Quasi Deterministic POMDPs and DecPOMDPs

(Extended Abstract)

Camille Besse & Brahim Chaib-draa
DAMAS Laboratory
Department of Computer Science and Software Engineering
Laval University, G1V 0A6, 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, it still captures 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.

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

Markov Models Design

Keywords

Enough Observability, Coordination, Complexity

1. INTRODUCTIVE EXAMPLES

In recent years, many problems have been modeled as POMDPs [2] and DET-POMDPs [3] and had been used for developing and evaluating various algorithms for planning under uncertainty and partial information. For space reasons, we present only two examples of 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} represents 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 [5].
- **Robot Navigation:** Consider an indoor robot in a $m \times n$ grid that must navigate from an initial position to a goal position while avoiding obstacles using only some noisy sensors on its position. The robot's moves are fairly deterministic but the observation of its current state is distorted by the noise on the sensors. The goal is to find a strategy for guiding the robot to its destination. This problem is easily extendable to several robots.

Let us now see the formal definition of the deterministic POMDP and the proposed variants.

Cite as: Quasi Deterministic POMDPs and Dec-POMDPs (Extended Abstract), C. Besse and B. Chaib-draa, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. XXX-XXX.

2. MODEL AND VARIANTS

2.1 Formal Models

Deterministic POMDPs were initially defined as follows [3]:

DEFINITION 1. *A Deterministic* POMDP *(*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 $a \in \mathcal{A}$ *of the agent;*
- Ω is the finite set of observations $z \in \Omega$ *of the agent and;-* $\mathcal{O}(z, a, s') : \Omega \times \mathcal{A} \times \mathcal{S} \mapsto \{0, 1\}$ *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 which state s' results from making a in 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 and $\boldsymbol{b}^0$ is the a priori knowledge about the state, i.e. the initial belief state, assumed non-deterministic.

Compared to DET-POMDPs, our proposed extended model presents changes on the observability function and is defined as follows:

DEFINITION 2. *A Quasi-deterministic* POMDP *(*QDET-POMDP*) is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{O}, \mathcal{R}, \boldsymbol{b}^0 \rangle$, where:*
- $\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \boldsymbol{b}^0$ 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 world is minimally observable and the probability of getting one of the observations is lower bounded in each state by one half;

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), only the study of the QDET-POMDP is necessary from which we will extend to the multiagent case. Indeed, one can consider that there exists in a QDET-DEC-POMDP a most likely observation of the state whatever the chosen joint action is, like in the monoagent case.

2.2 Enough-Observable models

Furthermore, in order to ensure completely the convergence of the agent's state knowledge, we propose to ensure the observability of this state through the observation function.

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:*

$\exists o_1 \in \Omega, \forall a \in \mathcal{A}, \forall s \in \mathcal{S}^{o_1},$

with $\mathcal{S}^{o_1} = \{s \in \mathcal{S}, o_1 \in \Omega | P(o_1|s,a) > P(o|s,a), \forall o \neq o_1\},$

then $|\Omega| = |\mathcal{S}|$ *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*

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

Where $\nu = \max_{s,a} \sum_{z\in\Omega} I(\theta > \mathcal{O}(z,a,s) > 0) < |\Omega|$.

Where n is the number of successful observations of the real underlying state, k the number of steps and ν represents the way the error spreads over the states.

Theorem 1 thus states that if the observation is good enough and if the error spreads over many states, 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.

Once the number n of MLOs 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 over k trials:

COROLLARY 2. *In any* QDET-POMDP *under enough-observability assumption, the probability that a belief state* $\boldsymbol{b}^k(s)$ *is* ε*-deterministic after* k *steps is:*

$$\exists s, \Pr(\boldsymbol{b}^k(s) \geqslant 1-\varepsilon) = \sum_{i=n}^{k} \left(\begin{array}{c} k \\ i \end{array}\right) \theta^i (1-\theta)^{k-i} \qquad (2)$$

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.

3. COMPLEXITY ANALYSIS

3.1 Mono-agent case

A major implication of Theorems 1 is the reduction of the complexity of general POMDPs problems when a QDET-POMDP is encountered. Indeed, [4] have shown that finite-horizon POMDPs are PSPACE-complete. However, fixing the horizon T to be constant, causes to complexity to fall down many steps in the polynomial hierarchy [7]. In the case of constant horizon POMDP, one can state:

PROPOSITION 3. *Finding a policy for a finite-horizon-k* POMDP*, that leads to an expected reward at least* C *is* $\Sigma^{\mathrm{P}}_{2k-1}$.

PROOF. To show that the problem is in $\Sigma^{\mathrm{P}}_{2k-1}$, the following algorithm using a $\Sigma^{\mathrm{P}}_{2k-2}$ 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 assumption:

COROLLARY 4. *Finding a policy for an infinite horizon* QDET-POMDP*, under enough-observability assumption, that leads to an expected reward at least* C *with probability* $1-\delta$*, is* $\Sigma^{\mathrm{P}}_{2k-1}$.

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 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.

3.2 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 [1]) and even to approximate it [6].

By restricting the model to be quasi-deterministic, and assuming that all agents still have enough-observability, one can state:

COROLLARY 5. *Finding a policy for an infinite horizon* QDET-DEC-POMDP*, under enough-observability assumption, leading to an expected reward at least* C *with probability* $1-\delta$*, is* PSPACE.

Note that the assumption of enough-observability 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. However, assuming a quasi-reliable communication system between agents is not so restrictive and induces naturally the enough-observability assumption.

4. CONCLUSION

We presented a new subclass of of POMDPs called enough-observable QDET-POMDPs that encompasses numerous decision problems where the environment is well defined and controlled but just partially observed. A study of their convergence properties leads to a significant improvement in terms of computational complexity.

5. REFERENCES

[1] 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.

[2] 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.

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

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

[5] 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.

[6] 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.

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