S}$ or $N^{as'}_{\psi} > N^{\epsilon}_{Z}$ for suitable thresholds N^{ϵ}_{S}, N^{ϵ}_{Z} that depend linearly on the number of states and joint observations, respectively.*

PROOF. Since an MPOMDP is a special case of POMDP, the BA-MPOMDP is a special case of BA-POMDP, thus this follows directly from Theorem 1. □

While this result is straightforward, the interesting part is that while N^{ϵ}_{Z} does depend on the number of joint observations, the count thresholds do not have any dependence on the number of joint actions. Therefore, for problems with few observations per agent, constructing this approximation might be feasible, even if there are many actions.

Of course, in general even a finite approximation of a BA-MPOMDP is intractable to solve optimally. Fortunately, online sample-based planning approaches in principle apply to this setting too. That is, the team of agents could perform online planning by considering a small finite horizon. After taking the joint action that resulted from the online planning phase, the environment makes a transition, the agents get observations, these observations are synchronized via communication and every agent computes the new 'joint belief'. Then this process repeats, etc.

The actual planning could take place in a number of ways: one agent could be designated the planner, which would require this agent to broadcast the computed joint action. Alternatively, each agent can in parallel perform an identical planning process (by, in the case of randomized planning, syncing the random number generators). Then each agent will compute the same joint action and execute its component. An interesting direction of future work is whether the planning itself can be done more effectively by distributing the task over the agents.

However, in the above, there is an additional bottleneck compared to the BA-POMDP: the number of joint actions and joint observations is exponential in the number of agents.

4. BA-BRM

In many real-world scenarios, instantaneous, noise and cost free communication may not be possible or practical. Similarly, in competitive domains, this type of communication often does not make sense. As a result, agents must learn during execution based solely on their own local information. In this section, we describe different way of applying Bayesian RL techniques in multiagent systems by giving a subjective description of the problem. That is, we describe the problem from a single agent's perspective by defining its *best-response model (BRM)*. We propose a Bayesian approach to online learning which represents the initial model and the initial policies for the other agents using priors and updates probability distributions over these models as the agent acts in the real world.

4.1 Best-Response Models

When making a subjective model from the perspective of a given agent, there are a number of assumptions one can make about uncertainty: First, we can assume that the agent is uncertain about only the transition and observation functions, but certain about the policies of the other agents. Second, we can assume that it also is uncertain about the other agents' policies, but that these are fixed. Finally, we can also assume that the other agents in turn are adaptive. We discuss these different assumptions in the subsections below, focusing on the first two. It is also possible that the agent is certain about T,O but uncertain about the policy of other agent. For such cases, the model we introduce for the second setting applies. Alternatively, such a setting might also be modeled using an I-POMDP (see Section 6).

4.2 Transition & Observation Uncertainty

Under the first assumption, agent i is uncertain about the transition and observation functions, but knows the (deterministic) policy π_j for all other agents.[2] Since those π_j map observation histories $\omega_j^t = \left(z_j^1,\ldots,z_j^t\right)$ to actions a_j^t at time t, agent i can model this situation using an augmented POMDP [16], which we will call the *best-response model*, in which states are tuples $\left\langle s,\vec{\omega}_{-i}^t\right\rangle$ of nominal states s and observation histories of other agents $\vec{\omega}_{-i}^t = \left\langle \vec{\omega}_1^t \ldots \vec{\omega}_{i-1}^t \vec{\omega}_{i+1}^t \ldots \vec{\omega}_n^t \right\rangle$. Given that this is a POMDP, we can incorporate uncertainty about transition and observation models by transforming it to its Bayes adaptive variant.

Formally, we define a *BA-BRM* as a tuple $\langle S_{BB},A_i,T_{BB},R_{BB},Z_i,O_{BB},h\rangle$ where

- A_i,Z_i are the sets of actions and observations of agent i;
- S_{BB} is the set of states $\left\langle s,\vec{\omega}_{-i}^t,\phi,\psi\right\rangle$ where ϕ is the vector of counts $\phi^a_{s\vec{\omega}s'\vec{\omega}'}$ and ψ is the vector of counts $\psi^a_{s'\vec{\omega}'z'}$;
- T_{BB} is the transition function (see below);
- R_{BB} is the reward function defined as $R_{BB}(s,\vec{\omega}_{-i}^t,\phi,\psi,a_i) = \sum_{a_{-i}} \pi_{-i}(a_{-i}|\vec{\omega}_{-i}^t)R(s,a_i,a_{-i})$;
- O_{BB} is the observation function. As earlier $O_{BB}((s,\vec{\omega}_{-i}^t,\phi,\psi),a_i,(s',\vec{\omega}_{-i}^{t+1},\phi',\psi'))$ is 1 iff the count vectors add correctly;
- h is the horizon.

The transition function is defined as

$$T_{BB}((s,\vec{\omega}_{-i}^t,\phi,\psi),a_i,(s',\vec{\omega}_{-i}^{t+1},\phi',\psi')) = \begin{cases} T_\phi^{s\vec{\omega}as'\vec{\omega}'} O_\psi^{s'\vec{\omega}'az} & ,\phi' = \phi + \delta^a_{s\vec{\omega}s'\vec{\omega}'},\ \psi' = \psi + \delta^a_{s'\vec{\omega}'z} \\ 0 & \text{otherwise.} \end{cases}$$

where we dropped sub- and superscripts that are clear from context. In this equation, the expected transition and observation functions are defined as

$$T_\phi^{s\vec{\omega}as'\vec{\omega}'} = \frac{\phi^a_{s\vec{\omega}s'\vec{\omega}'}}{\sum_{s''\vec{\omega}''}\phi^a_{s\vec{\omega}as''\vec{\omega}''}} \tag{4.1}$$

$$O_\psi^{s'\vec{\omega}',a,z} = \frac{\psi^a_{s'\vec{\omega}',z}}{\sum_{z'}\psi^a_{s'\vec{\omega}',z'}} \tag{4.2}$$

The former count ratio converges to $P(s',z_{-i}|s,a_i,a_{-i})$ (where a_{-i} is the action profile specified by π_{-i} for $\vec{\omega}_{-i}$) which is the true probability of s',$\vec{\omega}_{-i}^{t+1}$ given s,$\vec{\omega}_{-i}^t$,a_i and π_{-i} for the true, but unknown, count vector ϕ^*. The latter ratio, for ψ^*, converges to $P(z_i|a_i,\vec{a}_{-i},s')$, the true probability of receiving observation z_i given s',$\vec{\omega}_{-i}^{t+1}$,a_i and π_{-i}.

We point out that for this formulation, all the BA-POMDP theory holds even with the inclusion of other agent histories as part of the state information. Nevertheless, this model assumes the policies of other agents are *fixed, known* and *deterministic*. This latter assumption can be removed. When the other agents use a *stochastic policy,* those π_j map *action-observation histories* $h_j^t = \left(a_j^0 z_j^1 \ldots a_j^{t-1} z_j^t\right)$ to actions a_j^t. For this case, we can trivially adapt the BA-BRM by replacing the $\vec{\omega}_{-i}^t$ by $\vec{h}_{-i}^t$.

4.3 Policy Uncertainty

The substitution of $\vec{\omega}_{-i}^t$ by $\vec{h}_{-i}^t$ for stochastic policies brings an interesting insight: two subsequent states $(s,\vec{h}_{-i}^t)$ and $(s',\vec{h}_{-i}^{t+1})$ specify what actions the other agents took in the previous step (since those are specified in the action-observation histories). As such, counting these transitions, in general

[2]These policies can be represented as look-up tables or be computational procedures themselves. We do make the restriction, however, that they do not depend on the policy followed by agent i.

may also allow us the learn about the policies of others if we have uncertainty about them.

That is, the expected transition can be calculated as

$$T_{\phi}^{s\vec{h}as'\vec{h}'} = \frac{\phi^a_{s\vec{h}s'\vec{h}'}}{\sum_{s'',\vec{h}''} \phi^a_{s\vec{h}s''\vec{h}''}}$$

with $\vec{h} = \vec{h}^t$ and $\vec{h}' = \vec{h}^{t+1}$, and therefore the true count vector ratio will converge to the probability

$$P(s^{t+1},\vec{h}^{t+1}|s^t,\vec{h}^t_{-i},a^t_i) = P(s^{t+1},(\vec{h}^t_{-i},\vec{a}^t_{-i},\vec{z}^{t+1}_{-i})|s^t,\vec{h}^t_{-i},a^t_i)$$
$$= \pi_{-i}(\vec{a}^t_{-i}|\vec{h}^t_{-i})P(s^{t+1}|s^t,\vec{a}^t)\sum_{z_i^{t+1}} P(\vec{z}^{t+1}|s^{t+1},\vec{a}^t) \quad (4.3)$$

Note that we only need to consider h' if it includes h. This value includes the probabilities induced by the policies of the other agents, allowing uncertainty with respect to the other agents' policies to also be represented.

An interesting aspect of this formulation is that it can be used to bound the loss of computing a best response to one particular policy while in fact the agent uses a different one. To show this, we assume that there is a single other agent and that for two policies π^x_j,π^y_j of agent j we have that

$$\forall_{a_j h_j} \quad \left|\pi^x_j(a_j|h_j) - \pi^y_j(a_j|h_j)\right| \leq \epsilon. \quad (4.4)$$

Assume that π^x_j is the true policy of agent j. In that case the ϕ count vectors converge to some ϕ^*_x that satisfies

$$\forall_{shas'h'} \quad \frac{\phi^{a,x}_{shs'h'}}{\mathcal{N}^{a,x}_{sh}} = \pi^x_j(a_j|h^t_j)P(s^{t+1},z^{t+1}_j|s^t,a^t_i,a^t_j)$$

(where $\mathcal{N}^a_{sh}$ denotes the normalization constant) while, when π^y_j is the true policy, these count ratios converge to

$$\pi^y_j(a_j|h^t_j)P(s^{t+1},z^{t+1}_j|s^t,a^t_i,a^t_j) = \frac{\phi^{a,y}_{shs'h'}}{\mathcal{N}^{a,y}_{sh}}$$

Additionally, we have that, independently of π_j the policy of the other agent, the ψ count ratios of the true hidden state converge to

$$P(z^{t+1}_i|a^t_i,a^t_j,s^{t+1},z^{t+1}_j) = \frac{\psi^a_{s'\vec{h}',z}}{\mathcal{N}^a_{s'\vec{h}'}}$$

Note that here $\vec{h}'_j$ specifies both a^t_j and z^{t+1}_j.

It follows that, upon convergence of these ratios, we have

$$\left|\frac{\phi^{a,x}_{shs'h'}}{\mathcal{N}^{a,x}_{sh}}\frac{\psi^a_{s'\vec{h}',z}}{\mathcal{N}^a_{s'\vec{h}'}} - \frac{\phi^{a,y}_{shs'h'}}{\mathcal{N}^{a,y}_{sh}}\frac{\psi^a_{s'\vec{h}',z}}{\mathcal{N}^a_{s'\vec{h}'}}\right|$$
$$= \left|\left(\pi^x_j(a_j|h_j) - \pi^y_j(a_j|h_j)\right)P(s^{t+1},z^{t+1}_j|s^t,a^t_i,a^t_j)\right|$$
$$P(z^{t+1}_i|a^t_i,a^t_j,s^{t+1},z^{t+1}_j)$$
$$\leq \epsilon P(s^{t+1},z^{t+1}_i,z^{t+1}_j|s^t,a^t_i,a^t_j)$$

Moreover,

$$\sum_s\sum_z \left|\frac{\phi^{a,x}_{shs'h'}}{\mathcal{N}^{a,x}_{sh}}\frac{\psi^a_{s'\vec{h}',z}}{\mathcal{N}^a_{s'\vec{h}'}} - \frac{\phi^{a,y}_{shs'h'}}{\mathcal{N}^{a,y}_{sh}}\frac{\psi^a_{s'\vec{h}',z}}{\mathcal{N}^a_{s'\vec{h}'}}\right| \leq$$
$$\sum_s\sum_z \epsilon P(s^{t+1},z^{t+1}_i,z^{t+1}_j|s^t,a^t_i,a^t_j) = \epsilon \quad (4.5)$$

That is, given that the difference between two policies is bounded, the difference between count vector ratios, and thus expected transition probabilities that they will induce is bounded as well. This can subsequently be used to bound the loss in value when optimizing against a wrong policy.

THEOREM 3. *Given ϕ^*_x, ϕ^*_y and ψ^*, the converged count vectors corresponding to two policies π^x_j,π^y_j of agent j that satisfy (4.4), then, for all stages-to-go t, then for any t-steps-to-go policy for agent i, the associated values are bounded:*

$$\max_{s\in S}\left|\alpha_t(s,\phi^*_x,\psi^*) - \alpha_t(s,\phi^*_y,\psi^*)\right| \leq \frac{\epsilon(\gamma-\gamma^t)\left\|R\right\|_\infty}{(1-\gamma)^2} \quad (4.6)$$

PROOF. Here, $\|R\|_\infty$ is the reward with greatest magnitude and γ is the discount factor. The proof is given in the appendix. □

The implication of this theorem is that if we compute a best-response against some policy π^x_j which differs from π^y_j, the true policy used by agent j, by at most ϵ, then that loss in value is bounded by (4.6). While this relates to bounds for model equivalence [33], no bounds on the loss in value for different policies of the other agent have been proposed. Also, we expect that these bound could have big implications for work on influence based abstraction [31], and, in particular, using approximate influences.

Finally, we point out that, when other agents are adaptive, the assumption of a unknown, but fixed policy is violated. In fact there are inherent limits to what can be learned by Bayesian learners that perform a best-response [32]. Nevertheless, methods such as Q-learning have been shown to be effective in such domains [25, 29]. We expect that it might be possible to deal with this issue by, for instance, performing discounting of counts while learning during execution. This is a fertile area for future research.

4.4 Solving BA-BRMs

BA-BRMs have an intractable (infinite) number of parameters, but again, the theory from [23] applies such that we can ensure that a solution that is boundedly optimal can be generated using a finite model.

THEOREM 4. *Given any BA-MPOMDP, $\epsilon > 0$ and horizon h, it is possible to construct a finite POMDP by removing states with count vectors ϕ,ψ that have $N^{s,a}_\phi > N^\epsilon_S$ or $N^{as'}_\psi > N^\epsilon_Z$ for suitable thresholds N^ϵ_S, N^ϵ_Z that depend linearly on the number of augmented states and individual observations, respectively.*

PROOF. Again, since a BRM is a special case of POMDP, the BA-BRM is a special case of BA-POMDP, thus this follows directly from Theorem 1. □

The result itself is straightforward. In this case, however, N^ϵ_Z only depends on the size of the individual observation set. However, this comes at a cost, since $N^{as'}_\psi$ now depends linearly on the number of *augmented* states, which is $O(|S|(|A_j||Z_j|)^{h(n-1)})$. In case of a single other agent, this means that complexity dependence on joint observations in the BA-MPOMDP is replaced by an exponential dependence on the horizon.

Even in this case, the model will often be large and difficult to learn. Sample-based planning can also be used in this scenario by transforming the BA-BRM into a BA-POMDP and solving it. The number of states may become large, but the number of number of actions in the BA-BRM remains the same as in the original Dec-POMDP model (unlike the

BA-MPOMDP). Communication can also be incorporated to coordinate the learning and improve its efficiency.

Prior distributions over environment and agent models can be represented as initial count vectors. As is clear from (4.3), the ϕ ratios correspond to (should converge to) the true probability $\pi_{-i}(\vec{a}^t_{-i}|\vec{h}^{t-1}_{-i})P(s'|s,\vec{a})\sum_{z_i} P(\vec{z}|s',\vec{a})$. If these probabilities can be estimated, the count vectors can be set to ratios representing this quantity. Then, the confidence in this estimation can be reflected in a scaling factor of the various counts. In this way, different aspects of the agent and environment models can have different parameters and confidence based on knowledge of the problem. In the absence of domain knowledge a uniform prior with small counts can be utilized.

5. EXPERIMENTAL EVALUATION

We performed a preliminary empirical evaluation of the BA-MPOMDP and BA-BRM models.

5.1 Sample-Based Planning

To test performance of the different models, we implemented a simulation of agents that interact with an environment over a number of episodes ($N_{episodes}$). Importantly, at the end of each episode, the belief over states is reset to the initial belief, but the belief over count vectors is maintained. That way, the agents learn across all the episodes.

The agents use an on-line sample-based planner to act: in each stage of each episode, the agent(s) perform sample-based planning in order to select a (joint) action. This action is subsequently executed, a transition and observation is sampled, the agents update their (joint) beliefs, and a new round of online planning is initiated, etc. The beliefs are represented using a particle filter (with $N_{particles}$ particles).

As the sample-based planner, we use Monte Carlo planning: the expected value for each (joint) action is evaluated using a number ($N_{samples}$) of Monte Carlo rollouts (i.e., using a random (joint) action selection) up to a particular lookahead planning horizon (which can be shorter than h). This planner also uses particle-based belief representations.[3]

5.2 Experimental Setup

We evaluate our BA-MPOMDP and BA-BRM approaches by performing online learning using the common decentralized tiger benchmark [16] for horizon 3. In this cooperative problem, two agents have the choice of listening or opening one of two doors. If both agents listen, they each hear a noisy signal of the tiger's location (0.85 probability of hearing the correct location). If either agent opens a door, there is a shared penalty for opening the door with the tiger and a shared reward for opening the other door (which has treasure behind it). If both agents choose to open the same door, they receive a reduced penalty or a greater reward. Whenever a door is opened the tiger transitions uniformly to be behind one of the doors.

For illustration, we show only error in the observation model, but unlike Ross et al. we do not assume the transition function is known. Instead, we assume we have high confidence in the transition parameters and reflect this in the transition count vectors, (ϕ were initialized as 1000 times the true probability as discussed above). The observation priors that were used are listed in Table 1.

For the BA-BRMs, the optimal pair of (deterministic) policies was found (using [2]) and one of these policies was used as the fixed policy of the other agent. This policy represents the other agent listening until it has heard the tiger on the same side twice and then opening the appropriate door. The count vectors for the transitions were set similarly to those above (with 1000 times the true transition probabilities) and the observation count vectors were set as in Table 1.

To determine performance, we consider model error and value. Model error is calculated as a sum of the L1 distances between the correct and estimated probabilities weighted by the probability associated with the count vectors (as described in [23]). The error and value produced are averaged over a number of simulations ($N_{simulations}$). Note that at the start of each simulation also the count vectors are reset, so there is no learning across simulations — these are only to determine average values. These experiments were run on a 2.5 GHz Intel i7 using a maximum of 2GB of memory.

5.3 Results

The errors in the observation functions for all methods are shown in Figure 1. For the BA-MPOMDP formulation, for which we have performed $N_{simulations} = 50$ simulations with $N_{episodes} = 50$, $N_{particles} = 200$, $N_{samples} = 500$, and a lookahead horizon of two. We see that the error decreases sharply and then decreases more slowly. This error will likely continue to improve if more episodes are completed. The nonmonotonic improvement is due to randomness in the sample values and the fact that only 50 simulations were used to average the data. Value for this problem from the initial belief and count vectors is approximately -6 (the value of listening at each step) and the optimal value for this problem with known parameters is 13.015 (listening twice and then collectively opening the appropriate door if the same observation was heard twice). After 50 episodes, the value was estimated and was found to be -4.5.

For the BA-BRM experiments, we used the following parameters: $N_{simulations} = 50$ simulations with $N_{episodes} = 50$, $N_{particles} = 100$, $N_{samples} = 200$ and a lookahead horizon of two. It is worth noting that the optimal value of this version of the problem with known parameters is 5.19. Notice that this is less than in the MPOMDP case because the agents can no longer coordinate to always open the same door at the same time.

The model error for BA-BRM with a known other agent policy is also shown in Figure 1. Similar to the BA-MPOMDP case, we see a decrease in model error over as the number of episodes increases. Sampling to determine the value produces a value of approximately -3.78 with the initial count vectors and -2.03 after 50 episodes.

For BA-BRM with unknown other agent policy, the settings were: $N_{simulations} = 10$ simulations with $N_{episodes} = 50$, $N_{particles} = 50$, $N_{samples} = 50$ and a lookahead horizon of one. This reduction in samples is due to the increased time required for updating and evaluation by considering unknown actions in the other agent histories. The value produced in this model after learning is unchanged from the initial value. It is likely that increased sampling would improve both uncertainty and the resulting value, suggesting the need for more computationally efficient methods.

[3]When the number of reachable states was small, we used a closed-form description of beliefs to speed up planning.

joint action	joint observation	count
both listen	both correct	5
	1 correct	2
	both incorrect	1
other	all	2

action	observation	count
listen	correct	3
	incorrect	2
open	all	2

Table 1: Observation prior counts for the BA-MPOMDP (left) and BA-BRM (right).

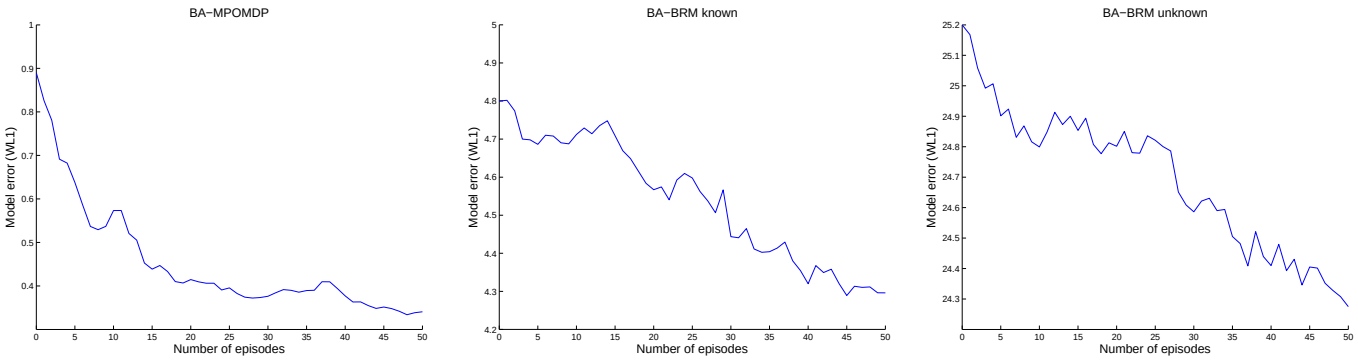

Figure 1: Model error (weighted L1) in the 2-agent tiger problem for the BA-MPOMDP and BA-BRM with known and unknown other agent policies

6. RELATED WORK

The BA-BRM is closely related to the framework of I-POMDPs [13, 33] in that it is also a model for describing an interactive setting from the perspective of a protagonist agent. There are a few differences, however. First, the I-POMDP treats another agent as an entity that (potentially) also reasons about the protagonist agent, but at a lower level. In contrast, the BA-BRM just considers the other agent as executing a particular policy. As such, there are no difficulties with infinite recursions of beliefs. Second, the I-POMDP does not consider uncertainty about the actual transition and observation model. However, we point out that since an I-POMDP is a special type of POMDP, it is possible to consider Bayes adaptive extensions that consider such uncertainty [18].

Other work that is out of the scope of this paper has developed other learning techniques for Dec-POMDPs. These approaches include model-free reinforcement learning methods using gradient-based methods to improve the policies [10, 20] and learning using local signals and modeling the remaining agents as noise [8]. Another approach has utilized communication and sample-based planning to generate best-response policies [3].

7. CONCLUSIONS

In this paper, we presented the first set of approaches for Bayesian reinforcement learning for multiagent systems with state uncertainty. We consider a centralized perspective where the team of agents is modeled as a multiagent POMDP, allowing Bayesian RL techniques from the POMDP literature to be applied. Because the centralized perspective assumes centralization or full communication between the agents, we also consider a decentralized perspective. In this perspective, we explore cases in which an agent knows the fixed policies of the others, but has uncertainty over the environment model and when an agent has uncertainty over the policies of the fixed agents and environment models. Each of these cases reflects realistic assumptions about real-world scenarios. We present proofs bounding the solution quality under these different assumptions. Our experimental results show a proof of concept for Bayesian RL in multiagent systems with state uncertainty, demonstrating how an agent can improve model estimates and performance while acting online.

These approaches can serve as the basis for many future work directions in multiagent learning. This could include the use of (delayed or noisy) communication to allow the best response model to update parameters based on information from the other agents. Similarly, additional domain or policy assumptions could be imposed to improve scalability. For instance, with transition and observation independence [4], models of others can be represented as mappings from local states to actions and using finite-state controllers [5], parameters can be limited by the size of the controllers. We also expect more efficient solutions for these models could be generated by more sophisticated sample-based planning methods such as [26], allowing greater scalability to larger domains and a larger number of agents. Lastly, even though our experiments consisted of a cooperative domain, our approach extends to competitive models and we plan to test its effectiveness in those problems as well.

Acknowledgements

Research supported in part by AFOSR MURI project #FA9550-091-0538 and in part by NWO CATCH project #640.005.003.

APPENDIX

An α-vector for a BA-POMDP, can be expressed as the immediate reward for the specified action a plus value of next stage vectors for some mapping $\alpha' : \mathcal{Z} \rightarrow \Gamma_{t-1}$ [24]. Therefore, for any policy

π_t, for all states s, we have that

$$
\begin{aligned}
&|\alpha_t^{\pi_t}(s,\phi_x,\psi_x) - \alpha_t^{\pi_t}(s,\phi_y,\psi_y)| \\
=&\gamma \Bigg| \sum_{s'} \sum_{z} \Big[\frac{\phi_x^{sas'}}{\mathcal{N}_x^{sa}} \frac{\psi_x^{as'z}}{\mathcal{N}_x^{as'}} (\alpha'(z)(s',\phi_x^\delta,\psi_x^\delta) - \alpha'(z)(s',\phi_y^\delta,\psi_y^\delta)) \\
&- \Big(\frac{\phi_y^{sas'}}{\mathcal{N}_y^{sa}} \frac{\psi_y^{as'z}}{\mathcal{N}_y^{as'}} - \frac{\phi_x^{sas'}}{\mathcal{N}_x^{sa}} \frac{\psi_x^{as'z}}{\mathcal{N}_x^{as'}} \Big) \alpha'(z)(s',\phi_y^\delta,\psi_y^\delta) \Big] \Bigg| \\
\leq&\gamma \sum_{s'} \sum_{z} \frac{\phi_x^{sas'}}{\mathcal{N}_x^{sa}} \frac{\psi_x^{as'z}}{\mathcal{N}_x^{as'}} \Big| \alpha'(z)(s',\phi_x^\delta,\psi_x^\delta) - \alpha'(z)(s',\phi_y^\delta,\psi_y^\delta) \Big| \\
&+ \gamma \sum_{s'} \sum_{z} \Bigg| \frac{\phi_y^{sas'}}{\mathcal{N}_y^{sa}} \frac{\psi_y^{as'z}}{\mathcal{N}_y^{as'}} - \frac{\phi_x^{sas'}}{\mathcal{N}_x^{sa}} \frac{\psi_x^{as'z}}{\mathcal{N}_x^{as'}} \Bigg| \Big| \alpha'(z)(s',\phi_y^\delta,\psi_y^\delta) \Big| \\
\leq&\gamma \max_{s',z} \Big| \alpha'(z)(s',\phi_x^\delta,\psi_x^\delta) - \alpha'(z)(s',\phi_y^\delta,\psi_y^\delta) \Big| \\
&+ \frac{\gamma \|R\|_\infty}{(1-\gamma)} \sum_{s'} \sum_{z} \Bigg| \frac{\phi_y^{sas'}}{\mathcal{N}_y^{sa}} \frac{\psi_y^{as'z}}{\mathcal{N}_y^{as'}} - \frac{\phi_x^{sas'}}{\mathcal{N}_x^{sa}} \frac{\psi_x^{as'z}}{\mathcal{N}_x^{as'}} \Bigg|
\end{aligned}
$$

Now, we can substitute in (4.5) and get

$$
\begin{aligned}
&|\alpha_t^{\pi_t}(s,\phi_x,\psi_x) - \alpha_t^{\pi_t}(s,\phi_y,\psi_y)| \\
\leq&\gamma \max_{s',z} \Big| \alpha'(z)(s',\phi_x^\delta,\psi_x^\delta) - \alpha'(z)(s',\phi_y^\delta,\psi_y^\delta) \Big| + \frac{\epsilon\, \gamma \|R\|_\infty}{(1-\gamma)}
\end{aligned}
$$

We note that this holds for an arbitrary π_t and thus for arbitrary $\langle a,\alpha' \rangle$. Now, define a recurrence by via the max:

$$
\begin{aligned}
&\max_{\alpha_t,s} |\alpha_t(s,\phi_x,\psi_x) - \alpha_t(s,\phi_y,\psi_y)| \\
\leq&\gamma \max_{s,a,z} \max_{\alpha_{t-1},s'} \Big| \alpha_{t-1}(s',\phi_x^\delta,\psi_x^\delta) - \alpha_{t-1}(s',\phi_y^\delta,\psi_y^\delta) \Big| + \frac{\epsilon\gamma \|R\|_\infty}{(1-\gamma)} \\
\leq& \sum_{\tau=1}^{t-1} \gamma^\tau \frac{\epsilon \|R\|_\infty}{(1-\gamma)} = \frac{\epsilon \|R\|_\infty}{(1-\gamma)} \sum_{\tau=1}^{t-1} \gamma^\tau = \frac{\epsilon \|R\|_\infty}{(1-\gamma)} \frac{\gamma - \gamma^t}{(1-\gamma)}.
\end{aligned}
$$

A. REFERENCES

[1] S. Abdallah and V. Lesser. Multiagent reinforcement learning and self-organization in a network of agents. In *AAMAS*, pages 172–179, 2007.

[2] C. Amato, J. S. Dibangoye, and S. Zilberstein. Incremental policy generation for finite-horizon DEC-POMDPs. In *ICAPS*, pages 2–9, 2009.

[3] B. Banerjee, J. Lyle, L. Kraemer, and R. Yellamraju. Sample bounded distributed reinforcement learning for decentralized POMDPs. In *AAAI*, 2012.

[4] R. Becker, S. Zilberstein, V. Lesser, and C. V. Goldman. Solving transition-independent decentralized Markov decision processes. *Journal of AI Research*, 22:423–455, 2004.

[5] D. S. Bernstein, C. Amato, E. A. Hansen, and S. Zilberstein. Policy iteration for decentralized control of Markov decision processes. *Journal of AI Research*, 34:89–132, 2009.

[6] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

[7] G. Chalkiadakis and C. Boutilier. Coordination in multiagent reinforcement learning: A Bayesian approach. In *AAMAS*, pages 709–716, 2003.

[8] Y.-H. Chang, T. Ho, and L. P. Kaelbling. All learning is local: Multi-agent learning in global reward games. In *NIPS 16*, 2004.

[9] M. Duff. *Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massassachusetts Amherst, 2002.

[10] A. Dutech, O. Buffet, and F. Charpillet. Multi-agent systems by incremental gradient reinforcement learning. In *IJCAI*, pages 833–838, 2001.

[11] Y. Engel, S. Mannor, and R. Meir. Reinforcement learning with gaussian processes. In *ICML*, pages 201–208, 2005.

[12] M. Ghavamzadeh and Y. Engel. Bayesian actor-critic algorithms. In *ICML*, pages 297–304, 2007.

[13] P. J. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:24–49, 2005.

[14] E. A. Hansen, D. S. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. In *AAAI*, pages 709–715, 2004.

[15] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:1–45, 1998.

[16] R. Nair, D. Pynadath, M. Yokoo, M. Tambe, and S. Marsella. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *IJCAI*, pages 705–711, 2003.

[17] R. Nair, P. Varakantham, M. Tambe, and M. Yokoo. Networked distributed POMDPs: a synthesis of distributed constraint optimization and POMDPs. In *AAAI*, 2005.

[18] B. Ng, K. Boakye, C. Meyers, and A. Wang. Bayes-adaptive interactive POMDPs. In *AAAI*, 2012.

[19] F. A. Oliehoek. Decentralized POMDPs. In M. Wiering and M. van Otterlo, editors, *Reinforcement Learning: State of the Art*, volume 12, pages 471–503. Springer, 2012.

[20] L. Peshkin, K.-E. Kim, N. Meuleau, and L. P. Kaelbling. Learning to cooperate via policy search. In *UAI*, pages 489–496, 2000.

[21] P. Poupart, N. Vlassis, J. Hoey, and K. Regan. An analytic solution to discrete bayesian reinforcement learning. In *ICML*, pages 697–704, 2006.

[22] D. V. Pynadath and M. Tambe. The communicative multiagent team decision problem: Analyzing teamwork theories and models. *Journal of Artificial Intelligence Research*, 16:389–423, 2002.

[23] S. Ross, B. Chaib-draa, and J. Pineau. Bayes-adaptive POMDPs. In *NIPS 19*, 2007.

[24] S. Ross, J. Pineau, B. Chaib-draa, and P. Kreitmann. A Bayesian approach for learning and planning in partially observable Markov decision processes. *Journal of Machine Learning Research*, 12:1729–1770, 2011.

[25] T. W. Sandholm and R. H. Crites. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, 37:147–166, 1995.

[26] D. Silver and J. Veness. Monte-carlo planning in large POMDPs. In *NIPS 23*, pages 2164–2172, 2010.

[27] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[28] W. T. L. Teacy, G. Chalkiadakis, A. Farinelli, A. Rogers, N. R. Jennings, S. McClean, and G. Parr. Decentralized Bayesian reinforcement learning for online agent collaboration. In *AAMAS*, pages 417–424, 2012.

[29] G. Tesauro and J. O. Kephart. Pricing in agent economies using multi-agent Q-learning. *Autonomous Agents and Multi-Agent Systems*, 5(3):289–304, 2002.

[30] N. Vlassis, M. Ghavamzadeh, S. Mannor, and P. Poupart. Bayesian reinforcement learning. In M. Wiering and M. van Otterlo, editors, *Reinforcement Learning: State of the Art*, volume 12. Springer, 2012.

[31] S. J. Witwicki and E. H. Durfee. Influence-based policy abstraction for weakly-coupled Dec-POMDPs. In *ICAPS*, 2010.

[32] H. P. Young. *Strategic Learning and Its Limits*. Oxford University Press, 2004.

[33] Y. Zeng, P. Doshi, Y. Pan, H. Mao, M. Chandrasekaran, and J. Luo. Utilizing partial policies for identifying equivalence of behavioral models. In *AAAI*, pages 1083–1088, 2011.

[34] C. Zhang, V. Lesser, and S. Abdallah. Self-organization for coordinating decentralized reinforcement learning. In *AAMAS*, pages 739–746, 2010.