Planning in Partially Observable Domains with Fuzzy Epistemic States and Probabilistic Dynamics

Nicolas Drougard[1], Didier Dubois[2],
Jean-Loup Farges[1], and Florent Teichteil-Königsbuch[1]

[1] Onera – The French Aerospace Lab
[2] IRIT

Abstract. A new translation from Partially Observable MDP into Fully Observable MDP is described here. Unlike the classical translation, the resulting problem state space is finite, making MDP solvers able to solve this simplified version of the initial partially observable problem: this approach encodes agent beliefs with fuzzy measures over states, leading to an MDP whose state space is a finite set of epistemic states. After a short description of the POMDP framework as well as notions of Possibility Theory, the translation is described in a formal manner with semantic arguments. Then actual computations of this transformation are detailed, in order to highly benefit from the factored structure of the initial POMDP in the final MDP size reduction and structure. Finally size reduction and tractability of the resulting MDP is illustrated on a simple POMDP problem.

1 Introduction

It is claimed that Partially Observable Markov Decision Processes (POMDPs) [17] finely models an agent acting under uncertainty in a partially hidden environment. However, solving a POMDP, *i.e.* the computation of an optimal strategy for the agent, is a really difficult task: the problem is PSPACE-complete [12]. Classical approaches try to solve this problem using Dynamic Programming [3], or via approximate computation. These include for instance heuristic searches [18] and Monte Carlo approaches [16].

The approach proposed here simplifies a POMDP problem before solving it. The transformation described leads to a fully observable MDP on a finite number of epistemic states, *i.e.* a problem modeling an agent acting under uncertainty in a fully observable environment [13]. As such a finite state space MDP problem is P-complete [12] this transformation qualifies as a simplification, and any MDP solver can return a policy for this translated POMDP.

Most of the POMDP algorithms draws upon the *agent belief* during the process, defined as the probability of the actual system state knowing all the system observations and agent actions from the beginning. This belief is updated at each time step using Bayes' rule and the new observation. The initial

belief, or *prior* probability distribution over the system states, takes part in the definition of the POMDP. However in practice, the initial system state can be unknown: for instance, in a robotic exploration context, the initial location of the agent, or initial presence of an entity in the scene. Defining the process with a uniform probability distribution as initial belief (*e.g.* over all locations or over entity presence) is a subjectivist answer [5], *i.e.* all probabilities are the same because no event is more plausible than another: it corresponds to equal betting rates. However the following belief updates will eventually mix up frequentist probability distributions defining the POMDP with the initial belief which is a subjective probability, and it does not always make sense.

More than only a simplification of the initial POMDP problem, the theoretical framework used here for the belief representation formally models an agent's knowledge about the system state: the proposed translation defines beliefs as *possibility distributions* over system states $s \in \mathcal{S}$: these kinds of distributions are denoted by π (counterpart of probability notation $\mathbf{p}$) and represent a fuzzy set of system states, as the indicator (chararacteristic) function of this set. Recall that the indicator function of a classical set $A \subseteq \mathcal{S}$ is $\mathbb{1}_A(s) = 1$ if $s \in A$ and 0 otherwise. Values of a fuzzy set indicator function π are chosen in a finite and totally ordered scale $\mathcal{L} = \{1 = l_1, l_2, \ldots, 0\}$ with $l_1 > l_2 > \ldots > 0$ *i.e.* $\pi : \mathcal{S} \to \mathcal{L}$. If $s \in \mathcal{S}$ is such that $\pi(s) = l_i$, s is in the fuzzy set described by π, with degree l_i. Possibilistic beliefs used in this work will represent fuzzy sets of possible states. If the current possibilistic belief coincide with the distribution $\pi(s) = 1\ \forall s \in \mathcal{S}$, all system states are totally possible, and it models therefore a total ignorance about the current system state: qualitative possibilistic beliefs can model agent initial ignorance. The perfect knowledge of the current state, say $\tilde{s} \in \mathcal{S}$, is encoded by a possibility distribution equal to the classical indicator function of the singleton $\pi(s) = \mathbb{1}_{\{s=\tilde{s}\}}(s)$. Between these two extrema, the current knowledge of the system is described by a set of entirely possible states, $\{s \in \mathcal{S} \text{ s.t. } \pi(s) = 1\}$, and successive sets of less plausible ones $\{s \in \mathcal{S} \text{ s.t. } \pi(s) = l_i\}$ down to the set of impossible states $\{s \in \mathcal{S} \text{ s.t. } \pi(s) = 0\}$.

The major originality of this work comes from the finiteness of the scale $\mathcal{L}$: the number of possible beliefs about the system state is, as well, finite (smaller than $\#(\mathcal{L}^{\mathcal{S}}) = (\#\mathcal{L})^{\#\mathcal{S}}$), while the set of all probability distributions over $\mathcal{S}$ is infinite. The translation described here leads then to an MDP whose finite state space is the set of possible possibilistic beliefs, or *epistemic states*.

In addition to POMDP simplification and knowledge modelling, this qualitative possibilistic framework offers some interesting properties: the possibilistic counterpart of Bayes' rule leads to a special belief behaviour. Indeed the agent can possibly change their mind radically and rapidly, and under some conditions the increased specificity of the belief distribution is enforced, *i.e.* the knowledge about the current state is non decreasing with time steps [6]. Finally, in order to fully define the resulting MDP, the translation has to attach a reward function to its states: as the new (epistemic) state of the problem is a possibility distribution, a dual measure, called necessity, can be computed from it. Defined as the

Choquet integral using the necessity measure, the reward of an epistemic state is a pessimistic evalutation of the actual reward.

However the number of possibilistic belief distributions, or *fuzzy epistemic states*, grows exponentially with the number of initial POMDP system states. The so called simplification of the problem does not transform the PSPACE POMDP problem into a polynomial one: as the new state space size is exponential in the previous one, the resulting problem is EXPTIME. The proposed translation tries to generate as few epistemic states as possible taking carefully into account potential factorized structures of the initial POMDP.

The first section is devoted to the presentation of the Markov Decision Processes, the main concern of this paper. Tools from Possibility Theory are also defined to make this paper self-contained. Follows a section describing the first contribution of this work, which is the translation itself, presented in a formal way. As the resulting state space of the built MDP is too big to make this problem tractable without factorization tricks in practice, the next section details the proper way to preprocess its attributes. Finally, the last section illustrates the relevance of this approach with a simple robotic mission problem.

2 Background

The work developed in this paper remains in the classical MDP and POMDP frameworks, which are recalled in this section: possibilistic material necessary to build the promised translation are then presented.

2.1 Markov Decision Processes

A Markov Decision Process (MDP) [1] is a well suited framework for sequential decision making under uncertainty, when the agent involved has a full knowledge of the actual system state. Such a process is formally defined by a 4-tuple $\langle \mathcal{S}, \mathcal{A}, T, r \rangle$ where $\mathcal{S}$ is a finite set of system states $s \in \mathcal{S}$. The finite set $\mathcal{A}$ consists of all actions $a \in \mathcal{A}$ available for the agent. The Markov dynamics of the system is described by the transition function $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. This function is defined as the transition probability distribution of the system states: if action $a \in \mathcal{A}$ is chosen by the agent, and the current system state is $s \in \mathcal{S}$, the next state $s' \in \mathcal{S}$ is reached with probability $T(s, a, s') = \mathbf{p}(s' \mid s, a)$. Finally, a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined to model the goal of the agent. Indeed, solving an infinite horizon MDP problem consists in computing a *strategy*, *i.e.* a function d defined on $\mathcal{S}$ and whose values are actions $a \in \mathcal{A}$, maximizing the expected discounted total reward: $\mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t \cdot r(s_t, d_t)\right]$ where $d_t = d(s_t)$ and $0 < \gamma < 1$ is the discount factor.

A Partially Observable MDP (POMDP) [17] makes a step further in the modeling flexibility, allowing the agent not to know which system state is the current one. The formal definition of a POMDP is the 7-tuple $\langle \mathcal{S}, \mathcal{A}, T, \Omega, O, r, b_0 \rangle$, where the system state $\mathcal{S}$, the set of actions $\mathcal{A}$, the transition function T and the reward function r remain the same as for the MDP definition. In this model, the current system state $s \in \mathcal{S}$ cannot be used as available information for the agent: the

agent knowledge about the actual system state comes from observations $o \in \Omega$, where Ω is a finite set. The observation function $O : \mathcal{S} \times \mathcal{A} \times \Omega \to [0,1]$ gives for each action $a \in \mathcal{A}$ and reached system state $s' \in \mathcal{S}$, the probability over possible observations $o' \in \Omega$: $O(s', a, o') = \mathbf{p}(o' \mid s', a)$. Finally, the initial belief $b_0 : \mathcal{S} \to [0,1]$ is the *prior* probability distribution over the state space $\mathcal{S}$: $b_0(s) = \mathbf{p}(s_0 = s)$, $\forall s \in \mathcal{S}$.

At a given time step $t > 0$, the agent belief is defined as the probability of the t^{th} system state s_t conditionned on all the past actions and observations, and with the prior b_0, *i.e.* $b_t(s) = \mathbf{p}_{s_0 \sim b_0}(s_t = s \mid a_0, o_1, \ldots, a_{t-1}, o_t)$. It can be easily recursively computed using Bayes' rule: at time step t, if the belief is b_t, chosen action $a \in \mathcal{A}$ and new observation $o' \in \Omega$, next belief is $b_{t+1}(s') \propto O(s', a, o') \cdot \sum_{s \in \mathcal{S}} T(s, a, s') \cdot b_t(s)$. Successive beliefs are computed with the observations perceived by the agent, and are then available during the process. Let us denote by $\mathbb{P}_{\mathcal{S}}$ the infinite set of probability distributions over $\mathcal{S}$: seen as an MDP whose states are probabilistic beliefs, an optimal strategy for the infinite horizon POMDP is looked for among strategies $d : \mathbb{P}_{\mathcal{S}} \to \mathcal{A}$ such that successive $d_t = d(b_t)$ maximize the expected discounted total reward, which can be rewritten

$$\mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t \cdot r(s_t, d_t)\right] = \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t \cdot r(b_t, d_t)\right], \tag{1}$$

defining $r(b_t, a) = \sum_{s \in \mathcal{S}} r(s, a) \cdot b_t(s)$ as the reward of belief b_t. As the focused problem (POMDP) has been formally defined, Possibilistic tools are now presented in the next section.

2.2 Possibility Theory

In our context, distributions defined in the Possibility Theory framework are valued in a totally ordered scale $\mathcal{L} = \{1 = l_1, l_2, \ldots, 0\}$ with $l_1 > l_2 > \ldots > 0$. A possibility measure Π defined on $\mathcal{S}$ is a fuzzy measure valued in $\mathcal{L}$, such that $\forall A, B \subset \mathcal{S}$, $\Pi(A \cup B) = \max\{\Pi(A), \Pi(B)\}$, $\Pi(\emptyset) = 0$ and $\Pi(\mathcal{S}) = 1$. It follows that this measure is entirely defined by the associated possibility distribution, *i.e.* the measure of the singletons: $\forall s \in \mathcal{S}$, $\pi(s) = \Pi(\{s\})$. Properties of this measure lead to the possibilistic normalization: $\max_{s \in \mathcal{S}} \pi(s) = \Pi(\mathcal{S}) = 1$. If $\overline{s}, \underline{s} \in \mathcal{S}$ are such that $\pi(\overline{s}) < \pi(\underline{s})$, it means that $\overline{s}$ is less plausible than $\underline{s}$. States with possibility degree 0, *i.e.* states $s \in \mathcal{S}$ such that $\pi(s) = 0$, are impossible (same meaning as $\mathbf{p}(s) = 0$), and those such that $\pi(s) = 1$ are entirely possible (but not necessary the most probable one).

After the introduction of a possibility measure over a set Ω, the joint possibility measure on $\mathcal{S} \times \Omega$ is defined in a qualitative way: $\forall A \subset \mathcal{S}$, $\forall B \subset \Omega$

$$\Pi(A, B) = \min\{\Pi(A \mid B), \Pi(B)\} = \min\{\Pi(B \mid A), \Pi(A)\}. \tag{2}$$

Note the similarities between Possibility and Probability Theory, replacing max by + and min by $\times$. Moreover, Possibility Theory has its own conditioning [9]:

$$\Pi(A \mid B) = \begin{cases} 1 \text{ if } \Pi(A, B) = \Pi(B) \\ \Pi(A, B) \text{ otherwise} \end{cases} \tag{3}$$

which is nothing more than the least specific measure fulfilling the condition described by Equation 2. It can also be seen more easily as the joint measure normalized in a possibilistic manner.

These tools from Qualitative Possibility Theory are enough to define the announced translation. Next section is then devoted to the building of an MDP with fuzzy epistemic states from a POMDP.

3 A Hybrid POMDP

As claimed by Zadeh, "most information/intelligent systems will be of hybrid type" [19]: the idea developped here is to use a granulated representation of the agent knowledge using possibilistic beliefs instead of probabilistic beliefs in the POMDP framework. The first advantage of this granulation is that strategy computations are performed reasoning on a finite set of possibilistic beliefs (called then epistemic states): the set of all possibility distributions defined over $\mathcal{S}$, denoted by $\Pi_{\mathcal{S}}$ is the set $\mathcal{L}^{\mathcal{S}}$ without non-normalized functions, and then

$$\#\Pi_{\mathcal{S}} = \#\mathcal{L}^{\#\mathcal{S}} - (\#\mathcal{L} - 1)^{\#\mathcal{S}}, \tag{4}$$

while the set of probability distributions over $\mathcal{S}$ is infinite. First, such beliefs are formally defined, as well as their own updates.

3.1 Possibilistic Belief

Consider that possibility distibutions similar to those used to define the initial POMDP are available: a transition distribution, giving the possibility degree of reaching $s' \in \mathcal{S}$ from $s \in \mathcal{S}$ using action $a \in \mathcal{A}$, $\pi(s' \mid s, a) \in \mathcal{L}$; as well as an observation one, giving the possibility degree of observing $o' \in \Omega$, in a system state $s' \in \mathcal{S}$ after the use of $a \in \mathcal{A}$, $\pi(o' \mid s', a) \in \mathcal{L}$. Indeed, this work is devoted to few kinds of practical problems: real problems modeled as POMDPs are often intractable. Our granulated approach is in this case a simplification of the initial POMDP, and possibility distributions are computed from the POMDP probability distributions, using a possibility-probability transformation [10]. On the other hand, some problems lead to POMDPs with partially defined probability distributions: some estimated probabilities have no strong guarantees. A more faithful representation is given with possibility distributions modeling the inherent imprecision, defining transition and observation possibility distributions.

Let $b_0^\pi : \mathcal{S} \to \mathcal{L}$ be an initial possibilistic belief, normalized as any possibility distribution: $\max_{s \in \mathcal{S}} b^\pi(s) = 1$. As in the probabilistic case, possibilistic belief can be defined recursively using the possibilistic belief update [6], derived from Bayes' rule based on the conditioning (3): at time step t, if the possibilistic belief is b_t^π, action $a \in \mathcal{A}$ and observation $o' \in \Omega$ specify the next belief

$$b_{t+1}^\pi(s') = u(b_t^\pi, a, o')(s') = \begin{cases} 1 \text{ if } \pi(o', s' \mid b_t^\pi, a) = \max\limits_{\tilde{s} \in \mathcal{S}} \pi(o', \tilde{s} \mid b_t^\pi, a) \\ \pi(o', s' \mid b_t^\pi, a) \text{ otherwise} \end{cases} \tag{5}$$

where the joint possibility distribution over $\Omega \times \mathcal{S}$ $\pi(o', s' \mid b_t^\pi, a)$ is equal to $\max_{s \in \mathcal{S}} \min\{\pi(o' \mid s', a), \pi(s' \mid s, a), b_t^\pi(s)\}$. Note that keeping a qualitative

view for the belief update, *i.e.* using the min operator to compute joint possibility distributions as defined in Equation 2, allows to reason on a finite set of beliefs, as no new values are created: the classical product is used in the quantitative part of the Possibility Theory, but is not considered in this work. Moreover, the use of the qualitative belief update has already been used in planning [7].

3.2 Setting up Transition Functions

If the agent selects the action $a \in \mathcal{A}$ in the epistemic state $b^\pi \in \Pi_\mathcal{S}$, the next epistemic state depends only on the next observation, as highlighted by possibilistic belief update (5). The probability distribution over observations conditionned on the reached state is part of the POMDP definition via the observation function O. The probability distribution over observations conditionned on the previous state is obtained using transition function T: $\mathbf{p}(o' \mid s, a) = \sum_{s' \in \mathcal{S}} O(s', a, o') \cdot T(s, a, s')$. This distribution and the possibilistic belief b^π about the system state, can lead to an approximated probability distribution over the next observations. Indeed, a probability distribution over the system state, $\overline{b^\pi} \in \mathbb{P}_\mathcal{S}$, can be derived from b^π using extension of Laplace principle. Then approximate distribution over $o' \in \Omega$ is defined as

$$\mathbf{p}(o' \mid b^\pi, a) = \sum_{s \in \mathcal{S}} \mathbf{p}(o' \mid s, a) \cdot \overline{b^\pi}(s). \tag{6}$$

Finally, summing over concerned observations, the transition probability distribution over epistemic states is defined as

$$\tilde{T}(b^\pi, a, (b^\pi)') = \mathbf{p}((b^\pi)' \mid b^\pi, a) = \sum_{o' \mid u(b^\pi, a, o') = (b^\pi)'} \mathbf{p}(o' \mid b^\pi, a). \tag{7}$$

A proper way to construct a probability distribution $\overline{b^\pi}$, from a possibility one b^π, is the use of the pignistic transformation [8], minimizing the arbitrariness in the translation into probability distribution: numbering system states with the order induced by distribution b^π, $1 = b^\pi(s_1) \geqslant b^\pi(s_2) \geqslant \ldots \geqslant b^\pi(s_{\#\mathcal{S}+1}) = 0$, with $s_{\#\mathcal{S}+1}$ an artificial state such that $\pi(s_{\#\mathcal{S}+1}) = 0$ introduced to simplify the formula,

$$\overline{b^\pi}(s_i) = \sum_{j=i}^{\#\mathcal{S}} \frac{b^\pi(s_j) - b^\pi(s_{j+1})}{j} \tag{8}$$

Note that this probability distribution corresponds to the center of gravity of the probability distributions family induced by the possibility measure defined by distribution b^π [10], and respects the Laplace principle of Insufficient Reason (ignorance leads to uniform probability).

3.3 Reward Aggregation

After the transition function, it remains to assign a reward to each epistemic state: in the classical probabilistic translation, the reward assigned to a belief b is

the reward expectation according to the probability distribution b: $\sum_{s\in\mathcal{S}} r(s,a)\cdot b(s)$. Here, the agent knowledge is represented with a possibility distribution b^π, which is less informative than a probability one: it accumulated uncertainty due to possibilistic discretization and due to possible agent ignorance. A way to define a reward being pessimistic about these uncertainties is to aggregate the reward using the dual measure of the possibility distribution, and the *Choquet integral*.

The dual measure of a possibility measure $\Pi : 2^{\mathcal{S}} \to \mathcal{L}$ is called *necessity measure* and is denoted by $\mathcal{N}$. This measure is defined by $\forall A \subseteq \mathcal{S}$, $\mathcal{N}(A) = 1-\Pi(\overline{A})$ where $\overline{A}$ is the complementary set of $A : \overline{A} = \mathcal{S}\setminus A$. Recall notation $\mathcal{L} = \{ l_1 = 1, l_2, l_3, \ldots, 0 \}$. For a given action $a \in \mathcal{A}$, reward values, $\{ r(s,a) \mid s \in \mathcal{S} \}$ are denoted by $\{ r_1, r_2, \ldots, r_k \}$ with $r_1 > r_2 > \ldots > r_k$, and $k \leqslant \#\mathcal{S}$. An artificial value $r_{k+1} = 0$ is also introduced to simplify the formulae.

The discrete Choquet integral of the reward function with the necessity measure $\mathcal{N}$ is defined, and then simplified, as follows:

$$Ch(r,\mathcal{N}) = \sum_{i=1}^{k} (r_i - r_{i+1}) \cdot \mathcal{N}\left(\{ r(s) \geqslant r_i \}\right) = \sum_{i=1}^{\#\mathcal{L}-1} (l_i - l_{i+1}) \cdot \min_{\substack{s\in\mathcal{S} \text{ s.t.} \\ \pi(s)\geqslant l_i}} r(s). \quad (9)$$

More on possibilistic Choquet integrals can be found in [4]. This reward aggregation using the necessity measure leads to a pessimistic estimation of the reward: as an example, the reward $\min_{s\in\mathcal{S}} r(s,a)$ is assigned to the total ignorance. Note that, if the necessity measure $\mathcal{N}$ is replaced by a probability measure $\mathbb{P}$, Choquet integral coincides with the expected reward based on $\mathbb{P}$.

3.4 MDP with Epistemic States

This section summarizes the complete translation using the final equations of the previous sections. This translation takes for input a POMDP: $\langle \mathcal{S}, \mathcal{A}, T, \Omega, O, r \rangle$ and returns an epistemic states based MDP: $\langle \tilde{\mathcal{S}}, \mathcal{A}, \tilde{T}, \tilde{r} \rangle$. The state space is $\tilde{\mathcal{S}} = \Pi_{\mathcal{S}}$. The (approximate) transition functions are $\tilde{T}$, such that $\forall (b^\pi, \tilde{b}^\pi) \in \Pi^2_{\mathcal{S}}$, $\forall a \in \mathcal{A}$, $\tilde{T}(b^\pi, a, \tilde{b}^\pi) = \mathbf{p}\left(\tilde{b}^\pi \,\middle|\, b^\pi, a \right)$ defined with Equations 6 and 7. The reward of a belief b^π is $\tilde{r}(a, b^\pi) = Ch\left(r(a,.), \mathcal{N}_{b^\pi}\right)$, defined with Equation 9 and where $\mathcal{N}_{b^\pi}$ is the necessity measure computed from b^π. Finally, as in the probabilistic framework (see Equation 1), the criterion of this MDP is the expected total reward: $\mathbb{E}_{(b_t^\pi)\sim\tilde{T}}\left[\sum_{t=0}^{+\infty} \gamma^t \cdot \tilde{r}\left(b_t^\pi, d_t\right) \right]$.

While the resulting state space is finite, only really small POMDP problems can be solved with this translation without computation tricks. Indeed, $\Pi_{\mathcal{S}}$ grows exponentially with the number of system states (see Equation 4), which makes the problem intractable even for state of the art MDP solvers.

Purely possibilistic counterparts of the (PO)MDPs, called Qualitative Possibilistic (PO)MDPs, have been already defined [14] and efficiently used for planning under uncertainty problems [7]. These π-(PO)MDPs are quite different from the model exposed in this paper. For instance, they do not use quantitative data

as probabilities or rewards. Dynamics is described in a purely qualitative possibilistic way. Frequentist information about the problem cannot be encoded: these frameworks are indeed dedicated to situations where the probabilistic dynamic of the studied system is lacking. Moreover, possible values of the reward function are chosen among the degrees of the qualitative possibilistic scale. A commensurability assumption between reward and possibility degrees, i.e. a meaning of why they share the same scale, is needed to use the criteria proposed in these frameworks. Our model bypass these demands: a real number is assigned to each possibilistic belief (epistemic state), using the Choquet integral, instead of a qualitative utility degree: it represents the reward got by the agent when reaching this belief (in an MDP fashion) as detailed in Section 3.3. Moreover, the dynamics of our process is described with probability distributions: approximate probabilistic transition functions between current and next beliefs, or epistemic states, are given in Section 3.2. Finally, our model can be solved by any MDP solver in practice: it becomes eventually a classical probabilistic fully observable MDP whose state space is the finite set $\Pi_{\mathcal{S}}$. Here, the term hybrid is used because the beliefs only are defined as possibility distributions, and all variables keep a probabilistic dynamic: the agent reasons based on a possibilistic analysis of the system state (the possibilistic belief, or epistemic state), and transition probability distributions are defined for its epistemic states.

4 Benefit from Factorization

This section carefully derive a tractable MDP problem from a factored POMDP [2]: the resulting MDP is equivalent to the former translation, but some factorization and computational tricks are described here to reduce its size and to make it factorized First, the definition of a factored POMDP is quickly outlined, followed by some notations about variable dependences helpful for describing how distributions are dealt with. Next, a classification of the state variables is made to strongly adapt computations according to the nature of the system state. The way how possibility distributions are defined is presented, and the description of the use of the possibilistic Bayes' rule in practice ends this section.

4.1 Factored POMDPs

Partially Observable Markov Decision Processes can be defined in a factorized way. The state space is described with Boolean variables of the set $\mathbb{S} = \{ s_1, \ldots, s_m \}$: $\mathcal{S} = s_1 \times \ldots \times s_m$. The notation $\mathbb{S}' = \{ s'_1, \ldots, s'_m \}$ is also used. The set of Boolean observation variables $\mathbb{O} = \{ o_1, \ldots, o_n \}$ describes also the observation space $\Omega = o_1 \times \ldots \times o_n$. For simplicity, and as state $s \in \mathcal{S}$ and observation $o \in \Omega$ notations are no longer reused in this paper, only variables are denoted with these letters from now: $s_j \in \mathbb{S}$ and $o_i \in \mathbb{O}$.

The factorized description continues defining, $\forall j \in \{1, \ldots, m\}$ and $\forall a \in \mathcal{A}$, a transition function $T_j^a(\mathbb{S}, s'_j) = \mathbf{p}\left(s'_j \mid \mathbb{S}, a\right)$, about the state variable s'_j. One observation function is also given for each observation variable: $O_i^a(\mathbb{S}', o'_i) = \mathbf{p}\left(o'_i \mid s'_1, \ldots, s'_m, a\right)$, $\forall i = 1, \ldots, n$ and $\forall a \in \mathcal{A}$. It is here understood that

$\mathbb{S}'$ are independent conditionned on $\mathbb{S}$ and the action a, and that $\{ o'_i \}_{i=1}^n$ are independent conditionned on $\mathbb{S}'$ and a.

4.2 Notations and Observation Functions

Transitions of the final MDP make it more handy if each variable depends on only few previous variables: the procedure to avoid blocking such simplifications brought by the structure of the initial POMDP during the translation, needs the following notations. In practice, for each $i \in \{ 1, \dots, n \}$ not all state variables influence observation variable o'_i; similarly, for each $j \in \{ 1, \dots, m \}$, not all current state variables influence next state variable s'_j: observation variable o'_i depends on some state variables which are called *parents* of o'_i as they appears as "parents nodes" in a Bayesian network illustrating dependencies of the process, and denoted by $\mathcal{P}(o'_i) = \left\{ s'_j \in \mathbb{S}' \text{ s.t. } o'_i \text{ depends on } s'_j \right\}$. As well, probability distributions of next state variable s'_j depend on some current state variables, denoted by $\mathcal{P}(s'_j) = \left\{ s_k \in \mathbb{S} \text{ s.t. } s'_j \text{ depends on } s_k \right\}$. It leads to the following rewriting of probability distributions: $T_j^a(\mathcal{P}(s'_j), s'_j) = \mathbf{p}\left(s'_j \,\middle|\, \mathcal{P}(s'_j), a \right)$ and $O_i^a(\mathcal{P}(o'_j), o'_i) = \mathbf{p}\left(o'_i \,|\, \mathcal{P}(o'_j), a \right)$. Finally, the following subset of $\mathbb{S}$ is useful to specify observation dynamics: $\mathcal{Q}(o'_i) = \left\{ s_k \in \mathbb{S} \text{ s.t. } \exists s'_j \in \mathcal{P}(o'_i) \text{ s.t. } s_k \in \mathcal{P}(s'_j) \right\} = \cup_{s'_j \in \mathcal{P}(o'_i)} \mathcal{P}(s'_j) \subseteq \mathbb{S}$. Probability distributions of variables $\mathcal{P}(o'_i)$ profit also from previous rewritings: thanks to state variables independences, $\forall i = 1, \dots, n$,

$$\mathbf{p}\left(\mathcal{P}(o'_i) \,|\, \mathbb{S}, a \right) = \prod_{s'_j \in \mathcal{P}(o'_i)} T_j^a(\mathcal{P}(s'_j), s'_j) = \mathbf{p}\left(\mathcal{P}(o'_i) \,|\, \mathcal{Q}(o'_i), a \right) \tag{10}$$

The observation probability distributions knowing previous state variables are

$$\forall i = 1, \dots, n, \quad \mathbf{p}\left(o'_i \,|\, \mathcal{Q}(o'_i), a \right) = \sum_{v \in 2^{\mathcal{P}(o'_i)}} \mathbf{p}\left(o'_i \,|\, v, a \right) \cdot \mathbf{p}\left(v \,|\, \mathcal{Q}(o'_i), a \right). \tag{11}$$

Therefore a possibilistic belief defined on $2^{\mathcal{Q}(o'_i)}$ is enough to get the approximate probability distribution of an observation variable: such an epistemic state, leads to a probability distribution $\overline{b^\pi}$ over $2^{\mathcal{Q}(o'_i)}$ via the pignistic transformation (8). The approximate probability distribution of the i^{th} observation variable, factorized counterpart of Equation 6, is: $\forall i = 1, \dots, n$,

$$\mathbf{p}\left(o'_i \,|\, b^\pi, a \right) = \sum_{v \in 2^{\mathcal{Q}(o'_i)}} \mathbf{p}\left(o'_i \,|\, v, a \right) \cdot \overline{b^\pi}(v). \tag{12}$$

4.3 State Variable Classification

State variables $s \in \mathbb{S}$ do not play the same role in the process: as already studied in the literature [11], some variables can be visible for the agent, and this *mixed-observability* leads to important computational simplifications. Moreover, some variables do not affect observation variables, and this structure profits in the final MDP complexity.

- A state variable s_j is said to be **visible**, if $\exists o_i \in \mathbb{O}$, observation variable, such that $\mathcal{P}(o'_i) = \{ s'_j \}$ and $\forall a \in \mathcal{A}$, $\mathbf{p}\left(o'_i \mid s'_j, a\right) = \mathbb{1}_{\{ o'_i = s'_j \}}$ *i.e.* if $o'_i = s'_j$ almost surely. The set of visible state variables is denoted by $\mathbb{S}_v = \{ s_{v,1}, s_{v,2}, \ldots, s_{v,m_v} \}$. The observation variables corresponding to the visible state variables can be removed from the set of observation variables: the number of observation variables becomes $\tilde{n} = n - m_v$.
- **Inferred hidden variables** are simply $\cup_{i=1}^{\tilde{n}} \mathcal{P}(o'_i)$, *i.e.* all hidden variables influencing (remaining) observation variables. The set of inferred hidden variables is $\mathbb{S}_h = \{ s_{h,1}, s_{h,2}, \ldots, s_{h,m_h} \}$ and contains possibly visible variables.
- **Non-inferred hidden variables** or **fully hidden variables**, denoted by $\mathbb{S}_f$, consists of hidden state variables which do not influence any observation, *i.e.* all remaining state variables. The fully hidden variables are denoted by $s_{f,1}, s_{f,2}, \ldots, s_{f,m_f}$, and the corresponding set is $\mathbb{S}_f$.

The classification allows to avoid some computations for visible variables: if $s_v \in \mathbb{S}_v$, and o_v is the associated observation, computations of the distribution over $\mathcal{P}(o'_v)$, Equation 10, and of the distribution over o'_v, Equation 11, are unnecessary: the distribution over s'_v $(= o'_v)$ needed is simply given by $T^a\left(\mathcal{P}(s'_v), s'_v\right)$. The counterpart of Equation 12 is then simply

$$\mathbf{p}\left(s'_v \mid b^\pi, a\right) = \sum_{2^{\mathcal{P}(s'_v)}} T^a\left(\mathcal{P}(s'_v), s'_v\right) \cdot \overline{b^\pi}(\mathcal{P}(s'_v)) \tag{13}$$

where $\overline{b^\pi}$ is the probability distribution over $2^{\mathcal{P}(s'_v)}$ extracted from the possibilistic belief over the same space, using pignistic transformation (8).

4.4 Beliefs Process Definition and Handling

This section is meant to define marginal beliefs instead of a global one, in order to profit of the structure of the initial POMDP. Possibilistic belief distributions have different definitions according to which class of state variables they concern.

As visible state variables are directly observed, there is no uncertainty over these variables. Two epistemic states (possibilistic belief distribution) are possible for visible state variable $s'_{v,j}$: $b'_{v,T}(s'_{v,j}) = \mathbb{1}_{\{ s'_{v,j} = \top \}}$ and $b'_{v,F}(s'_{v,j}) = \mathbb{1}_{\{ s'_{v,j} = \bot \}}$. As a consequence, one Boolean variable $\beta'_{v,j} \in \{ \top, \bot \}$ per visible state variables is enough to represent this belief distribution in practice: if $s'_{v,j} = \top$, then next belief is $b' = b'_{v,T}$ represented by belief variable assignment $\beta'_{v,j} = \top$, otherwise, next belief is $b' = b'_{v,F}$, and $\beta'_{v,j} = \bot$.

For each $i \in 1, \ldots, \tilde{n}$, each inferred hidden variable constituting $\mathcal{P}(o'_i)$ is an input of the same possibilistic belief distribution: non-normalized belief is

$$\forall i = 1, \ldots, \tilde{n}, \quad \tilde{b'}(\mathcal{P}(o'_i)) = \max_{v \in 2^{\mathcal{Q}(o'_i)}} \min\left\{ \pi\left(o'_i, \mathcal{P}(o'_i) \mid v, a\right), b(v) \right\}, \tag{14}$$

where the joint possibility distributions over $o'_i \times \mathcal{P}(o'_i)$ are $\pi\left(o'_i, \mathcal{P}(o'_i) \mid \mathcal{Q}(o'_i), a\right)$ $= \min\left\{ \pi\left(o'_i \mid \mathcal{P}(o'_i), a\right), \min_{s'_j \in \mathcal{P}(o'_i)} \pi\left(s'_j \mid \mathcal{P}(s'_j), a\right) \right\}$. The possibilistic normalization, $\forall w \in 2^{\mathcal{P}(o'_i)}$, $b'(w) = \begin{cases} 1 \text{ if } w \in \operatorname{argmax}_{v \in 2^{\mathcal{P}(o'_i)}} \tilde{b'}(v); \\ \tilde{b'}(w) \text{ otherwise.} \end{cases}$ finalizes this

rewriting of the belief update (5). In practice, if $l = \#\mathcal{L}$, and $p_i = \#\mathcal{P}(o'_i)$, the number of belief states is $l^{2^{p_i}} - (l-1)^{2^{p_i}}$, and then the number of belief variables is $n_{h,i} = \lceil \log_2(l^{2^{p_i}} - (l-1)^{2^{p_i}}) \rceil$. A belief variable of an inferred hidden state variable is denoted by β_h.

For each $j \in 1, \ldots, m_f$, non-normalized belief defined on fully hidden variable $s_{f,j}$ is

$$\tilde{b}'(s'_{f,j}) = \max_{v \in 2^{\mathcal{P}(s'_{f,j})}} \min \left\{ \pi \left(s'_{f,j} \,\middle|\, v, a \right), b(v) \right\}, \tag{15}$$

which leads to the actual new belief b' after the possibilistic normalization. As each fully hidden variable is considered independently from the others, the number of belief variables is $n_f = \lceil \log_2(l^2 - (l-1)^2) \rceil = \lceil \log_2(2l-1) \rceil$. A belief variable of a fully hidden state variable is denoted by β_f. Finally, the global epistemic state is $b'(\mathbb{S}') = \min \left\{ \min_{j=1}^{m_v} b'(s'_{v,j}), \min_{i=1}^{\tilde{n}} b'(\mathcal{P}(o'_i)), \min_{k=1}^{m_f} b'(s'_{f,k}) \right\}$.

Note that the belief over the inferred hidden variables (14) and the distribution over observation variables (12), need a belief distribution over $\mathcal{Q}(o'_i) \subseteq \mathbb{S}$. As well, the belief over the fully hidden state variables (15) needs a belief distribution over variables $\mathcal{P}(s'_{f,j}) \subseteq \mathbb{S}$. Moreover, a belief distribution over $\mathcal{P}(s'_{v,i}) \subseteq \mathbb{S}$ is needed to define an approximate probability distribution over visible state variables (13). These beliefs can be computed marginalizing the global belief using the max operator over unused state variables.

5 Solving a POMDP with a Discrete MDP Solver

A practical version of the factored MDP achieved in the previous section is described here. A concrete POMDP problem and the resulting MDP illustrate then the state space size reduction of our detailed possibilistic translation.

5.1 Resulting Factored MDP

A belief update depends only on the next observation (see Equation 5): the transition of a belief is then deterministic conditionned on the next observation. A simple trick is used to keep this determinism in the final MDP: a $flipflop$ Boolean variable is introduced, changing its state at each step, denoted by f. It artificially divides a classical time step of the POMDP into two phases. During the first phase, called *the observation generation phase*, non-identity transition functions are the probability distributions over observation variables (12) and visible state variables (13). During the second phase, called *the belief update phase*, non-identity transition functions are the deterministic transitions of the belief variables: variables β_v are updated knowing value of corresponding visible variable s_v; variables $\beta_h^1, \ldots, \beta_h^{n_{h,i}}$ are updated knowing value of observation variables o_i, and using update (14); finally, variables $\beta_f^1, \ldots, \beta_f^{n_f}$ are updated using update (15). The state space is then defined as: $\mathcal{S} = f \times s_v^1 \times \ldots \times s_v^{m_v} \times o^1 \times \ldots \times o^{\tilde{n}} \times \beta_v^1 \times \ldots \times \beta_v^{m_v} \times \beta_h^1 \times \ldots \times \beta_h^{\tilde{n}} \times \beta_f^1 \times \ldots \times \beta_f^{m_f}$, where $\forall i = 1, \ldots, \tilde{n}$, β_h^i represents Boolean variables $\beta_h^{1,i}, \ldots, \beta_h^{n_{h,i},i}$, and $\forall k = 1, \ldots, m_f$, β_f^j represents Boolean variables $\beta_f^{1,j}, \ldots, \beta_f^{n_f,j}$. Figure 1 is the Influence Diagram

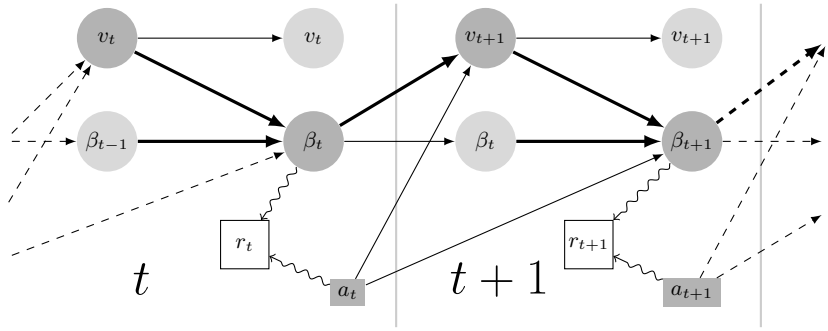

Fig. 1. ID of the resulting MDP: thickest arrows are non-identity transitions.

(ID) of the resulting MDP where β_t represents all belief variables, and v_t the visible variables: the flipflop variable f, observations and visible state variables. The resulting MDP is a factored MDP thanks to the flipflop trick.

5.2 For a Concrete POMDP

A problem inspired by the RockSample problem [18] is described in this section to illustrate the factorized possibilistic discretization of the agent belief, from a factored POMDP: a rover is navigating in a place described by a finite number of locations $l_1, \ldots, l_n$, and where stand m rocks. Some of these m rocks have an interest in the scientific mission of the rover, and it has to sample them. However, sampling a rock is an expensive operation. The rover is thus fitted with a long range sensor making him able to estimate if the rock has to be sampled. Finally operating time of the rover is limited, but its battery level is available.

Variables of this problem can now be set, and classified as in Section 4.3: as the battery level is directly observable by the agent (the rover), the set of visible state variables consists of the Boolean variables encoding it: $\mathbb{S}_v = \{ B_1, B_2, \ldots, B_k \}$. The agent knows the different locations of the rocks, however the nature of each rock is estimated. The set of inferred hidden state variables consists of m Boolean variables R_i encoding the nature of the i^{th} rock, $\top$ for "scientifically good" and $\bot$ otherwise: $\mathbb{S}_h = \{ R_1, R_2, \ldots, R_m \}$. When the i^{th} rock is observed using the sensor, it returns a noisy observation of the rock in $\{ \top, \bot \}$, modeled by the Boolean variable O_i: the set of observation variables is then $\mathbb{O} = \{ O_1, O_2, \ldots, O_m \}$. Finally, no localization equipment is provided: the agent estimates its location from its initial information, and its actions. Each location of the rover is formally described by a variable L_j, which equals $\top$ if the rover is at the j^{th} location, and $\bot$ otherwise. The set of fully hidden variables consists thus of these n variables: $\mathbb{S}_f = \{ L_1, L_2, \ldots, L_n \}$. Initial location is known, leading to a deterministic initial belief: $b_0^\pi(\mathbb{S}_h) = 1$ if $L_1 = \top$ and $L_j = \bot \; \forall j \neq 1$, and 0 otherwise. However initial nature of each rock is not known. Instead of a uniform probability distribution, the Possibility Theory allows to represent the initial ignorance about rock natures with the belief $b_0^\pi(\mathbb{S}_h) = 1$, for each variable assignment.

Classical POMDP solvers are based on probabilistic beliefs over the state space defined by $\mathbb{S}_h$, $\mathbb{S}_f$ and even $\mathbb{S}_v$ if Mixed-Observability [11] is not taken into account. The approach presented in this paper leads to an MDP with a finite space of epistemic states. Finally, the factorization tricks lead to a reduction of the state space size: with a flat translation of this POMDP, $\lceil \log_2(\#\mathcal{L}^{2^{n+m+k}} - (\#\mathcal{L}-1)^{2^{n+m+k}}) \rceil$ Boolean variables are necessary. Taking advantage of the POMDP structure, the resulting state space is encoded with $1 + 2k + m + (m+n)\lceil \log_2(2\#\mathcal{L}-1) \rceil$ Boolean variables: the flipflop variable, the visible variables and associated beliefs variables, the observation variables, and the belief variables associated to the fully hidden and inferred hidden variables. Moreover, the dynamic of the resulting MDP is factored, and lot of transitions are deterministic, thanks to the flipflop variable trick. These simplifying structures are beneficial to the MDP solvers, leading to faster computations.

6 Conclusion

This paper describes a hydrid translation of a POMDP into a finite state space MDP one. The Qualitative Possibility Theory is used to maintain an epistemic state during the process: the belief space has a granulated representation, instead of a continuous one as in the classical translation. The resulting MDP is entirely defined computing transition and reward functions over these epistemic states. Definitions of these functions use respectively the pignistic transformation, used to recover a probability distribution from an epistemic state, and the Choquet integral with respect to the necessity, making the agent pessimistic about its ignorance. A practical way to implement this translation is then described: with these computation tricks, a factored POMDP leads to a factored and tractable MDP. This promising approach will be tested on the POMDPs of the IPPC competition [15] in a future work: provided problem descriptions are indeed in the form of the factored POMDPs introduced in Section 4.

References

1. Bellman, R.: A Markovian Decision Process. Indiana Univ. Math. J. 6, 679–684 (1957)
2. Boutilier, C., Poole, D.: Computing optimal policies for partially observable decision processes using compact representations. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference, AAAI 96, IAAI 96, Portland, Oregon, August 4-8, 1996, Volume 2. pp. 1168–1175 (1996), http://www.aaai.org/Library/AAAI/1996/aaai96-173.php
3. Cassandra, A., Littman, M.L., Zhang, N.L.: Incremental pruning: A simple, fast, exact method for partially observable markov decision processes. In: In Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence. pp. 54–61. Morgan Kaufmann Publishers (1997)
4. Cooman, G.D.: Integration and conditioning in numerical possibility theory. Ann. Math. Artif. Intell. 32(1-4), 87–123 (2001), http://dx.doi.org/10.1023/A:1016705331195

5. De Finetti, B.: Theory of probability: a critical introductory treatment. Wiley series in probability and mathematical statistics. Probability and mathematical statistics, Wiley (1974), http://books.google.fr/books?id=aRbvAAAAMAAJ
6. Drougard, N., Teichteil-Konigsbuch, F., Farges, J.L., Dubois, D.: Qualitative Possibilistic Mixed-Observable MDPs. In: Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13). pp. 192–201. AUAI Press, Corvallis, Oregon (2013)
7. Drougard, N., Teichteil-Königsbuch, F., Farges, J., Dubois, D.: Structured possibilistic planning using decision diagrams. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada. pp. 2257–2263 (2014), http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8553
8. Dubois, D.: Possibility theory and statistical reasoning. Computational Statistics and Data Analysis 51, 47–69 (2006)
9. Dubois, D., Prade, H.: The logical view of conditioning and its application to possibility and evidence theories. International Journal of Approximate Reasoning 4(1), 23 – 46 (1990), http://www.sciencedirect.com/science/article/pii/0888613X90900070
10. Dubois, D., Prade, H., Sandri, S.: On possibility/probability transformations. In: Proceedings of Fourth IFSA Conference. pp. 103–112. Kluwer Academic Publ (1993)
11. Ong, S.C.W., Png, S.W., Hsu, D., Lee, W.S.: Planning under uncertainty for robotic tasks with mixed observability. Int. J. Rob. Res. 29(8), 1053–1068 (Jul 2010)
12. Papadimitriou, C., Tsitsiklis, J.N.: The complexity of markov decision processes. Math. Oper. Res. 12(3), 441–450 (Aug 1987), http://dx.doi.org/10.1287/moor.12.3.441
13. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, Inc., New York, NY, USA, 1st edn. (1994)
14. Sabbadin, R.: A possibilistic model for qualitative sequential decision problems under uncertainty in partially observable environments. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. pp. 567–574. UAI'99, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1999), http://dl.acm.org/citation.cfm?id=2073796.2073860
15. Sanner, S.: Probabilistic track of the 2011 international planning competition. http://users.cecs.anu.edu.au/∼ssanner/IPPC_2011 (2011)
16. Silver, D., Veness, J.: Monte-carlo planning in large pomdps. In: Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A. (eds.) Advances in Neural Information Processing Systems 23, pp. 2164–2172. Curran Associates, Inc. (2010), http://papers.nips.cc/paper/4031-monte-carlo-planning-in-large-pomdps.pdf
17. Smallwood, R.D., Sondik, E.J.: The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon, vol. 21. INFORMS (1973)
18. Smith, T., Simmons, R.: Heuristic search value iteration for POMDPs. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. pp. 520–527. UAI '04, AUAI Press, Arlington, Virginia, United States (2004), http://dl.acm.org/citation.cfm?id=1036843.1036906
19. Zadeh, L.A.: Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. Soft Comput. 2(1), 23–25 (1998), http://dx.doi.org/10.1007/s005000050030