Quasi Deterministic POMDPs and DecPOMDPs

Camille Besse & Brahim Chaib-draa
DAMAS Laboratory
Department of Computer Science and Software Engineering
Laval University, G1K 7P4, Quebec (Qc), Canada
{besse,chaib}@damas.ift.ulaval.ca

ABSTRACT

In this paper, we study a particular subclass of partially observable models, called quasi-deterministic partially observable Markov decision processes (QDET-POMDPs), characterized by deterministic transitions and stochastic observations. While this framework does not model the same general problems as POMDPs, they still capture a number of interesting and challenging problems and have, in some cases, interesting properties. By studying the observability available in this subclass, we suggest that QDET-POMDPs may fall many steps in the complexity hierarchy. An extension of this framework to the decentralized case also reveals a subclass of numerous problems that can be approximated in polynomial space. Finally, a sketch of ε-optimal algorithms for these classes of problems is given and empirically evaluated.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed AI—*Multiagent systems*; G.3 [**Mathematics of Computing**]: Probability and statistics—*Markov processes*

General Terms

Design, Experimentation

Keywords

partial observability, determinism, coordination

1. INTRODUCTION

AI planning was initially conceived as a deterministic problem where a sequence of actions has to be decided in order to achieve a goal state with desirable values from an original state. This problem was thoroughly studied in AI with important contributions as A*, GRAPHPLAN, and others [13].

However, this deterministic model has strong limitations on the type of problem that can be represented. Thus, one cannot represent situations where actions have non-deterministic outcomes or where states are not completely observable. In such cases, one must resort to Markov Decisions Processes (MDPs [15]) when the state is fully observable and Partially Observable Markov Decisions Processes (POMDPs [6]) otherwise. However, with this expressiveness comes an increase of complexity, specially for POMDPs, and thus this gain in generality involves a cost in the ability to solve the final problem. For instance, POMDPs offer one of the most expressive frameworks and are thus widely used for sequential decision making under partial observability [11], but the current known algorithms scale very poorly as the planning horizon grows. This reality is even more true in a decentralized setting where each agent has to anticipate every possible past and future of other agents.

Indeed, a major difficulty of decision-theoretic domains mainly depends on the definition of observability of agents in the problem. For example, considering the different markovian models (MDPs, POMDPs) and their various cooperative multiagent extensions (MMDPs [5], MTDP [17], DEC-POMDPs [3]), the difference between *fully* and *partially* observable models truly shows the exponential increase in worst-case complexity.

Nevertheless, numbers of problems that involve partial observability have a common characteristic: they have actions with deterministic outcomes and the observation generated is also deterministic. Indeed, these problems have recently been used in many proposals for planning with incomplete information, e.g. [14], and are used for learning partially observable models [1].

These models were briefly discussed in [12], under the name of deterministic POMDPs (DET-POMDPs) for which some important theoretical results were obtained. Littman first showed that a DET-POMDP can be mapped into an MDP with an exponential number of states and then be solved with standard algorithms for MDPs. Second, he showed that optimal non-stationary policies of polynomial size can be computed in non-deterministic polynomial time and finally that optimal stationary policies can be computed in polynomial space. Since then, up to our knowledge, no paper was published on this subject except [4] that extends these results by defining a specific subclass of DET-POMDPs, that have the so-called *polynomial diameter* property, that can be solved in non-deterministic polynomial time. Bonet also linked the DET-POMDP framework to the AND/OR tree search algorithms, arguing that this type of algorithm is more efficient than standard POMDP algorithms for this subclass of POMDPs.

Given this role of DET-POMDPs in recent research and motivated by the quest of amenable models for decision making under partial observability, we extend the work of Littman and Bonet in order to bridge a part of the gap between DET-POMDPs and POMDPs, by studying the subclass of POMDPs with deterministic transitions by actions and particular stochastic observations. We thus present a specific subclass of widely used POMDPs, called quasi-DET-POMDPs (QDET-POMDPs) and two particular subclasses of partial observability. A theoretical analysis suggests that ε-approximating these subclasses falls many steps in complexity in the polynomial hierarchy. We also extend these results to the multiagent case revealing a drastic improvement of the complexity in case of decentralized decisions.

This paper is organized as follows. First, examples of challeng-

AAMAS 2010 Workshop on Multi-agent Sequential Decision-Making in Uncertain Domains, May 11, 2010, Toronto, Canada.

ing problems are given (in the next section) that motivate our research, as mono or multiagent problems. In Sect. 3, a formal definition of the models and the variants of observability are given. In Sect. 4, main theoretical results are described and the complexity of the subclass is presented for both mono and multiagent models in Sect. 5. Finally, some experimental results are presented in Sect. 6 before discussing the significance of this work in Sect. 7.

2. EXAMPLES

Many problems have been modeled as POMDPs and DET-POMDPs and had been used for developing and evaluating various algorithms for planning under uncertainty and partial information. For space reasons, we present only few examples of some problems that may be modeled as a QDET-POMDP:

Diagnosis: The aim of diagnosis is to identify one of the m states of a system (e.g. a patient) using n noisy binary tests. An instance consists of a $m \times n$ stochastic matrix T where each T_{ij} represent the probability that test j is positive in the state i. The goal is to find the sequence of tests that will identify almost surely the state of the studied system [16]. In this example, the model is quasi-deterministic since the transition is deterministic (only one state) and observations are results of each test and are thus stochastic.

Indoor Robot: Consider an indoor arm robot working on a supply chain that moves boxes from one conveyor to another. Even if its actuators are nearly deterministic, its captors of surrounding activity may be noisy and may provide stochastic observations. In this domain we may also want to coordinate several robots and humans on complexe assembling tasks.

Fault Detection Consider a synthetic deterministic system that evolves through the interaction of an agent (e.g. a coffee machine). The aim of the problem is to identify the sequence of interactions that leads the system to fail while only receiving noisy or partial observations of its internal state. This problem is also extendable at the case where multiple agents interact at the same time with the system. Here again, transitions are deterministic while observations are stochastic.

All of these problems can be modeled as QDET-POMDPs or QDET-DEC-POMDPs. Let us now see the formal definition of the deterministic POMDP and its variants.

3. MODEL AND VARIANTS

Deterministic POMDPs were initially defined as follows [12]:

DEFINITION 1. *A Deterministic Partially Observable Markov Decision Process (*DET-POMDP*) is a tuple* $\langle \mathcal{S}, \mathcal{A}, \Omega, \mathcal{T}, \mathcal{O}, \mathcal{R}, \gamma, \boldsymbol{b}^0 \rangle$*, where:*

- $\mathcal{S}$ *is a finite set of* states $s \in \mathcal{S}$*;*
- $\mathcal{A}$ *is the finite set of* actions *of the agent and* $a \in \mathcal{A}$*, denotes an action;*
- Ω *is the finite set of* observations *of the agent and* $z \in \Omega$*, denotes an observation;*
- $\mathcal{O}(z, a, s') : \Omega \times \mathcal{A} \times \mathcal{S} \mapsto \{0, 1\}$ *is the* deterministic observation function *indicating wether or not the agent gets observation* z *when the world falls in state* s' *after executing action* a*;*
- $\mathcal{T}(s, a, s') : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \{0, 1\}$ *is the* deterministic transition function *indicating wether or not making action* a *in state* s *results in state* s'*;*
- $\mathcal{R}(s, a) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ *is the* reward *perceived by the agent when the world falls into state* s *after executing action* a*;*
- γ *is the discount factor;*
- $\boldsymbol{b}^0$ *is the a priori knowledge about the state, namely the initial belief state, assumed non-deterministic, i.e. where no state has probability one.*

Note that the initial *belief state* $\boldsymbol{b}^0$, which describes the different possibilities for the initial state, is crucial. Indeed, if the initial state were known, and since the transition function is deterministic, then all the future states will also be known, and the DET-POMDP is then reduced to the well studied problem of deterministic planning in AI [13].

Compared to deterministic POMDPs, our proposed extended model presents changes on the observability function, it is called Quasi-deterministic Partially Observable Markov Decision Process and it is defined as follows:

DEFINITION 2. *A Quasi-deterministic Partially Observable Markov Decision Process (*QDET-POMDP*) is a tuple* $\langle \mathcal{S}, \mathcal{A}, \Omega, \mathcal{T}, \mathcal{O}, \mathcal{R}, \gamma, \boldsymbol{b}^0 \rangle$*, where:*

- $\mathcal{S}, \mathcal{A}, \Omega, \mathcal{T}, \mathcal{R}, \gamma, \boldsymbol{b}^0$ *are the same as in Definition 1;*
- $\mathcal{O}(z, a, s') : \Omega \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ *is the* observation function *indicating the probability of getting observation* z *when the world falls in* s' *after executing* a*;*
  Moreover, $\forall\, s' \in \mathcal{S}$, $a \in \mathcal{A}$, $\exists\, z \in \Omega$, *s.t.* $\mathcal{O}(z, a, s') \geqslant \theta > \frac{1}{2}$*, i.e. the probability of getting one of the observations is lower bounded in each state by at least one half;*

First, let us notice that θ is just a lower bound on the probability of observing each state and thus can be eventually greater in some states. Notice also that the planning horizon is not set *a priori*. This is due – as we will see in Section 4 – to an interesting convergence property of this model with some other assumptions to a very low entropy belief state after a fixed number of steps.

Second, to handle the multiagent case in both definitions, simply consider a set of agents where each agent i has its own action set $\mathcal{A}_i$ and where the joint action set $\boldsymbol{\mathcal{A}}$ is the product of all the agents' action sets. The transition and the observation functions are then just defined over the joint action set, and the condition on minimal observability is defined for all joint action $\boldsymbol{a}$ in $\boldsymbol{\mathcal{A}}$.

However, assuming that all agents have the same observability capacity (and hence the same observation space), and considering that in a QDET-DEC-POMDP there exist a most likely observation of the state whatever the chosen joint action is, in the same way as in the monoagent case, only the study of the QDET-POMDP is necessary from which we will extend to the multiagent case. For the ease of explanations, we will thus restrain the theoretical study to the monoagent case and then later in the paper, extend the results to the multiagent case.

Let us now see the optimality criteria and the variants of these models.

Optimality Criteria and Variants

As our goal is to compute a policy that permits an agent to perform *optimally*, we consider the **maxexp** optimality criteria that maximizes the expected discounted reward of a policy. The value of a

policy π is thus computed using:

$$V_\pi(\boldsymbol{b}^0) = \mathop{\mathrm{E}}_{s\sim\boldsymbol{b}^0}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}(s^t,\pi(s^t))\,\middle|\,s^0=s,\ \pi\right]$$

The considered variants of this model are related to the observation model as follows:

Unobservable models in which $|\Omega| = 1$ and thus no information is retrieved about the state. This class is a subclass of the so-called conformant problem in planning [9].

Fully Observable models in which $\Omega = \mathcal{S}$ and $\mathcal{O}(z,a,s') = 1$ iff $z = s'$. This class is exactly the classic fully observable MDPs where only the initial state is unknown (e.g. qMDP [11]).

Non-unobservable models in which $|\Omega| > 1$. This class is exactly the complement of unobservable problems. Among this class of problems, we distinguish:

- **Enough-observable models** in which $\Omega = \mathcal{S}$. This class regroups all the linear but noisy observation problems where the state itself is perceived but with an additive noise. This class regroups for example all control problems where the state is perceived through noisy sensors.
- **Factored-observable models** in which $|\Omega| = |\mathcal{X}| \times |\mathcal{D}_{\mathrm{X}}|$. Where $\mathcal{X}$ is the set of state variables and $\mathcal{D}_{\mathrm{X}}$ is the domain of variable X. The state space is then given by $\mathcal{S} = \bigtimes_{\mathrm{X}\in\mathcal{X}}\mathcal{D}_{\mathrm{X}}$. This class is similar to the previous one using additive noise but restricting the number of observations along the "dimensions" of the state space. Indeed, as the state space is assumed structured, the agent can use this structure to learn about at least one dimension at each time step. The previous class is equivalent to this class but with only one dimension.

General models which include previous cases, do not assume anything on the observation function.

As the fully observable, the unobservable and the general cases were extensively studied in the literature [13, 11], we will not consider them in the remaining of the paper. However, the enough-observable and the factored-observable cases present an interesting avenue since many of the quasi-deterministic problems mentioned earlier are very often factored or at least enough-observable.

We will show in the next section that these problems actually are easier than the general problems by bounding the history needed to identify the underlying state with high probability. Such bounds will indeed induce a complexity reduction of the problem and we will present this result in section 5.

4. THEORETICAL ANALYSIS

In this section, a lower bound on the number of steps to ensure convergence to a certain belief is given.

As mentioned earlier in the paper, a way to represent compactly the full history of observations during the planning process is the *belief state* [20]. This is a probability distribution over the states that represents the belief of the agent to be in each state through probabilities. We denote by $\boldsymbol{b}^t(s) = \Pr(s|z^t,a^t,\boldsymbol{b}^{t-1})$ the probability of being in state s at step t given that observation z^t was perceived and action a^t was performed in the belief state $\boldsymbol{b}^{t-1}$. This probability is computed using Bayes' rule:

$$\boldsymbol{b}^t(s) = \frac{\mathcal{O}(z^t,a^t,s)\sum_{s'\in\mathcal{S}}\mathcal{T}(s',a^t,s)\boldsymbol{b}^{t-1}(s')}{\sum_{s''\in\mathcal{S}}\mathcal{O}(z^t,a^t,s'')\sum_{s'\in\mathcal{S}}\mathcal{T}(s',a^t,s'')\boldsymbol{b}^{t-1}(s')} \tag{1}$$

Using a matrix representation, Equation (1) can be rewritten:

$$\boldsymbol{b}^k(s) = \frac{D_kT_{a^k}\cdots D_1T_{a^1}\boldsymbol{b}^0}{\mathbb{1}^\top D_kT_{a^k}\cdots D_1T_{a^1}\boldsymbol{b}^0} \tag{2}$$

Where $\boldsymbol{b}^0$ is the initial belief, T_{a^t} are transition matrices according to action a^t, D_i are diagonal matrices with the terms on the diagonal corresponding to the probability to observe z_i given each state, and $\mathbb{1}$ a $|\mathcal{S}|$-dimensional vector of ones.

In order to show the convergence of the belief state to a single state with high probability, let us first state that this probability depends on the number n of succeeded observations among k steps in a non-unobservable context. Nevertheless, non-unobservability is not a sufficient condition to ensure this convergence. Let us now study how n varies regarding to the proposed variants on the observability.

4.1 Enough-Observable models

Enough-observable models ensure that there is only one most likely observation (MLO) in each state and that each state's MLO is not the MLO of any other state:

DEFINITION 3. *An enough-observable* QDET-POMDP *is a* QDET-POMDP *where following assumptions holds:*

$$\exists o_1\in\Omega,\ \forall a\in\mathcal{A},\ \forall s\in\mathcal{S}^{o_1},$$
$$\textit{with } \mathcal{S}^{o_1} = \{s\in\mathcal{S}, o_1\in\Omega | P(o_1|s,a) > P(o|s,a), \forall o\neq o_1\},$$
$$\textit{then } |\Omega| = |\mathcal{S}| \textit{ and } |\mathcal{S}^{o_1}| = 1$$

Here, $\mathcal{S}^{o_1}$ is the set of states where o_1 is the MLO.

Considering this definition, one can state our first main result:

THEOREM 1. *Under the enough-observability assumption,* $\boldsymbol{b}^k(s) \geqslant 1-\varepsilon$ *iff*

$$n \geqslant \frac{1}{2\ln\frac{\nu\theta}{(1-\theta)}}\ln\left[\frac{1-\varepsilon}{\varepsilon}\left(1+\nu^{1-\frac{k}{2}}\right)\right] + \frac{k}{2} \tag{3}$$

Where $\nu = \max_{s,a}\sum_{z\in\Omega} I(\theta > \mathcal{O}(z,a,s) > 0) < |\Omega|$ *the maximum number of "bad" observations that can be perceived in a state.*

PROOF SKETCH. In the worst case, the probability of observing the real underlying state is always minimal and equals to θ at each step. Moreover, if the failed observations obtained always support the second most likely state, it results in an increasing of the probability to potentially be in this state. According to Equation (2) and using determinism of transitions, which induces that transition matrices are permutation matrices, one must show that:

$$\frac{\theta^n\frac{(1-\theta)^p}{\nu^p}}{\theta^n\frac{(1-\theta)^p}{\nu^p} + \theta^p\frac{(1-\theta)^n}{\nu^n} + (\nu-1)\frac{(1-\theta)^k}{\nu^k}} \geqslant 1-\varepsilon \tag{4}$$

Where n is the number of successful observations of the real underlying state and $p = k - n$ the number of failures. The numerator is obtained by obtaining n times a "good" observation and p times a "bad" one during the execution. The denominator sum over all states the same sequence of observation where the first term if for the most likely state, the second term for the second most likely state and the third term for the rest of possible states according to the number of "bad" observations ν. We assume here that the probability to get a "bad" observation is uniform. This assumption is justified by the *maximum-entropy principle* which states that according to the current knowledge, the highest entropy distribution – the uniform in our case – is the most appropriate one. Solving this inequality leads to Equation (3). The extensive derivation of the equations is given in Appendix A. □

Roughly speaking, ν represents also the way the error spreads over the false states and Theorem 1 states that if the observation is good enough (with a large probability θ to observe the real underlying state) and if the error spreads over many states (with a large ν), then it suffices to have one half of the observations plus one to be the real underlying state to converge to a deterministic belief state.

The uniform assumption on bad observations is justified by the maximum entropy principle but can be omitted without loss of generality. Indeed, taking ν equals to 1 induces that the agent observes either the true underlying state or the second most likely state of his belief which is truly the worst case for our agent. The resulting equation would be exactly the same as equation (3) but where the term $\nu^{1-\frac{k}{2}}$ would be also equals to one which would then slightly loosen the bound.

4.2 Factored models

In a more general way than enough-observable models, factored-observable models ensure that each value of each variable is sufficiently often observed so that the factored state can de determined in a finite number of steps:

DEFINITION 4. *A factored-observable* QDET-POMDP *is a* QDET-POMDP *where following assumption holds:*

- *The state space is* factored *in μ state variables: $\mathcal{S} = \bigtimes_{\mathrm{x}\in\mathcal{X}} \mathcal{D}_{\mathrm{x}}$ and observations possible are $\Omega = \bigcup_{\mathrm{x}\in\mathcal{X}} \mathcal{D}_{\mathrm{x}}$.*
- *The sum of probabilities of observing one state's variables' real values is lower bounded by $\theta > \frac{1}{2}$.*

This definition implies that, in the worst case, for each state variable, there is a probability $\frac{\theta}{\mu}$ to observe its real value and a probability $\frac{1-\theta}{|\Omega|-\mu}$ to observe anything else. Note also that this definition is a generalization of Definition 3 which is the case $\mu = 1$. This statement leads to the following theorem:

THEOREM 2. *Under the factored-observability assumption, $\boldsymbol{b}^k(s) \geqslant 1-\varepsilon$ iff*

$$n \geqslant \frac{1}{2\ln\frac{(|\Omega|-\mu)\theta}{\mu(1-\theta)}} \ln\left[\frac{1-\varepsilon}{\varepsilon}\left(1+|\mathcal{S}|-\mu\right)\right] + \frac{k}{2} \qquad (5)$$

PROOF SKETCH. The proof follows exactly the same arguments as in Theorem 1. □

Once the number n of most likely observations is lower bounded, finding the probability to achieve at least this number is simply an application of the binomial distribution to have at least n successes on k trials:

COROLLARY 3. *In any* QDET-POMDP *under enough-observability* ***or*** *factored-observability assumptions, the probability that a belief state $\boldsymbol{b}^k(s)$ is ε-deterministic after k steps is:*

$$\exists s, \Pr(\boldsymbol{b}^k(s) \geqslant 1-\varepsilon) = \sum_{i=n}^{k} \binom{k}{i} \theta^i (1-\theta)^{k-i} \qquad (6)$$

In other words, this indicates that to be certain (with a small δ) to have a deterministic belief state (with a small ε) we may have to explore a large horizon if θ is too small (e.g. near 0.5).

Let us now derive the worst case complexity from these bounds.

5. COMPLEXITY ANALYSIS

In this section, new complexity results induced from Theorems 1 and 2 are given.

5.1 Mono-agent case

A major implication of Theorems 1 and 2 is the reduction of the complexity of general POMDPs problems when a QDET-POMDP is encountered. Indeed, [15] have shown that finite-horizon POMDPs are PSPACE-complete. This roughly speaking rests on the fact that the agent has to choose an action that, given any observation, leads to the choice of another action and so on, on a polynomially bounded horizon T. However, fixing the horizon T to be constant, causes to complexity to fall down many steps in the polynomial hierarchy [21]. Polynomial hierarchy consists in the generalized class of problems that uses oracles. Stockmeyer [21] defined $\Sigma_2^{\mathrm{P}} = \mathrm{NP}^{\mathrm{NP}}$ as the class of decision problems that can be solved in polynomial time by a non-deterministic Turing machine using a NP-oracle. The "canonical" problem for this complexity class (which is SAT for NP) is 2-QBF for Σ_2^{P}; 2-QBF is the problem of deciding whether the following quantified boolean formula is true: $\exists\vec{a}\forall\vec{b}\,\phi(\vec{a},\vec{b})$. Stepping up in the polynomial hierarchy (e.g. Σ_3^{P}) means adding another quantifier for another set of variables of the Boolean formula (e.g. verifying if $\exists\vec{a}\forall\vec{b}\exists\vec{c}\,\phi(\vec{a},\vec{b},\vec{c})$ is true); and so on. Thus, in the case of constant horizon POMDP, one can state:

PROPOSITION 4. *Finding a policy for a finite-horizon-k* POMDP, *that leads to an expected reward at least C is $\Sigma_{2k-1}^{\mathrm{P}}$.*

PROOF. To show that the problem is in $\Sigma_{2k-1}^{\mathrm{P}}$, the following algorithm using a $\Sigma_{2k-2}^{\mathrm{P}}$ oracle can be used: guess a policy for $k-1$ steps with the oracle and then verify that this policy leads to an expected reward at least C in polynomial time by verifying the $|\Omega|^k$ possible histories, since k is a constant. □

As QDET-POMDPs are a subclass of POMDPs and since fixing $1-\delta$, the wanted probability to be in a ε-deterministic belief state, induces a constant horizon under enough-observability or factored-observability assumptions:

COROLLARY 5. *Finding a policy for an infinite horizon* QDET-POMDP, *under enough-observability or factored-observability assumptions, that leads to an expected reward at least C with probability $1-\delta$, is $\Sigma_{2k-1}^{\mathrm{P}}$.*

PROOF. To show that this problem is in $\Sigma_{2k-1}^{\mathrm{P}}$, the following algorithm gives the ε-optimal policy in polynomial time assuming a $\Sigma_{2k-2}^{\mathrm{P}}$ oracle: guess a policy for $k-1$ steps with the oracle and then verify that this policy leads to an expected reward at least C by computing the belief in each leaf of the tree of observations and by adding the optimal expected value of the underlying MDP since the belief is deterministic in each of these leaves. □

Practically, finding a probably approximatively correct ε-optimal policy for a QDET-POMDP thus implies using a k-QMDP algorithm that computes exactly k exact backups of a POMDP and that then uses the policy of the underlying MDP for the remaining steps (eventually infinite).

To sum up, by fixing the wanted probability (1-δ) to be in a ε-deterministic belief state, one can upper-bound the horizon on which it is necessary to plan, from which one can ensure that following the optimal policy of the underlying POMDP will perform well. Now, let us see how can this result can be extended to decentralized decision making.

5.2 Multi-agent case

Concerning the DEC-POMDPs, the improvement is much greater. Indeed, DEC-POMDPs are known to be exceptionally hard to solve optimally in the finite horizon case (NEXP-complete [3]) and even to approximate [18]. Moreover, these approximate solutions of

θ	ν	k	$n \geqslant$
0.6	3	75	40
0.6	10	59	31
0.6	100	50	26
0.7	3	22	13
0.7	10	19	11
0.7	100	14	8
0.8	3	9	6
0.8	10	6	4
0.8	100	6	4

Table 1: Enough-Observable bound.

θ	μ	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{S}\rvert$	k	$n \geqslant$
0.6	2	10	100	84	44
0.6	3	5	125	98	52
0.6	10	6	10^6	112	60
0.7	2	10	100	30	17
0.7	3	5	125	33	19
0.7	10	6	10^6	39	23
0.8	2	10	100	13	8
0.8	3	5	125	16	10
0.8	10	6	10^6	20	13

Table 2: Factored-observable bound.

the infinite horizon general case, usually based on finite state controllers [2], converge to local optima without any guarantee on the solution found.

By restricting the model to be quasi-deterministic, and assuming that all agents still have enough-observability or factored-observability, one can also find a great complexity reduction:

COROLLARY 6. *Finding a policy for an infinite horizon* QDET-DEC-POMDP*, under enough-observability or factored-observability assumptions, that leads to an expected reward at least C with probability $1-\delta$, is* PSPACE.

PROOF. To show that this problem is PSPACE, the following algorithm gives the ε-optimal policy in polynomial space: expand all possible action-observation history up to horizon k (done in space $\mathcal{O}((|\boldsymbol{\mathcal{A}}||\boldsymbol{\Omega}|)^k)$), and then compute for each leaf of the constructed tree the expected value of the reached quasi-deterministic belief using the underlying MMDP infinite horizon optimal policy. Finally, propagate the value back to the root to verify if an expected reward of C is obtained. □

Note that the assumption of enough-observability means here that each agent perceives the same *complete* state while not necessarily obtaining the same observation. On one hand, this assumption seems less applicable in DEC-POMDPs than in POMDPs since many internal values of the agents are also in the joint state of the DEC-POMDP and thus are not necessarily observable as the example in the next Section will illustrate. On the other hand, assuming a quasi-reliable communication system between agents is not so restrictive and induces naturally the enough-observability assumption by encompassing the communication noise into the observation noise.

In fact, the enough-observability assumption relates closely to the work of Goldman *et al.* [8] on DEC-MDPs where agents, when communicating their observations, have access to the real underlying state. This nonetheless differs on one point; while in DEC-MDPs each agent observes completely its own part of the state, granting the set of agents the complete observability of the state through communication, enough-observability assumption only assure that each agent *may* observe the true underlying state with probability at least θ. In other words, a DEC-MDP with complete communication is a DEC-POMDP under the enough-observability assumption with $\theta = 1$.

Let us now put figures on Theorems 3 and 4 and present a new decentralized fire fighting problem that meets our assumption requirements.

6. EXPERIMENTAL ANALYSIS

In this section we first provide examples of horizon that can be induced by the proposed bounds and then provide an example where such bounds could be applied.

6.1 Bounds' efficiency

To give an idea of the efficiency of the proposed bounds, we define $\delta > 0$ such that $\Pr(\boldsymbol{b}^K(s) \geqslant 1-\varepsilon) \geqslant 1-\delta$. Table 1 and 2 give, for $\varepsilon = 10^{-3}$, $\delta = 10^{-1}$ and different values of θ, the probability of observation, ν, the error spreading factor, μ, the number of state variables, and the domains' size of variables, the value of the bound on the horizon k and the number of successes needed n given that the probability of having both is above $1-\delta$.

As expected, horizons needed to converge are greater in the factored case than in the enough-observable case for similar state and observation spaces since the agent, at each time step, gets less information about the current state. Actually, observations discriminate among subsets of states but not among states themselves like in the previous case. However, as the number of observations is much less than in the previous case, current algorithms may have less difficulty in this type of problems. An empirical study of their difference should be interesting as a research avenue[1].

Furthermore, it is interesting to notice that, as soon as the probability θ to observe the real underlying state is above 0.8 in the enough-observable case, it suffices to compute the optimal policy for the first four steps to have a ε-deterministic belief state with at least 90% probability.

6.2 Infinite Horizon QDET-DEC-POMDP

As suggested by the complexity proofs of corollary 5 and 6, an algorithm that would solve ε-optimally the infinite horizon QDET-DEC-POMDP problem under enough observability or factored observability consists in computing a regular policy for the finite k-horizon problem and then use the optimal policy of the underlying MMDP for the remaining of the task. Such an algorithm would use for example the Dynamic Programming method [10]. However, in practice, solving a k-horizon DEC-POMDP even quasi-deterministic is still very hard and some other approximation should be used. We thus present in this section some results of a specific fire fighting problem where agents have to coordinate to make a bucket chain from a fire hydrant to a fire (e.g. Figure 4).

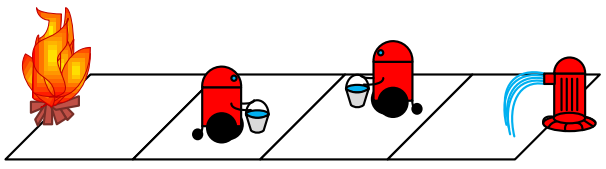

Figure 4: The bucket chain problem.

[1]Comparing results for the fire fighting problem between factored and enough observable models will be available in the final version of the paper.

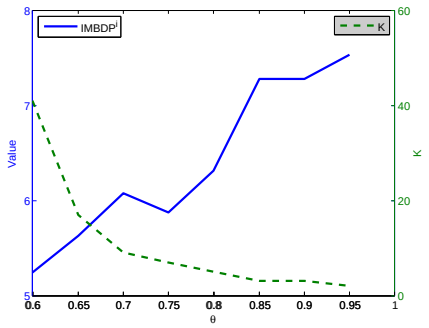

Figure 1: Expected value and estimated horizon for different value of θ.

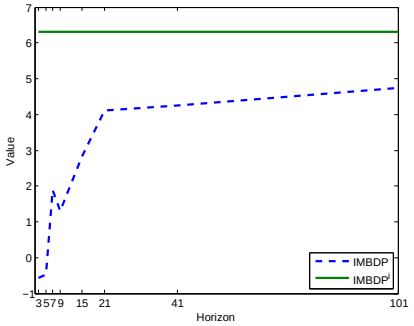

Figure 2: Finite horizon discounted expected value for various horizon lengths.

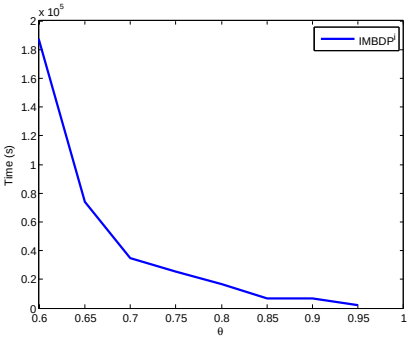

Figure 3: Computational time for different value of θ.

The general version of this problem is stated as follows: two agents are located on a linear grid and can carry a bucket. They can go *right*, *left*, or *throw water*, each action incurring a small penalty (-0.1 per agent). As soon as they go on the rightmost place, they automatically refill their bucket with probability φ. Moving actions have also a probability φ to succeed, the action *throw water* is always successful. Agents can give a bucket one to each other when on an adjacent place through the same *throw water* action. Each time a bucket is emptied on the most left place, a reward is obtained (1 per agent). Agents are initially placed at random without any water. Agents are assumed to have a noisy observation on their position (and have a θ probability to observe it and $\frac{1-\theta}{2}$ to observe one of the adjacent state) as well as they receive a noisy communication of the other agent indicating its position and its bucket state. As the problem could be an infinite horizon problem a discount factor $\gamma = 0.95$ has been used in the experiments. A value $\varphi = 1$ was used to meet the deterministic transition requirements. When $\varphi = 1$, the number of reachable states is 49 (7 for each agent) and thus 49 observations can be obtained.

Concerning the algorithm, we used an adapted version of Improved Memory Bounded Dynamic Programming (IMBDP) [19] so that it used the optimal expected value of the underlying MMDP at the leaves of the tree search. This algorithm is denoted by IMBDP^i on figures.

We ran several simulations using various parameters. Figure 1 shows the expected of IMBDP^i on the bucket chain problem for various values of θ ranging from 0.6 to 0.95. Parameters of IMBDP^i were $maxTree = 5$ and $maxObs = 1$. Increasing these parameters does not significantly improve the expected value while significantly deteriorating the time and space performances. As expected, as long as θ increases to one, the necessary planning horizon decreases to two and the infinite horizon expected value increases near to the value of the underlying MMDP's optimal policy.

Figure 2 shows the finite horizon discounted expected values for $\theta = 0.8$ at various horizons ranging from 3 to 101 showing that the algorithm tends to the infinite horizon expected value. The standard IMBDP algorithm was used with the same parameters as above ($maxTree = 5$ and $maxObs = 1$).

Finally, Figure 3 shows the computational time needed to compute an approximate infinite horizon policy using the exact same parameters for the algorithm. As expected, the time decreases as θ grows since the needed horizon also decreases.

Let us now discuss the different assumptions of the proposed models and their significance.

7. DISCUSSION

Quasi-deterministic models encompass numerous decision problems where the environment is well defined and controlled but just partially observed. The presented results can then be applied in a large number of applications ranging from web agents to fault detection systems. However, the approach still has some limitations.

First of all, presented results are probabilistic. Agents are guaranteed to converge to an ε-deterministic belief state with probability $1 - \delta$. Diminishing the probability δ induces an exponential increase of the planning horizon k and thus an increase of the computational needs.

Second, deterministic transition functions are not the mostly used transition functions in the Markovian community since its probabilistic roots. However, many real problems are in fact deterministic but do not use Markovian models for the same reason.

Third, We did not talk about policies nor actions throughout the paper, this relies on the fact that we assume that observation perceived are the same whatever the action is chosen. Relaxing this assumption leads to the problem of balancing information gathering and reward gathering like the *exploration-exploitation* dilemma in reinforcement learning. A concern that is beyond the scope of this paper but that we intend to prospect in future researches.

Last, it is not clear how the assumptions on the observability of the agents restrain the field of applications of this work although it is clear that it may be applied in many indoor robotics applications where each sensor is very often dedicated to one component of the state of the robot, independently of the policy chosen.

Concerning the observability assumptions, they are the key point of this paper. The majority of problems in the literature assume either full observability or partial observability. This paper proposes two other classes of observability that fill in the gap between these two extremes. Indeed, many of real problems are enough-observable or even factored-observable and one usually has to assume partial observability in these cases. This proposition of a new subclass of partial observability may stimulate the development of specific algorithms for these subclasses based on and/or graphs for example [4].

Moreover, in the multiagent case, the proposed work assumed some sort of communications between agents, allowing them to exchange their partial view of the world in order to provide each agent a noisy but global view of the state of the system. Even if this kind of communication is very often employed in true applications of multiagent systems, all previous works on communication in decentralized POMDPs (e.g. [7]) focused on the act of communicating and not on the goals of doing so nor on the content of the

exchanged messages. Indeed, a natural insight of communicating is to inform others what they do not know. By communicating its own observations to all other agents, the problem of decentralized POMDP would reduce to a "simple" multiagent POMDP where every single agent shares the same belief state and thus does not have to plan for all over possible policies of other agents. This work thus opens many interesting communication modeling avenues in multiagent Markovian models where multiagent does not necessarily means NEXP-complete.

8. CONCLUSION AND FUTURE WORK

To summarize, we proposed in this paper an extension of the DET-POMDP framework to stochastic observability, called QDET-POMDP, that bridges a part of the gap between DET-POMDPs and general POMDPs. A further extension to multiagent systems has also been proposed. A study of their convergence properties leads to a significant improvement in terms of computational complexity. Empirical performances are also presented through a new decentralized problem of fire fighting with communication between agents.

Concerning future work, we are currently working on extending current bounds on the horizon to problems where transitions are stochastic but can still be lower bounded. Model and algorithms that take communication of agents into account are also one of our research avenues. Finally, in our opinion, the problem of balancing information-reward gathering should be one of the next research areas in the community studying partially observable environments and models.

9. REFERENCES

[1] E. Amir and A. Chang. Learning Partially Observable Deterministic Action Models. *J. Artif. Intell. Res.*, 33:349–402, 2008.

[2] D. S. Bernstein, C. Amato, E. A. Hansen, and S. Zilberstein. Policy Iteration for Decentralized Control of Markov Decision Processes. *J. of AI Res. (JAIR)*, 34:89–132, 2009.

[3] D. S. Bernstein, S. Zilberstein, and N. Immerman. The Complexity of Decentralized Control of Markov Decision Processes. In *Proc. of UAI*, pages 32–37, 2000.

[4] B. Bonnet. Deterministic POMDPs Revisited. In *Proc. of Uncertainty in AI*, 2009.

[5] C. Boutilier. Sequential Optimality and Coordination in Multiagent Systems. In *Proc. of the Inter. Joint Conf. on AI*, pages 478–485, 1999.

[6] A. Cassandra, L. Kaelbling, and M. Littman. Acting Optimally in Partially Observable Stochastic Domains. In *Proc. of Assoc. for the Adv. of AI*, pages 1023–1028, 1994.

[7] C. V. Goldman, M. Allen, and S. Zilberstein. Decentralized Language Learning through Acting. In *Proc. of the Inter. Joint Conf. on AAMAS*, pages 1006–1013, 2004.

[8] C. V. Goldman and S. Zilberstein. Decentralized Control of Cooperative Systems: Categorization and Complexity Analysis. *J. Artif. Intell. Res.*, 22:143–174, 2004.

[9] R. P. Goldman and M. S. Boddy. Expressive planning and explicit knowledge. In *Proc. of the 3rd Inter. Conf. on Artif. Intel. Planning Systems*, pages 110–117, 1996.

[10] E. A. Hansen, D. S. Bernstein, and S. Zilberstein. Dynamic Programming for Partially Observable Stochastic Games. In *Proc. of Assoc. for the Adv. of AI*, pages 709–715, 2004.

[11] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and Acting in Partially Observable Stochastic Domains. *Artif. Intell.*, 101(1-2):99–134, 1998.

[12] M. Littman. *Algorithms for Sequential Decision Making*. PhD thesis, Dept. of Comp. Sc., Brown University, 1996.

[13] D. Nau, M. Ghallab, and P. Traverso. *Automated Planning: Theory & Practice*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.

[14] H. Palacios and H. Geffner. From Conformant into Classical Planning: Efficient Translations that May Be Complete Too. In *Proc. of the 17th Int. Conf. on Automated Planning and Scheduling*, pages 264–271, 2007.

[15] C. Papadimitriou and J. Tsisiklis. The Complexity of Markov Decision Processes. *Math. Oper. Res.*, 12(3):441–450, 1987.

[16] K. Pattipati and M. Alexandridis. Application of Heuristic Search and Information Theory to Sequential fault Diagnosis. *IEEE Trans. on Sys., Man and Cyber.*, 20(4):872–887, 1990.

[17] D. V. Pynadath and M. Tambe. The Communicative Multiagent Team Decision Problem: Analyzing Teamwork Theories and Models. *J. Artif. Intell. Res.*, 16:389–423, 2002.

[18] Z. Rabinovich, C. V. Goldman, and J. S. Rosenschein. The Complexity of Multiagent Systems: the Price of Silence. In *Proc. of AAMAS*, pages 1102–1103, 2003.

[19] S. Seuken and S. Zilberstein. Improved Memory-Bounded Dynamic Programming for Dec-POMDPs. In *Proc. of Uncertainty in AI*, 2007.

[20] E. J. Sondik. *The optimal control of Partially Observable Markov Processes*. PhD thesis, Stanford University, 1971.

[21] L. J. Stockmeyer. The Polynomial-Time Hierarchy. *Theor. Comput. Sci.*, 3(1):1–22, 1976.

10. APPENDIX A

PROOF OF THEOREM 1.

$$
\begin{aligned}
& \frac{\theta^n \frac{(1-\theta)^p}{\nu^p}}{\theta^n \frac{(1-\theta)^p}{\nu^p} + \theta^p \frac{(1-\theta)^n}{\nu^n} + (\nu-1)\frac{(1-\theta)^k}{\nu^k}} \geqslant 1-\varepsilon \\
\Leftrightarrow \quad & \frac{\nu^n\theta^n(1-\theta)^p}{\nu^n\theta^n(1-\theta)^p + \nu^p\theta^p(1-\theta)^n + (\nu-1)(1-\theta)^k} \geqslant 1-\varepsilon \\
\Leftrightarrow \quad & \frac{\nu^p\theta^p(1-\theta)^n}{\nu^n\theta^n(1-\theta)^p} + \frac{(\nu-1)(1-\theta)^k}{\nu^n\theta^n(1-\theta)^p} \leqslant \frac{1}{1-\varepsilon} - 1 \\
\Leftrightarrow \quad & \nu^{k-2n}\theta^{k-2n}(1-\theta)^{2n-k} + (\nu-1)\frac{\theta^{-n}\nu^{-n}}{(1-\theta)^{-n}} \leqslant \frac{\varepsilon}{1-\varepsilon} \\
\Leftrightarrow \quad & \frac{\nu^{k-2n}\theta^{k-2n}}{(1-\theta)^{k-2n}}\left[1 + (\nu-1)\frac{\nu^{-p}\theta^{-p}}{(1-\theta)^{-p}}\right] \leqslant \frac{\varepsilon}{1-\varepsilon} \\
\Leftrightarrow \quad & (k-2n)\ln\frac{\nu\theta}{(1-\theta)} + \ln\left[1 + (\nu-1)\frac{\nu^{-p}\theta^{-p}}{(1-\theta)^{-p}}\right] \leqslant \ln\frac{\varepsilon}{1-\varepsilon} \\
\Leftrightarrow \quad & (k-2n)\ln\frac{\nu\theta}{(1-\theta)} \leqslant \ln\frac{\varepsilon}{1-\varepsilon} - \ln\left[1 + (\nu-1)\frac{(1-\theta)^p}{\nu^p\theta^p}\right] \\
\Leftrightarrow \quad & (2n-k)\ln\frac{\nu\theta}{(1-\theta)} \geqslant \ln\frac{1-\varepsilon}{\varepsilon} + \ln\left[1 + (\nu-1)\frac{(1-\theta)^p}{\nu^p\theta^p}\right] \\
\Leftrightarrow \quad & (2n-k) \geqslant \frac{\ln\frac{1-\varepsilon}{\varepsilon}}{\ln\frac{\nu\theta}{(1-\theta)}} + \frac{1}{\ln\frac{\nu\theta}{(1-\theta)}}\ln\left[1 + (\nu-1)\frac{(1-\theta)^p}{\nu^p\theta^p}\right] \\
\Leftrightarrow \quad & n \geqslant \frac{\ln\frac{1-\varepsilon}{\varepsilon}}{2\ln\frac{\nu\theta}{(1-\theta)}} + \frac{1}{2\ln\frac{\nu\theta}{(1-\theta)}}\ln\left[1 + (\nu-1)\frac{(1-\theta)^p}{\nu^p\theta^p}\right] + \frac{k}{2} \qquad (7)
\end{aligned}
$$

but since $2 \leqslant \nu \leqslant |\mathcal{S}| - 1$, $n > \frac{k}{2}$ and $\frac{1-\theta}{\theta} < 1$,

$$
\begin{aligned}
\ln\left[1 + \frac{|\mathcal{S}|-2}{\nu^{k-n}}\frac{(1-\theta)^{k-n}}{\theta^{k-n}}\right] & \leqslant \ln\left[1 + \frac{|\mathcal{S}|-2}{\nu^{\frac{k}{2}}}\left(\frac{1-\theta}{\theta}\right)^{\frac{k}{2}}\right] \\
& \leqslant \ln\left[1 + \nu^{1-\frac{k}{2}}\right]
\end{aligned}
$$

From which, ensuring Eqn. (3) also ensures Eqn. (7), □