Purely Epistemic Markov Decision Processes

Régis Sabbadin
INRA-UBIA
31326 Castanet Tolosan Cedex, France
sabbadin@toulouse.inra.fr

Jérôme Lang
IRIT - Université Paul Sabatier
31062 Toulouse Cedex, France
lang@irit.fr

Nasolo Ravoanjanahary
INRA-UBIA
31326 Castanet Tolosan Cedex, France
ravoanja@irit.fr

Abstract

Planning under uncertainty involves two distinct sources of uncertainty: uncertainty about the effects of actions and uncertainty about the current state of the world. The most widely developed model that deals with both sources of uncertainty is that of Partially Observable Markov Decision Processes (POMDPs). Simplifying POMDPs by getting rid of the second source of uncertainty leads to the well-known framework of fully observable MDPs. Getting rid of the *first* source of uncertainty leads to a less widely studied framework, namely, decision processes where actions cannot change the state of the world and are only intended to bring some information about the (static) state of the world. Such "purely epistemic" processes are very relevant, since many practical problems (such as diagnosis, database querying, or preference elicitation) fall into this class. However, it is not known whether this specific restriction of POMDP is computationally simpler than POMDPs. In this paper we establish several complexity results for purely epistemic MDPs (EMDPs). We first show that short-horizon policy existence in EMDPs is PSPACE-complete. Then we focus on the specific case of EMDPs with reliable observations and show that in this case, policy existence is "only" NP-complete; however, we show that this problem cannot be approximated with a bounded performance ratio by a polynomial-time algorithm.

Introduction

Most real-world planning problems are pervaded with uncertainty. This uncertainty comes from two distinct sources: uncertainty about the effects of actions and uncertainty about the current state of the world (or partial observability). The first source of uncertainty is linked to the dynamics of the system, whereas the second one is purely static. The most widely developed model that deals with both sources of uncertainty is that of Partially Observable Markov Decision Processes (POMDPs) (Smallwood & Sondik 1973). Solving POMDP is a very hard task: policy existence (that is, the problem of determining whether there exists a policy for a given problem whose expected utility exceeds a given threshold) in short-horizon (resp. long-horizon) POMDPs is PSPACE-complete (resp. EXPSPACE) (Papadimitriou & Tsitsiklis 1987; Mundhenk *et al.* 2000). When all actions are deterministic, then the complexity of policy existence falls down and is only NP-complete (Littman 1996).

Simplifying POMDPs by getting rid of uncertainty about the current state of the world leads to the well-known framework of *fully observable Markov decision processes*, which are known to be tractable, unlike general POMDPs. Getting rid of uncertainty about the effects of actions on the state of the world leads to a less widely studied framework, namely, *decision processes where actions are only intended to bring some information about the (static) state of the world*. In this paper we focus on such "purely epistemic" Markov decision processes (EMDPs for short).

Remark first that this specific class of POMDPs is very relevant in practice. Indeed, many important problems fall into this class, including *preference elicitation*, with possible application to electronic commerce (see (Boutilier 2002) for a POMDP formulation of preference elicitation); *diagnosis* (in medicine or general systems); *database querying*; *games* (such as Mastermind and naval battleship)...

Clearly, such problems can be formulated as POMDPs and solved using generic algorithms tailored for general POMDPs. But doing this would not allow for benefiting from the possible computational benefits obtained by getting rid of the uncertainty about the effects of actions. This raises an important question: to what extent is this specific class of POMDPs computationally easier than POMDPs in general? This paper considers this issue in detail and gives some first answers by establishing several complexity results for EMDPs. We first show that short-horizon policy existence in EMDPs is PSPACE-complete – which means that as far as this problem is concerned, EMDPs are as complex as general POMDPs. Then we focus on a specific case of EMDPs where the observations returned by information-gathering actions are fully reliable. In this case, policy existence is NP-complete, thus no simpler than policy existence in deterministic POMDP, and finding an optimal policy for such D-EMDP is an NPO-complete optimization problem[1]. We show that this problem is not in APX, which means that there exists no polynomial-time algorithm returning approximate solutions with bounded performance ratio.

[1] Recall that NPO is the counterpart of the decision class NP for optimization problems (Ausiello *et al.* 1999).

Epistemic Markov Decision Processes

Definition

Epistemic Markov Decision Processes (EMDPs) are intended to provide us with a framework for epistemic planning. In EMDPs, current knowledge about the state of the world is modelled by a *belief state* b (a probability distribution over the set $\mathcal{S}$ of possible physical states). The agent has a set $\mathcal{A}$ of available actions. Performing one of these actions produces an observation $o \in \mathcal{O}$, where $\mathcal{O}$ is a predefined set of possible outcomes, but does *not* change the (unknown) state of the world. Such actions are called *purely epistemic*. Typical purely epistemic actions are truth tests (such as checking directly whether a given component is faulty or not), value tests (or measurements), queries to agents or databases, etc.

When a purely epistemic action a is performed in state s, its effect is described by an observation function O, which can be:

- *Deterministic*, in which case O is a mapping from $\mathcal{S} \times \mathcal{A}$ to $\mathcal{O}$: observation $o = O(s, a)$ is returned whenever a is performed in s.
- *Stochastic*, in which case o is drawn from a probability distribution $p(\cdot|s, a)$ over $\mathcal{O}$. The probability of observing o when a is applied in s is denoted $O(s, a, o)$.

An example of stochastic epistemic action is a noisy measurement, which may return several possible values for the same true value to be measured.

For technical reasons that will be made clear soon, in addition to these purely epistemic actions we consider one more action, called $stop$. This action is *terminating*: no further action can be applied after $stop$ has been applied. Furthermore, we will require later that every branch of a policy ends with $stop$. Whenever a non-terminating action a is performed in some current belief state b, a cost $c(a) \geq 0$ will incur, which depends on a only (and not on b). When the action $stop$ is performed, a reward (positive or negative) $\mathcal{R}(b)$ incurs. $\mathcal{R}(b)$ measures to which extent the epistemic content of the final belief state b is satisfactory. Finally, we let $\mathcal{R}(b, a) = -c(a)$ if $a \neq stop$ and $\mathcal{R}(b, stop) = \mathcal{R}(b)$. *Note that* $\mathcal{R}(b, a) \leq 0$ *whenever* $a \neq stop$. Here are some specific cases for this function $\mathcal{R}(.)$[2]:

- let $r : \mathcal{S} \to \mathbb{R}^+$, and $\mathcal{R}(b) = r(s)$ if $b(\{s\}) = 1$ and 0 otherwise: a reward is obtained only when the state is *disambiguated* (and this reward may depend on the state).
- let $\{\langle X_1, \alpha_1\rangle, \ldots, \langle X_q, \alpha_q\rangle\}$ where each X_i is a nonempty subset of S and each α_i a positive real number. Then $\mathcal{R}(b) = \max\{\alpha_i | b(X_i) = 1\}$ if there exists a i such that $b(X_i) = 1$ (and $\mathcal{R}(b) = 0$ otherwise).
- $\mathcal{R}(b) = -h(b)$, where $h(b) = \sum_{s\in S} -b(s)log\big(b(s)\big)$.

For computational reasons, we require that $\mathcal{R}$ can be expressed in space polynomial in $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{O}|$ and computed also in polynomial time from this polynomial size input. For instance, in the first case we only need to store $r(s)$ for each $s \in S$, which takes space $O(|\mathcal{S}|)$ and from which $\mathcal{R}(b)$ is computable in time $O(|\mathcal{S}|)$.

Formally, an EMDP is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, O, \mathcal{R}, b_0, H\rangle$:

- $\mathcal{S}, \mathcal{A}$ and $\mathcal{O}$ are finite sets of *states*, *actions* and *observations*, respectively,
- O is the *observation function*: $O(s, a, o) = p(o|s, a)$ is the probability of observing o after taking action a while being in s[3].
- $\mathcal{R}$ is the *reward function*.
- b_0 (initial *belief state*) is a probability distribution over $\mathcal{S}$.
- $H \in \mathbb{N}$ is the *horizon* of the EMDP[4].

It is clear that an EMDP is a special case of finite-horizon POMDP, where the transition function T is defined as follows: for any $a \in \mathcal{A}$, $T(s, a, s') = 1$ if $s = s'$, and $T(s, a, s') = 0$ if $s \neq s'$.

Belief state update in EMDP

In an EMDP, observations modify our current knowledge about the state of the world: only the *belief* changes, not the state of the world itself (which departs from the general POMDP setting). Recall that in a POMDP, when an action $a \in \mathcal{A}$ is performed in belief state $b \in \mathcal{B}$ and an observation $o \in \mathcal{O}$ results, a new belief state $b^o_a \in \mathcal{B}$ is computed as :

$$\forall s' \in \mathcal{S}, b^o_a(s') = \frac{p(o|s', a) \sum_{s\in\mathcal{S}} \big(T(s, a, s')b(s)\big)}{p(o|b, a)}.$$

where $p(o|b, a)$ is a normalizing factor independent of s'.

Taking into account the transition simplification, the EMDP belief state update becomes:

$$b^o_a(s') = \frac{p(o|s', a)b(s')}{p(o|b, a)}, \quad \forall a, o, b \text{ and } \forall s' \in \mathcal{S}, \qquad (1)$$

where the normalizing factor $p(o|a, b)$ is computed as:

$$p(o|b, a) = \sum_{s'\in\mathcal{S}} p(o|s', a)b(s'), \quad \forall a, o, b.$$

[2]The reader may wonder why the obvious choice $\mathcal{R}(b) = \sum_{s\in S} b(s)r(s)$ does not figure on this list. The reason is that rewards should be attached to the *epistemic content* of belief states, and should increase with the amount of information, which totally departs from expected utility: for instance, suppose $S = \{s_1, s_2\}$; then, as soon as $u(s_1) < u(s_2)$, the belief state in which we know for sure that the true state is s_1 would have the minimal reward among all belief states, which violates this principle that more informed belief states are more desirable. This principle of monotonicity of $\mathcal{R}$ with respect to information would deserve more attention, but we leave its study for further work.

[3]Remark that, since the state of the world does not change when performing action a, $O(s, a, o)$ can be interpreted both as the probability of observing o when taking action a while being in s and as the probability of observing o when taking action a and ending up in s.

[4]Throughout this paper, the problems considered have a *finite horizon*. As far as purely epistemic problems are concerned, this does not really make sense to consider infinite-horizon decision problems. Moreover, the most interesting classes of EMDPs we consider next imply that the horizon is not only finite but bounded by the number of available actions. For similar reasons, we do not consider discount factors.

Computing optimal policies for EMDP

In a finite-horizon EMDP (as in a POMDP), a policy δ can be represented by a tree τ_δ of depth H, where each node is labelled with an action, and the edges leaving a node labelled with action a represent all possible observation outcomes after a is performed. Moreover, all terminal nodes, and only terminal nodes, are labelled by *stop*. Equivalently, a policy can be viewed as a partial function δ mapping finite (and possibly empty) sequences of pairs consisting of an action and an observation to an action: $\delta()$ is the first action of the policy, and $\delta(a_1, o_1, \ldots, a_i, o_i)$ is the action to be performed after actions $a_1, \ldots, a_i$ have been performed and have resulted in the observations $o_1, \ldots, o_i$. The utility of a policy δ applied in initial belief state b_0 is the expectation of the sum of rewards incurred along all branches:

$$V^\delta(b_0) = E\Big[\sum_{t=0}^{H} \mathcal{R}(b_t, a_t)|b_0, \delta\Big]$$

Given an initial belief state b_0 and a policy tree τ_δ, we can build iteratively the corresponding tree of reachable belief states. To the root of the tree is attached belief state b_0, and for any node of associated belief b and action $a = \tau_\delta(b)$, the belief attached to the successor node via edge labelled o is b_a^o. The utility of policy δ can then be computed *backwards*, attaching rewards to the (belief states) leaves of the tree, and costs to every edges. An *optimal policy* δ^* for a finite-horizon EMDP (a policy maximizing $V^\delta(b_0)$) can of course also be computed backwards. However, one difficulty with finite-horizon POMDP in general is that the size required for expressing H-steps finite horizon policies grows exponentially with the horizon H considered, as grows the number of belief states that can be reached from the initial belief state. So, the backwards induction algorithm just described is of exponential (time and space) complexity.

Let us now consider the POLICY EXISTENCE problem: given a problem $\mathcal{P}$ and a real number v, is there a policy for $\mathcal{P}$ whose expected value is at least v? Recall that for general POMDPs, this problem is PSPACE-complete for problems with a short-term horizon, that is, when $H \leq |\mathcal{S}|$ (Papadimitriou & Tsitsiklis 1987; Mundhenk *et al.* 2000)[5]. In the remainder of the paper, we show that policy existence for EMDPs is also PSPACE-complete in the short-term case (and, a fortiori, PSPACE-hard in the more general finite-horizon case). This result is rather negative, since it shows that EMDPs are as hard to solve as general POMDP. However, we will show that in the specific case where actions are *deterministic*, i.e. provide us with reliable observations, (a) any EMDP has a short-term horizon, and (b) policy existence problem is NP-complete.

Complexity of policy existence in finite-horizon EMDPs

The results of this section rely on the following complexity result by (Conitzer & Sandholm 2003). Define STATE DISAMBIGUATION (SD) as the following problem:

[5]Equivalently, POLICY EXISTENCE is PSPACE-complete when H is encoded in unary in the input.

Definition 1 (STATE-DISAMBIGUATION) *We are given:*

- *A set* $\Theta = \{\theta_1, \ldots, \theta_n\}$ *of possible states of the world and a uniform probability distribution* p *over* Θ.
- *A utility function* $u : \Theta \to [0; +\infty[$. $u(\theta_i)$ *is the utility of knowing* for sure *that the state of the world is* θ_i.
- *A set* $\mathcal{Q} = \{q_1, \ldots, q_r\}$ *of queries.* $q_j = \{q_{j1}, \ldots, q_{jm_j}\}$ *is a set of subsets of* Θ*, such that* $\bigcup_{1 \leq k \leq m_j} q_{jk} = \Theta$*. If the true state of the world is* θ_i *and* q_j *is asked, an answer is chosen (uniformly) randomly among the answers* q_{jk} *containing* θ_i.
- *A maximum number* N *of queries that can be asked and a target real value* $G > 0$.

The STATE DISAMBIGUATION *problem consists in deciding whether there exists a policy asking at most* N *queries that gives expected utility at least* G*. If* $\pi_\delta(\theta_i)$ *denotes the probability of identifying* θ_i *by using policy* δ*, the SD problem amounts to deciding whether there exists* δ *such that* $\sum_{1 \leq i \leq n} p(\theta_i)\pi_\delta(\theta_i)u(\theta_i) \geq G$.

Proposition 1 *(Conitzer & Sandholm 2003)* SD *is* PSPACE*-hard, even if* $N \leq n$.

Remark that (Conitzer & Sandholm 2003) does not contain any result about the exact complexity of SD (but only the PSPACE-hardness result). This exact complexity actually depends on the way N is represented in the input: if N is expressed in unary, or if N is required to be bounded by a polynomial function of $n + r$, then SD can easily be shown to be in PSPACE, which leads to the following result:

Corollary 1 *Short-term* STATE DISAMBIGUATION *is* PSPACE*-complete*[6].

From Proposition 1 we now derive the following result:

Proposition 2 POLICY EXISTENCE *for short-term EMDP is* PSPACE*-complete.*

Proof Membership to PSPACE is classical (and similar to the membership proof to PSPACE of the QBF problem, since both problems have the same structure, namely, they consist in searching through a polynomial-depth tree). Hardness comes from the following polynomial reduction from SD. Let $SD = \{\Theta, p, u, \mathcal{Q}, N, G\}$ with $N \leq |\Theta|$. Define the following short-term EMDP policy existence problem: $\mathcal{P} = \{\mathcal{S}, \mathcal{A}, \mathcal{O}, O, \mathcal{R}, b_0, H, G\}$, where: $\mathcal{S} = \Theta$, $\mathcal{A} = \mathcal{Q} \cup \{stop\}$, $\mathcal{O} = \bigcup_{1 \leq j \leq r, 1 \leq k \leq m_j} \{q_{jk}\}$, $b_0 = p$ and $H = N$. When query $q_j \in \mathcal{Q}$ is asked in state θ_i, observation q_{jk} is returned with probability $O(\theta_i, q_j, q_{jk})$ defined as in the SD problem. Action $stop$ does not return any observation. $\mathcal{R}(b, q_j) = 0, \forall q_j \in \mathcal{Q}$ (there are no action costs) and $\mathcal{R}(b, stop) = u(\theta_i)$ if $b(\theta_i) = 1$ for some i and $\mathcal{R}(b, stop) = 0$ otherwise. Then, we show that there exists a policy δ for SD of expected utility at least G if and only if there exists a corresponding policy for $\mathcal{P}$ of expected utility at least G.

First, notice that any policy δ for SD can be represented by the same decision tree of length at most N for $\mathcal{P}$. In addition, branches issuing from query actions have the same probability attached to them in both SD and $\mathcal{P}$. So, if a leaf of the tree is a disambiguated state $\{\theta_i\}$ in both SD and $\mathcal{P}$, the probability $\pi_\delta(\theta_i)$

[6]When N is expressed compactly (in binary), then the problem is most probably not in PSPACE but can only be shown to be in EXPSPACE.

is the same in both cases. Finally, since utilities incurred when state θ_i is disambiguated in the SD case and when *stop* is applied in the same situation in $\mathcal{P}$ are equal, the utility of policy δ in $\mathcal{P}$ is $\sum_{1\leq i\leq n} p(\theta_i)\pi_\delta(\theta_i)u(\theta_i)$ as in the SD case.

Thus, any policy has the same value in both SD and $\mathcal{P}$, which implies that $\mathcal{P}$ solves SD. As a consequence, POLICY EXISTENCE in short-term EMDPs is PSPACE-hard. □

Deterministic EMDP

We study here a subclass of EMDP where the observation function is deterministic, i.e. $o = O(s, a) \in \mathcal{O}$. If we write $S_a^o = \{s \in \mathcal{S}, O(s, a) = o\}$ then for any non-terminating action a, $\{S_a^o\}_{o\in O(S,a)}$ forms a partition of S. In this case, the belief state update (equation 1) writes

$$b_a^o(s) = \frac{b(s)}{Pr(o|a,b)} = \frac{b(s)}{b(S_a^o)} \text{ if } s \in S_a^o \text{ and } 0 \text{ otherwise} \quad (2)$$

In this section we will show that (i) policies for finite-horizon D-EMDP can always be expressed in size bounded by $|\mathcal{S}|H$, (ii) the value of any policy can be computed in polynomial time and (iii) a polynomial *approximation-preserving* reduction can be built from the MINIMUM SET COVER problem to finite-horizon D-EMDP. From (i) and (ii) we get that short-term D-EMDP belongs to NP. From (iii) we get that it is NP-complete and even does not admit any constant-ratio polynomial-time approximation, since MINIMUM SET COVER does not (Lund & Yannakakis 1994).

Polynomial-space policy expression in D-EMDPs

As already noticed, the size required for expressing H-steps finite horizon policies for POMDP grows exponentially with H. The same difficulty holds for finite-horizon EMDP in general, however, we show that it is not the case for finite-horizon D-EMDP. First, we show that belief state update (equation 2) implies that the belief states that can be reached at any time step have a very specific form.

Proposition 3 *Let a D-EMDP $\mathcal{P}$ be given, and let $(a_1, o_1, \ldots, a_t, o_t)$ with $t \leq H$ be a possible sequence of state / observation pairs. Let $b_{a_1,\ldots,a_t}^{o_1,\ldots,o_t}$ be the belief state after the sequence has been observed. We have:*

$$b_{a_1,\ldots,a_t}^{o_1,\ldots,o_t}(s) = \frac{b(s)}{b\left(\bigcap_{i=1}^t S_{a_i}^{o_i}\right)} \text{ if } s \in \bigcap_{i=1}^t S_{a_i}^{o_i} \text{ and 0 otherwise.}$$

Proof We prove the result by induction. It is true for $t = 1$ as a result of equation 2. Then, assume it is true for $t < H$ and let us show it is true for $t + 1$. Applying Eq 2 to $b = b_{a_1,\ldots,a_t}^{o_1,\ldots,o_t}$, we get

$$b_{a_1,\ldots,a_{t+1}}^{o_1,\ldots,o_{t+1}}(s) = \frac{b_{a_1,\ldots,a_t}^{o_1,\ldots,o_t}(s)}{b_{a_1,\ldots,a_t}^{o_1,\ldots,o_t}\left(S_{a_{t+1}}^{o_{t+1}}\right)} \text{ if } s \in S_{a_{t+1}}^{o_{t+1}} \text{ and 0 otherwise.}$$

The induction hypothesis implies $b_{a_1,\ldots,a_t}^{o_1,\ldots,o_t}(s) = \frac{b(s)}{b\left(\bigcap_{i=1}^t S_{a_i}^{o_i}\right)}$

if $s \in \bigcap_{i=1}^t S_{a_i}^{o_i}$ and 0 otherwise.

Furthermore, $b_{a_1,\ldots,a_t}^{o_1,\ldots,o_t}\left(S_{a_{t+1}}^{o_{t+1}}\right) = \frac{b\left(S_{a_{t+1}}^{o_{t+1}} \bigcap \left(\bigcap_{i=1}^t S_{a_i}^{o_i}\right)\right)}{b\left(\bigcap_{i=1}^t S_{a_i}^{o_i}\right)}$.

Thus, $b_{a_1,\ldots,a_{t+1}}^{o_1,\ldots,o_{t+1}}(s) = \frac{b(s)}{b\left(\bigcap_{i=1}^{t+1} S_{a_i}^{o_i}\right)}$ if $s \in \bigcap_{i=1}^{t+1} S_{a_i}^{o_i}$ and 0 otherwise. □

From Proposition 3, we get that $\forall t \leq H$ and for any possible partial trajectory $(a_1, o_1, \ldots, a_t, o_t)$, the current belief state $b_{a_1,\ldots,a_t}^{o_1,\ldots,o_t}$ is uniquely determined by the initial belief state b and the set of states $\left(\bigcap_{i=1}^t S_{a_i}^{o_i}\right)$. Now, given a policy δ, let us define $S_\delta^{o_1,\ldots,o_t}$ by $S_\delta^\emptyset = S$ and

$$S_\delta^{o_1,\ldots,o_t,o_{t+1}} = S_\delta^{o_1,\ldots,o_t,o_t} \cap S_{\delta(a_1,o_1,\ldots,a_t,o_t)}^{o_{t+1}} = \bigcap_{i=1}^t S_{a_i}^{o_i}$$

where for every i, $a_i = \delta(a_1, o_1, \ldots, a_{i-1}, o_{i-1})$. At any time step $t > 1$, if the current belief state is represented by $S_\delta^{o_1,\ldots,o_{t-1}}$ then the next action is $\delta(a_1, o_1, \ldots, a_{t-1}, o_{t-1})$.

The next result also holds:

Proposition 4 *For every $t \leq H$, the set of nonempty subsets of $\left\{S_\delta^{o_1,\ldots,o_t}\right\}_{(o_1,\ldots,o_t)\in\mathcal{O}^t}$ forms a partition of $\mathcal{S}$.*

Proof By induction on t. First, this is true for $S_\delta^\emptyset$. Now, assume that the set of nonempty subsets of $\left\{S_\delta^{o_1,\ldots,o_t}\right\}_{(o_1,\ldots,o_t)\in\mathcal{O}^t}$ forms a partition of $\mathcal{S}$. Then, $\left\{S_\delta^{o_1,\ldots,o_t,o_{t+1}}\right\}_{o_{t+1}\in\mathcal{O}}$ is a partition of $S_\delta^{o_1,\ldots,o_t}$ for any $(o_1, \ldots, o_t)$. Therefore, the set of nonempty subsets of $\left\{S_\delta^{o_1,\ldots,o_{t+1}}\right\}_{(o_1,\ldots,o_{t+1})\in\mathcal{O}^{t+1}}$ is a partition of $\mathcal{S}$. □

Proposition 4 implies that, at any time step t, all the reachable belief states are distinct, and thus, $|\mathcal{S}|$ is an upper bound of the number of reachable belief states at time t. This in turn implies that the policy tree corresponding to a finite-horizon, D-EMDP has size bounded by $|\mathcal{S}|H$.

Example 1 *Let $S = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{O} = \{o_1, o_2, o_3\}$ and $\mathcal{A} = \{a, b, c\}$, where all three actions are deterministic and thus defined by $S_a^{o_1} = \{1\}$, $S_a^{o_2} = \{2, 3, 4, 5\}$, $S_a^{o_3} = \{6\}$, $S_b^{o_1} = \{1, 2, 3\}$, $S_b^{o_2} = \{4, 5, 6\}$, $S_c^{o_1} = \{1, 3, 5\}$, and $S_c^{o_2} = \{2, 4, 6\}$. Let $b_0 = \langle 0.2, 0.2, 0.2, 0.2, 0.1, 0.1\rangle$. Then $b_a^{o_1} = \langle 1, 0, 0, 0, 0, 0\rangle$; $b_a^{o_2} = \langle 0, \frac{2}{7}, \frac{2}{7}, \frac{2}{7}, \frac{1}{7}, 0\rangle$; $b_{a,c}^{o_2,o_1} = \langle 0, 0, \frac{2}{3}, 0, \frac{1}{3}, 0\rangle$. Figure 1 shows a policy (left) and the supports of the reachable belief states. Notice that δ succeeds in disambiguating the state of the world.*

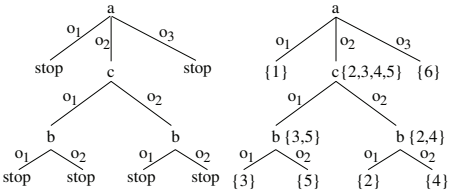

Figure 1: Left: policy. Right: Reachable belief states.

Proposition 4 does not yet allow for claiming that finite-horizon EMDP have a polynomial-space policy representation, since H may be exponentially large in the size of the input. However, the following result shows that a D-EMDP always has a short optimal policy.

Proposition 5 *Let $\mathcal{P}$ be a D-EMDP and δ a policy for $\mathcal{P}$. Then there exists a policy δ' such that $depth(\delta') \leq |\mathcal{A}| + 1$ and $V_{\delta'}(b_0) \geq V_\delta(b_0)$.*

Proof When actions are deterministic, applying the same action twice on the same branch of a policy δ will give the same observation. Therefore, on a given branch of δ, if some action a appears, applying a again later will not change the belief state, so any occurrence of a other than the first one may be removed without changing the belief state. Let δ' be the policy obtained by removing such action occurrences. The belief states obtained in δ and δ' are the same; δ and δ' may differ only on action costs. Since action costs are never negative (applying an action never increases the global utility), we have $V_{\delta'}(b_0) \geq V_\delta(b_0)$. Lastly, since each action $a \in \mathcal{A}$ appears at most once on every branch of δ', the depth of δ' is at most $|\mathcal{A}| + 1$. □

This fact allows to reduce the search for the optimal policy to policies of depth at most $|\mathcal{A}| + 1$. Therefore, any finite-horizon D-EMDP can be reduced to an equivalent short-term horizon D-EMDP (which horizon is $|\mathcal{A}| + 1$). We therefore get the following corollary:

Corollary 2 *For any finite-horizon D-EMDP problem $\mathcal{P} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, O, \mathcal{R}, b, H \rangle$, there exists an optimal policy whose depth in bounded by $|\mathcal{A}| + 1$.*

The latter result, together with Proposition 4, finally proves the following corollary:

Corollary 3 *Any finite-horizon EMDP $\mathcal{P} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, O, \mathcal{R}, b, H \rangle$ has an optimal policy that can be represented in space $O(|\mathcal{S}||\mathcal{A}|)$.*

Policy existence for D-EMDP is NP-complete

We first show that the problem of finding a policy $\delta = \{\delta_1, \ldots, \delta_H\}$ optimal with respect to $\mathcal{P}$ is in NP.

Lemma 1 POLICY EXISTENCE *for D-EMDP is in* NP.

Proof Let $\mathcal{P} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, O, \mathcal{R}, b, H \rangle$ be a short-term D-EMDP problem and v be a utility threshold. Due to Corollary 2, without loss of generality we may assume that $H \leq |\mathcal{A}| + 1$. We know from Corollary 3 that there exists an optimal policy δ that can be expressed in space $|\delta| = O(|\mathcal{S}||\mathcal{A}|)$. We now show that the value $V^\delta(b)$ can be computed in polynomial time. Indeed, this value can be computed backwards by solving the equations $V_t^\delta(b) = \mathcal{R}(b, \delta(b)) + \sum_{o \in \mathcal{O}} \left(p(o|b,a) V_{t+1}^\delta(b_a^o) \right)$, where $V_{H+1}^\delta(b) = 0, \forall b \in \mathcal{B}$. Since the number of belief states which can be reached in H time steps or less from the initial belief state b is in $O(|\mathcal{S}||\mathcal{A}|)$, they can all be generated in polynomial time and the backwards computation also takes polynomial time. So, checking whether $V^\delta(b) \geq v$ can be done in polynomial time, and the policy existence problem D-EMDP is in NP. □

NP-hardness of POLICY EXISTENCE for D-EMDP is now shown by a polynomial reduction from MINIMUM SET COVER (MSC), which is known to be NP-hard (Karp 1972).

Definition 2 (MINIMUM SET COVER) *An instance of the* MINIMUM SET COVER *(*MSC*) problem is composed of a finite set $S = \{s_1, \ldots, s_n\}$ and a collection $\mathcal{C} = \{C_1, \ldots, C_m\}$ of subsets of S. A* solution *for the* MSC *problem is a subset $\mathcal{C}' \subseteq \mathcal{C}$ which covers S, i.e. $\bigcup_{C \in \mathcal{C}'} C = S$. The measure m for the* MSC *problem is $m(\mathcal{C}') = |\mathcal{C}'|$.*

The proposed reduction computes a D-EMDP instance $\mathcal{P}_{\mathsf{MSC}}$ for any instance MSC:

Definition 3 *Let $\mathsf{MSC} = \langle \mathcal{S}, \mathcal{C} \rangle$ be an instance of Minimum Set Cover. We define the corresponding instance of policy existence in D-EMDP $\mathcal{P}_{\mathsf{MSC}} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, O, \mathcal{R}, b_0, H, G \rangle$:*

- *$\mathcal{S} = \{s_1, \ldots, s_n\}$ is identical to the* MSC *state space.*
- *$\mathcal{A} = \{a_1, \ldots, a_m\}$ where a_j corresponds to set C_j in the* MSC *problem.*
- *$\mathcal{O} = \{y, n\}$ and the deterministic function O is such that $O(s_i, a_j) = y$ if $s_i \in C_j$ and $O(s_i, a_j) = n$ if $s_i \notin C_j$.*
- *$b_0(s_i) = 1/n, \forall s_i \in \mathcal{S}$ and $H = m - 1$.*
- *$\forall b, a_j, \mathcal{R}(b, a_j)$ is defined as follows: let $M = nm$; then $\mathcal{R}(b, stop) = M$ if there is a subset C_j such that $b(C_j) = 1$; and $\mathcal{R}(b, stop) = 0$ otherwise. Then, $c(a_j) = 1, \forall a_j$.*
- *$G = M - m + 1$.*

We now show the following:

Lemma 2 *There exists a policy δ for $\mathcal{P}_{\mathsf{MSC}}$ such that $V_\delta(b_0) \geq G$ if and only if* MSC *is a positive instance of* MINIMUM SET COVER.

Proof $\Leftarrow$ Let $\mathcal{C}' = \{C_{i_1}, \ldots, C_{i_q}\}$ be a set cover of S of size at most m (thus, $q \leq m$). Then consider the following policy $\delta_{\mathcal{C}'}$:

a_{i_1}; `if` $o = yes$ `then` $stop$
`else` a_{i_2}; `if` $o = yes$ `then` $stop$
(...)
`else` $a_{i_{q-1}}$; $stop$

Clearly, the depth of $\delta_{\mathcal{C}'}$ is $q - 1 \leq m - 1$. Consider any terminal node of $\delta_{\mathcal{C}'}$. If the last action performed before $stop$ was a_{i_j} with $j < q - 1$, then the fact that the observation returned after a_{i_j} is yes implies that the final belief state b is such that $b(C_{i_j}) = 1$, therefore, $\mathcal{R}(b, stop) = M$ (b is a "good" belief state). If the last action performed before $stop$ was $a_{i_{q-1}}$, then there are two possibilities. If the returned observation is yes, then the final belief state b is such that $b(C_{i_{q-1}}) = 1$, therefore, $\mathcal{R}(b, stop) = M$. If the returned observation is no, then we know that the true state of the world is not in any of the sets $C_1, \ldots, C_{q-1}$; therefore, because $\mathcal{C}$ is a set cover of S, the true state of the world must be in C_q, henceforth $b(C_q) = 1$, and again $\mathcal{R}(b, stop) = M$. Thus, the reward associated with every terminal node is M, and each branch contains at most $m - 1$ non-stop actions, therefore, $V_{\delta_{\mathcal{C}'}}(b_0) \geq M - m + 1$.

$\Rightarrow$ Let δ be a policy of depth at most $m - 1$ and such that $V_\delta(b_0) \geq M - m + 1$. Let $a_{i_1}, \ldots, a_{i_q}$ the actions appearing on the rightmost branch of δ. Then we claim that there exists $C_{i_{q+1}}$ such that $\{C_{i_1}, \ldots, C_{i_{q+1}}\}$ is a set cover of S (of size at most m). Suppose it is not the case, that is, there exists a state s that does not belong to any of the $\{C_{i_1}, \ldots, C_{i_{q+1}}\}$. If the true state of the world is s then the obtained branch will be the rightmost branch of δ, and the reward associated with the final belief state will be 0. Since the prior probability of s is $\frac{1}{n}$, the value of δ would be at most $(1 - \frac{1}{n}).M = M - m$, which contradicts the assumption that $V_\delta(b_0) \geq M - m + 1$. Therefore, there exists a set cover of S of size at most m. □

From Lemmas 1 and 2 we get the following result:

Proposition 6 POLICY EXISTENCE *for D-EMDP is* NP*-complete.*

Finally, optimal policies for D-EMDPs are hard to approximate.

Proposition 7 *Finding a D-EMDP optimal policy is not in APX.*

Proof MSC does not belong to APX (Lund & Yannakakis 1994), i.e. no polynomial algorithm exists which approximates a MSC optimal solution within any constant factor. Since the reduction used in the proof of Lemma 2 is *approximation preserving*[7], finding an optimal policy for a D-EMDP is not in APX. □

Concluding remarks

In this paper we have defined the class of epistemic POMDPs, which are relevant for *information gathering* problems. Even though an EMDP can be seen as a degenerate case of POMDP, we have shown that solving an EMDP is as hard as solving a POMDP in general (PSPACE-complete for short-term horizons), except when knowledge-gathering actions are deterministic, in which case the policy existence problem is only NP-complete.

One of the works most related to ours is (Conitzer & Sandholm 2003), where the state disambiguation problem is shown to be PSPACE-hard. Our work extends theirs by providing a unique framework for meta-reasoning problems (finite-horizon EMDP).

Another related result can be found in (Littman 1996), where policy existence for *deterministic* POMDP (D-POMDP) is shown to be NP-complete. As a particular case of D-POMDP, D-EMDP are in NP as well. Even though D-EMDP seem simpler than D-POMDP at first glance, we show that they are NP-complete as well and not in APX, which implies that D-POMDP are not in APX either. The fact that D-EMDP belongs to NP could have been derived directly from Littman's result, however, our direct proof allows to prove a result concerning the size of optimal policies which contrasts with Littman's result about D-POMDP. More precisely, we have shown that whatever horizon is considered (polynomial in $|S|$, finite, or even infinite), optimal policies for D-EMDP have size $O(|S||A|)$, at most. For D-POMDP in general, this size is in $O(H|S|)$ and is only bounded by $(1 + |S|)^{|S|}$ (infinite horizon case).

Admittedly, our results are rather negative. However we believe they are significant for two reasons. First, before we established these results, we did not know whether the NP-hardness result for deterministic POMDP (Littman 1996) was still holding under the additional restriction that actions are purely epistemic; given the practical importance and the specificity of such purely epistemic MDPs, it is significant to know that these are not simpler than deterministic POMDPs. This, together with the PSPACE-completeness result in the general case, is a contribution to making more complete the general complexity landscape of planning under uncertainty. Second, these negative results suggest that it might not be necessarily suboptimal to reuse policy construction algorithms for general POMDPs for solving epistemic planning problems.

Finally, a related stream of work is planning under partial observability with compact action descriptions. Most planning languages use concise representations of states and actions, which makes the plan existence problem more complex than when the state and action spaces are flat (as in our paper). Rintanen (Rintanen 2004) shows that the plan existence problem for propositional non-probabilistic planning with partial observability is 2-EXP-complete and the special case with deterministic operators is EXPSPACE-complete. Obviously, the next step would consist in considering epistemic planning problems with propositional representation and identify the complexity of the plan existence problem.

Finally, even though not approximable within a constant factor in polynomial time, EMDP (and structured EMDP) may admit approximate solution algorithms. We plan to study algorithms using heuristics based on *information theory* principles (such as entropy, or the value of information).

[7] To show this, it is enough to notice that the reduction preserves the solution quality measure.

References

Ausiello, G.; Crescenzi, P.; Gambosi, G.; Kann, V.; Marchetti-Spaccamela, A.; and Protasi, M. 1999. *Complexity and Approximation – Combinatorial Optimization Problems and their Approximability Properties*. Springer-Verlag.

Boutilier, C. 2002. A pomdp formulation of preference elicitation problems. In *AAAI/IAAI*, 239–246.

Conitzer, V., and Sandholm, T. 2003. Definition and complexity of some basic metareasoning problems. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*, pp. 1099–1106.

Karp, R. M. 1972. *Complexity of Computer Computations*. R. E. Miller and J. W. Thatcher (eds.). Plenum Press, New York. chapter Reducibility among combinatorial problems.

Littman, M. L. 1996. *Algorithms for Sequential Decision Making*. Ph.D. Dissertation, Department of Computer Science, Brown University.

Lund, C., and Yannakakis, M. 1994. On the hardness of approximating minimization problems. *Journal of the ACM* 41:960–981.

Mundhenk, M.; Goldsmith, J.; Lusena, C.; and Allender, E. 2000. Complexity of finite-horizon markov decision process problems. *J. of the ACM* 47, Issue 4:681–720.

Papadimitriou, C. H., and Tsitsiklis, J. N. 1987. The complexity of Markov decision processes. *Mathematics of operations research* 12(3).

Rintanen, J. 2004. Complexity of planning with partial observability. In *Proceedings of the Fourteenth International Conference on Automated Planning and Scheduling (ICAPS 2004)*, 345–354. AAAI.

Smallwood, R., and Sondik, E. 1973. The optimal control of partially observed markov processes over the finite horizon. *Operations Research* 21:1071–1088.