Structured Possibilistic Planning Using Decision Diagrams

Nicolas Drougard, Florent Teichteil-Königsbuch, Jean-Loup Farges, Didier Dubois

Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

Structured Possibilistic Planning Using Decision Diagrams

Nicolas Drougard, Florent Teichteil-Königsbuch
Jean-Loup Farges
Onera – The French Aerospace Lab
2 avenue Edouard Belin
31055 Toulouse Cedex 4, France

Didier Dubois
IRIT – Paul Sabatier University
118 route de Narbonne
31062 Toulouse Cedex 4, France

Abstract

Qualitative Possibilistic Mixed-Observable MDPs (π-MOMDPs), generalizing π-MDPs and π-POMDPs, are well-suited models to planning under uncertainty with mixed-observability when transition, observation and reward functions are not precisely known and can be qualitatively described. Functions defining the model as well as intermediate calculations are valued in a finite possibilistic scale $\mathcal{L}$, which induces a *finite* belief state space under partial observability contrary to its probabilistic counterpart. In this paper, we propose the first study of factored π-MOMDP models in order to solve large structured planning problems under qualitative uncertainty, or considered as qualitative approximations of probabilistic problems. Building upon the SPUDD algorithm for solving factored (probabilistic) MDPs, we conceived a symbolic algorithm named PPUDD for solving factored π-MOMDPs. Whereas SPUDD's decision diagrams' leaves may be as large as the state space since their values are real numbers aggregated through additions and multiplications, PPUDD's ones always remain in the finite scale $\mathcal{L}$ via $\min$ and $\max$ operations only. Our experiments show that PPUDD's computation time is much lower than SPUDD, Symbolic-HSVI and APPL for possibilistic and probabilistic versions of the same benchmarks under either total or mixed observability, while still providing high-quality policies.

Introduction

Many sequential decision-making problems under uncertainty can be easily expressed in terms of Markov Decision Processes (MDPs) (Bellman 1957). Partially Observable MDPs (POMDPs) (Smallwood and Sondik 1973) take into account situations where the system's state is not totally visible to the agent. Finally, the Mixed Observable MDP (MOMDP) framework (Ong et al. 2010; Araya-López et al. 2010) generalizes both previous ones by considering that states can be expressed in terms of two parts, one visible and the other hidden to the agent, which reduces the dimension of the infinite belief space. With regard to POMDPs, exact dynamic programming algorithms like incremental pruning (Cassandra, Littman, and Zhang 1997) can only solve very small problems: many approximation algorithms have been proposed to speed up computations while controlling the quality of the resulting policies (Pineau, Gordon, and Thrun 2003; Smith and Simmons 2004; Kurniawati, Hsu, and Lee 2008). In this paper, we proceed quite differently: we start with an approximated *qualitative* model that we exactly solve.

Qualitative possibilistic MDPs (π-MDPs), POMDPs (π-POMDPs) and more broadly MOMDPs (π-MOMDPs) were introduced in (Sabbadin 2001; Sabbadin, Fargier, and Lang 1998; Sabbadin 1999; Drougard et al. 2013). These possibilistic counterparts of probabilistic models are based on Possibility Theory (Dubois and Prade 1988) and more specifically on Qualitative Decision Theory (Dubois, Prade, and Sabbadin 2001; Dubois and Prade 1995). A possibility distribution, classically models imprecision or lack of knowledge about the model. Links between probabilities and possibilities are theoretically well-understood: possibilities and probabilities have similar behaviors for problems with low entropy probability distributions (Dubois et al. 2004). Using Possibility Theory instead of Probability Theory in MOMDPs necessarily leads to an approximation of the initial probabilistic model (Sabbadin 2000), where probabilities and rewards are replaced by qualitative statements that lie in a finite scale (as opposed to continuous ranges in the probabilistic framework), which results in simpler computations. Furthermore, in presence of partial observability, this approach benefits from computations on *finite* belief state spaces, whereas probabilistic MOMDPs tackle *infinite* ones. It means that the same algorithmic techniques can be used to solve π-MDPs, π-POMDPs or π-MOMDPs. What is lost in precision of the uncertainty model is saved in computational complexity.

Nevertheless, existing works on π-(MO)MDPs do not totally take advantage of the problem structure, i.e. visible or hidden parts of the state can be themselves factored into many state variables, which are flattened by current possibilistic approaches. In probabilistic settings, factored MDPs and Symbolic Dynamic Programming (SDP) frameworks (Boutilier, Dearden, and Goldszmidt 2000; Hoey et al. 1999) have been extensively studied in order to reason directly at the level of state variables rather than state space in extension. However, factored probabilistic MOMDPs have not yet been proposed to the best of our knowledge, probably because of the intricacy of reasoning with a mixture of a finite

state subspace and an infinite belief state subspace due to the probabilistic model – contrary to the possibilistic case where both subspaces are finite. The famous algorithm SPUDD (Hoey et al. 1999) solves factored probabilistic MDPs by using symbolic functional representations of value functions and policies in the form of Algebraic Decision Diagrams (ADDs) (Bahar et al. 1997), which compactly encode real-valued functions of Boolean variables: ADDs are directed acyclic graphs whose nodes represent state variables and leaves are the function's values. Instead of updating states individually at each iteration of the algorithm, states are aggregated within ADDs and operations are symbolically and directly performed on ADDs over many states at once. However, SPUDD suffers from manipulation of potentially huge ADDs in the worst case: for instance, expectation involves additions and multiplications of real values (probabilities and rewards), creating other values in-between, in such a way that the number of ADD leaves may equal the size of the state space, which is exponential in the number of state variables. Therefore, the work presented here is motivated by the simple observation that *symbolic operations with **possibilistic** MDPs would necessarily limit the size of ADDs*: indeed, this formalism operates over a *finite* possibilistic scale $\mathcal{L}$ with only $\max$ and $\min$ operations involved, which implies that all manipulated values remain in $\mathcal{L}$.

This paper begins with a presentation of the π-MOMDP framework. Then we present our first contribution: a Symbolic Dynamic Programming algorithm for solving factored π-MOMDPs named Possibilistic Planning Using Decision Diagram (PPUDD). This contribution alone is insufficient, since it relies on a *belief-state* variable whose number of values is exponential in the size of the state space. Therefore, our second contribution is a theorem to factorize the belief state itself in many variables under some assumptions about dependence relationships between state and observation variables of a π-MOMDP, which makes our algorithm more tractable while still exact and optimal. Finally, we experimentally assess our approach on possibilistic and probabilistic versions of the same benchmarks: PPUDD against SPUDD and APRICODD (St-aubin, Hoey, and Boutilier 2000) under total observability to demonstrate that generality of our approximate approach does not penalize performances on restrictive submodels; PPUDD against symbolic HSVI (Sim et al. 2008) and APPL (Kurniawati, Hsu, and Lee 2008; Ong et al. 2010) under mixed-observability.

Qualitative Possibilistic MOMDPs

Qualitative Possibilistic Mixed-Observable MDPs (π-MOMDPs) have been first formulated in (Drougard et al. 2013). Let us define $\mathcal{L} = \{0, \frac{1}{k}, \ldots, \frac{k-1}{k}, 1\}$, the fixed possibilistic scale ($k \in \mathbb{N}^*$). A *possibility distribution* over the state space $\mathcal{S}$ is a function $\pi : \mathcal{S} \to \mathcal{L}$ which verifies the possibilistic normalization: $\max_{s \in \mathcal{S}} \pi(s) = 1$. This distribution ranks plausibilities of events: $\pi(s) < \pi(s')$ means that s is less plausible than s'. A π-MOMDP is defined by a tuple $\langle \mathcal{S} = \mathcal{S}_v \times \mathcal{S}_h, \mathcal{A}, \mathcal{L}, T^\pi, \mathcal{O}, \Omega^\pi, \mu \rangle$ where:

- $\mathcal{S} = \mathcal{S}_v \times \mathcal{S}_h$ is a finite set of states composed of states in $\mathcal{S}_v$ visible to the agent and states in $\mathcal{S}_h$ hidden to it;
- $\mathcal{A}$ is a finite set of actions;
- $T^\pi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathcal{L}$ is a possibility transition function s.t. $T^\pi(s, a, s') = \pi(s' \mid s, a)$ is the *possibility degree* of reaching state s' when applying action a in state s;
- $\mathcal{O}$ is a finite set of observations;
- $\Omega^\pi : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \mapsto \mathcal{L}$ is an observation function s.t. $\Omega^\pi(s', a, o') = \pi(o' \mid s', a)$ is the *possibility degree* of observing o' when applying action a in state s';
- $\mu : \mathcal{S} \mapsto \mathcal{L}$ is a preference distribution that models *qualitative* agent's goals, i.e. $\mu(s) < \mu(s')$ means that s is less preferable than s'.

The influence diagram of a π-MOMDP is depicted in Figure 1. This framework includes totally and partially observable problems: π-MDPs are π-MOMDPs with $\mathcal{S} = \mathcal{S}_v$ (state is entirely visible to the agent); π-POMDPs are π-MOMDPs with $\mathcal{S} = \mathcal{S}_h$ (state is entirely hidden).

The state's hidden part may initially not be entirely known: an estimation can be however available, expressed in terms of a possibility distribution $\beta_0 : \mathcal{S}_h \to \mathcal{L}$ called *initial possibilistic belief state*. For instance, if $\forall s_h \in \mathcal{S}_h$, $\beta_0(s_h) = 1$, the initial hidden state is completely unknown; if $\exists \overline{s}_h \in \mathcal{S}_h$ such that $\forall s_h \in \mathcal{S}_h$, $\beta_0(s_h) = \delta_{\overline{s}_h, s_h}$ (Kronecker delta), the initial hidden state is known to be $\overline{s}_h$. Let us use the prime notation to label the symbols at the next time of the process (e.g. β' for the next belief) and the unprime one at the current time (e.g. a for the current action). By using the possibilistic version of Bayes' rule (Dubois and Prade 1990), the belief state's update under mixed-observability is (Drougard et al. 2013) $\forall s'_h \in \mathcal{S}_h$:

$$\beta'(s'_h) = \begin{cases} 1 \text{ if } s'_h \in \underset{s'_h \in \mathcal{S}_h}{\operatorname{argmax}} \pi(o', s'_v, s'_h \mid s_v, \beta, a) > 0 \\ \pi(o', s'_v, s'_h \mid s_v, \beta, a) \text{ otherwise,} \end{cases} \tag{1}$$

denoted by $\beta' = U(\beta, a, s_v, s'_v, o')$. The set of all beliefs over $\mathcal{S}_h$ is denoted by B^π. **Note that B^π is finite of size** $\#\mathcal{L}^{\#\mathcal{S}_h} - (\#\mathcal{L} - 1)^{\#\mathcal{S}_h}$ unlike continuous probabilistic belief spaces $B \subsetneq [0, 1]^{\mathcal{S}_h}$. It yields a belief finite-state π-MDP over $\mathcal{S}_v \times B^\pi$, named *state space accessible to the agent*, whose transitions are (Drougard et al. 2013):

$$\pi(s'_v, \beta' \mid s_v, \beta, a) = \max_{\substack{s'_h \in \mathcal{S}_h \\ o' \mid \beta' = U(\beta, a, s_v, s'_v, o')}} \pi(o', s'_v, s'_h \mid s_v, \beta, a).$$

Finally, preference over (s_v, β) is defined such that this paired state is considered as good if it is necessary (according to β) that the system is in a good state: $\mu(s_v, \beta) = \min_{s_h \in \mathcal{S}_h} \max\{\mu(s_v, s_h), 1 - \beta(s_h)\}$.

A *stationary policy* is defined as a function $\delta : \mathcal{S}_v \times B^\pi \to \mathcal{A}$ and the set of such policies is denoted by Δ. For $t \in \mathbb{N}$,

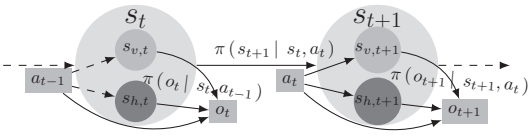


Figure 1: Dynamic influence diagram of a (π-)MOMDP

the set of t-length trajectories starting in state $x = (s_v, \beta) \in \mathcal{S}_v \times B^\pi$ by following policy δ is denoted by $\mathcal{T}_t^\delta(x)$. For a trajectory $\tau \in \mathcal{T}_t^\delta(x)$, $\tau(t')$ is the state visited at time step $t' \leqslant t$ and its quality is defined as the preference of its terminal state: $\mu(\tau) = \mu(\tau(t))$. The *value (or utility) function* of a policy δ in a state $x = (s_v, \beta) \in \mathcal{S}_v \times B^\pi$ is defined as the optimistic quality of trajectories starting in x:

$$V^\delta(x) = \max_{t=0}^{+\infty} \max_{\tau \in \mathcal{T}_t^{(\delta)}(x)} \min\{\pi(\tau \mid x_0 = x, \delta), \mu(\tau(t))\}$$

By replacing $\max$ by $\sum$ and $\min$ by $\times$, one can easily draw a parallel with probabilistic MDPs' expected criterion with terminal rewards. The optimal value function is defined as: $V^*(x) = \max_{\delta \in \Delta} V^\delta(x)\,,\ x = (s_v, \beta) \in \mathcal{S}_v \times B^\pi$. As proved in (Drougard et al. 2013), there exists an optimal stationary policy $\delta^* \in \Delta$, which is optimal over all history-dependent policies and independent from the initial state, which can be found by dynamic programming **if** there exists an action $\overline{a}$ such that $\pi(\, s'_v, \beta' \mid\, s_v, \beta, \overline{a}\,) = \delta_{(s_v,\beta),(s'_v,\beta')}$ (Kronecker delta). This assumption is satisfied if $\pi(\, s' \mid\, s, \overline{a}\,) = \delta_{s,s'}$ (state does not change) and $\pi(\, o' \mid\, s', \overline{a}\,) = 1\ \forall s', o'$ (agent does not observe). Action $\overline{a}$ is similar to the discount factor in probabilistic MOMDPs; it allows the following dynamic programming equation to converge in at most $\#\mathcal{S}_v \times \#B^\pi$ iterations to the optimal value function V^*:

$$V^*_{t+1}(x) = \max_{a \in \mathcal{A}} \max_{x' \in \mathcal{S}_v \times B^\pi} \min\{\pi(x' \mid x, a), V^*_t(x')\}, \quad (2)$$

with initialization $V^*_0(x) = \mu(x)$. This hypothesis is yet not a constraint in practice: in the returned optimal policy δ^*, action $\overline{a}$ is only used for goals whose preference degree is greater than possibility degree of transition to better goals. For a given action a and state s, we note $q^a(x) = \max_{x' \in \mathcal{S}_v \times B^\pi} \min\{\pi(x' \mid x, a), V^*_t(x')\}$ which is known as the *Q-value function*.

This framework does not consider s_v nor β to be themselves factored into variables, meaning that it does not tackle *factored* π-MOMDPs. In the next section, we present our first contribution: the first symbolic algorithm to solve factored possibilistic decision-making problems.

Solving factored π-MOMDPs using symbolic dynamic programming

Factored MDPs (Hoey et al. 1999) have been used to efficiently solve structured sequential decision problems under probabilistic uncertainty, by symbolically reasoning on functions of states via decision diagrams rather than on individual states. Inspired by this work this section sets up a symbolic resolution of factored π-MOMDPs, which assumes that $\mathcal{S}_v$, $\mathcal{S}_h$ and $\mathcal{O}$ are each cartesian products of variables. According to the previous section, it boils down to solving a finite-state belief π-MDP whose state space is in the form of $\mathcal{S}_{v,1} \times \cdots \times \mathcal{S}_{v,m} \times B^\pi$, where each of those state variable spaces is finite. We will see in the next section how B^π can be further factorized thanks to the factorization of $\mathcal{S}_h$ and $\mathcal{O}$. While probabilistic belief factorization in (Boyen and Koller 1999; Shani et al. 2008) is approximate,

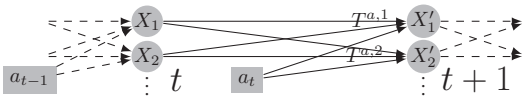


Figure 2: DBN of a factored π-MDP

the one presented here relies on some assumptions but is exact. For now, as finite state variable spaces of size K can be themselves factored into $\lceil \log_2 K \rceil$ binary-variable spaces (see (Hoey et al. 1999)), we can assume that we are reasoning about a factored belief-state π-MDP whose state space is $X = (X_1, \ldots, X_n), n \in \mathbb{N}^*$ and $\forall i, \#X_i = 2$.

Dynamic Bayesian Networks (DBNs) (Dean and Kanazawa 1989) are a useful graphical representation of process transitions, as depicted in Figure 2. In DBN semantics, $parents(X'_i)$ is the set of state variables on which X'_i depends. We assume that $parents(X'_i) \subset X$, but methods are discussed in the literature to circumvent this restrictive assumption (Boutilier 1997). In the possibilistic settings, this assumption allows us to compute the joint possibility transition as $\pi(\, s'_v, \beta' \mid\, s_v, \beta, a\,) = \pi(\, X' \mid\, X, a\,) = \min_{i=1}^n \pi(\, X'_i \mid\, parents(X'_i), a\,)$. Thus, a factored π-MOMDP can be defined with transition functions $T^{a,i} = \pi(\, X'_i \mid\, parents(X'_i), a\,)$ for each action a and variable X'_i. Each transition function can be compactly encoded in an Algebraic Decision Diagram (ADD) (Bahar et al. 1997). An ADD, as illustrated in Figure 3a, is a directed acyclic graph which compactly represents a real-valued function of binary variables, whose identical sub-graphs are merged and zero-valued leaves are not memorized. The possibilistic update of dynamic programming, *i.e.* Equation 2, can be rewritten in a symbolic form, so that states are now globally updated at once instead of individually ; the Q-value of an action $a \in \mathcal{A}$ can be decomposed into independent computations thanks to the following proposition:

Proposition 1. *Consider the current value function $V^*_t : \{0,1\}^p \to \mathcal{L}$. For a given action $a \in \mathcal{A}$, let us define:*
- $q^a_0 = V^*_t(X'_1, \cdots, X'_n)$,
- $q^a_i = \max_{X'_i \in \{0,1\}} \min\left\{(\, X'_i \mid\, parents(X'_i), a\,), q^a_{i-1}\right\}$,
Then, the possibilistic Q-value of action a is: $q^a = q^a_n$.

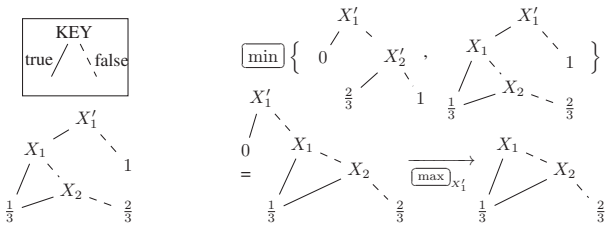
Proof.

$$\begin{aligned}
q^a &= \max_{(s'_v,\beta') \in \mathcal{S}_v \times B^\pi} \min\{\pi(s'_v, \beta' | s_v, \beta, a), V^*_t(s'_v, \beta')\} \\
&= \max_{X' \in \mathcal{S}_v \times B^\pi} \min\left\{\min_{i=1}^n \pi(\, X'_i \mid\, parents(X'_i), a\,), V^*_t(X')\right\} \\
&= \max_{X'_n \in \{0,1\}} \min\Big\{\pi(\, X'_n \mid\, parents(X'_n), a\,), \cdots \\
&\qquad \max_{X'_2 \in \{0,1\}} \min\{\pi(\, X'_2 \mid\, parents(X'_2), a\,), \\
&\qquad\quad \max_{X'_1 \in \{0,1\}} \min\{\pi(X'_1 \mid\, parents(X'_1), a), V^*_t(X')\}\} \cdots\Big\}
\end{aligned}$$

where the last equation is due to the fact that, for any variables $x, y \in \mathcal{X}, \mathcal{Y}$ finite spaces, and any functions $\varphi : \mathcal{X} \to \mathcal{L}$ and $\psi : \mathcal{Y} \to \mathcal{L}$, we have:

$$\max_{y \in \mathcal{Y}} \min\{\varphi(x), \psi(y)\} = \min\{\varphi(x), \max_{y \in \mathcal{Y}} \psi(y)\} \qquad \square$$

The Q-value of action a, represented as an ADD, can be then iteratively regressed over successive post-action state



(a) ADD encoding $T^{a,1}$ of Fig. 2

(b) Symbolic regression of the current Q-value ADD combined with the transition ADD of Figure 3a

Figure 3: Algebraic Decision Diagrams for PPUDD

variables $X'_i, 1 \leqslant i \leqslant n$. The following notations are used to make it explicit that we are working with symbolic functions encoded as ADDs:
- $\boxed{\min}\{f, g\}$ where f and g are 2 ADDs;
- $\boxed{\max}_{X_i} f = \boxed{\max}\{f^{X_i=0}, f^{X_i=1}\}$, which can be easily computed because ADDs are constructed on the basis of the Shannon expansion: $f = \overline{X_i} \cdot f^{X_i=0} + X_i \cdot f^{X_i=1}$ where $f^{X_i=1}$ and $f^{X_i=0}$ are sub-ADDs representing the positive and negative Shannon cofactors (see Fig. 3a).

Figure 3b illustrates the possibilistic regression of the Q-value of an action for the first state variable X_1 and leads to the intuition that ADDs should be far smaller in practice under possibilistic settings, since their leaves lie in $\mathcal{L}$ instead of $\mathbb{R}$, thus yielding more sub-graph simplifications.

Algorithm 1 is a symbolic version of the π-MOMDP Value Iteration Algorithm (Drougard et al. 2013), which relies on the regression scheme defined in Proposition 1. Inspired by SPUDD (Hoey et al. 1999), PPUDD means *Possibilistic Planning Using Decision Diagrams*. As for SPUDD, it needs to swap unprimed state variables to primed ones in the ADD encoding the current value function before computing the Q-value of an action a (see Line 5 of Algorithm 1 and Figure 3b). This operation is required to differentiate the next state represented by primed variables from the current one when operating on ADDs.

We mentioned at the beginning of this section that belief variable B^π could be transformed into $\lceil \log_2 K \rceil$ binary variables where $K = \#\mathcal{L}^{\#\mathcal{S}_h} - (\#\mathcal{L} - 1)^{\#\mathcal{S}_h}$. However, this K can be very large so we propose in the next section a method to exploit the factorization of $\mathcal{S}_h$ and $\mathcal{O}$ in order to factorize B^π itself into small belief subvariables, which will decompose the possibilistic transition ADD into an aggregation of smaller ADDs. Note that PPUDD can solve π-MOMDPs even if this belief factorization is not feasible, but it will manipulate bigger ADDs.

Algorithm 1: PPUDD

1 $V^* \leftarrow 0$; $V^c \leftarrow \mu$; $\delta \leftarrow \overline{a}$;
2 **while** $V^* \neq V^c$ **do**
3 $V^* \leftarrow V^c$;
4 **for** $a \in \mathcal{A}$ **do**
5 $q^a \leftarrow$ swap each X_i variable in V^* with X'_i ;
6 **for** $1 \leqslant i \leqslant n$ **do**
7 $q^a \leftarrow \boxed{\min}\{q^a, \pi(X'_i \mid parents(X'_i), a)\}$;
8 $q^a \leftarrow \boxed{\max}_{X'_i} q^a$;
9 $V^c \leftarrow \boxed{\max}\{q^a, V^c\}$;
10 update δ to a where $q^a = V^c$ and $V^c > V^*$;
11 **return** (V^*, δ) ;

π-MOMDP belief factorization

Factorizing the belief variable requires three structural assumptions on the π-MOMDP's DBN, which are illustrated by the Rocksample benchmark (Smith and Simmons 2004).

Motivating example. A rover navigating in a $N \times N$ grid has to collect scientific samples from interesting ("good") rocks among R ones and then to reach the exit. It is fitted with a noisy long-range sensor that can be used to determine if a rock is "good" or not:
- $\mathcal{S}_v$ consists of all the possible locations of the rover in addition to the exit ($\#\mathcal{S}_v = N^2 + 1$),
- $\mathcal{S}_h$ consists of all the possible natures of the rocks ($\mathcal{S}_h = \mathcal{S}_{h,1} \times \ldots \times \mathcal{S}_{h,R}$ with $\forall 1 \leqslant i \leqslant R, \mathcal{S}_{h,i} = \{good, bad\}$),
- $\mathcal{A}$ contains the (deterministic) moves in the 4 directions, checking rock i $\forall 1 \leqslant i \leqslant R$ and sampling the current rock,
- $\mathcal{O} = \{o_{good}, o_{bad}\}$ are the possible sensor's answers for the current rock.

The more the rover is close to the checked rock, the better it observes its nature. The rover gets the reward $+10$ (resp. -10) for each good (resp. bad) sampled rock, and $+10$ when it reaches the exit.

In the possibilistic model, the observation function is approximated using a critical distance $d > 0$ beyond which checking a rock is uninformative: $\pi(o'_i \mid s'_i, a, s_v) = 1$ $\forall o'_i \in \mathcal{O}_i$. The possibility degree of erroneous observation becomes zero if it stands at the checked rock, and lowest non zero possibility degree otherwise. Finally, as possibilistic semantics does not allow sums of rewards, an additional visible state variable $s_{v,2} \in \{1, \ldots, R\}$ which counts the number of checked rocks is introduced. Preference $\mu(s)$ equals qualitative dislike of sampling $\frac{R+2-s_{v,2}}{R+2}$ if all rocks are bad and location is terminal, zero otherwise. The location of the rover is finally denoted by $s_{v,1} \in \mathcal{S}_{v,1}$ and the visible state is then $s_v = (s_{v,1}, s_{v,2}) \in \mathcal{S}_{v,1} \times \mathcal{S}_{v,2} = \mathcal{S}_v$.

Observations $\{o_{good}, o_{bad}\}$ for the current rock can be equivalently modeled as a cartesian product of observations $\{o_{good_1}, o_{bad_1}\} \times \cdots \times \{o_{good_R}, o_{bad_R}\}$ for each rock. By using this equivalent modeling, state and observation spaces are both respectively factored as $\mathcal{S}_{v,1} \times \ldots \times \mathcal{S}_{v,m} \times \mathcal{S}_{h,1} \times \ldots \times \mathcal{S}_{h,l}$ and $\mathcal{O} = \mathcal{O}_1 \times \ldots \times \mathcal{O}_l$, and we can now map each observation variable $o_j \in \mathcal{O}_j$ to its hidden state variable $s_{h,j} \in \mathcal{S}_{h,j}$. It allows us to reason about DBNs in the form of Figure 4, which expresses three important assumptions that will help us factorize the belief state itself:
1. all state variables $s_{v,1}, s_{v,2}, \ldots, s_{h,1}, s_{h,2}, \ldots$ are independent post-action variables (no arrow between two state variables at the same time step, e.g. $s_{v,2}$ and $s_{h,1}$);
2. a hidden variable does not depend on previous other hid-

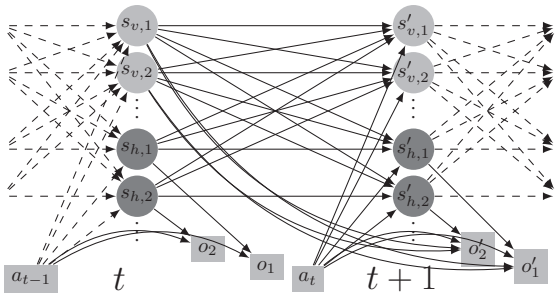


Figure 4: DBN of a factored belief-independent π-MOMDP

den variables: the nature of a rock is independent from the previous nature of other rocks (e.g. no arrow from $s_{h,1}$ to $s'_{h,2}$);
3. an observation variable is available for each hidden state variable. It does not depend on other hidden state variables nor current visible ones, but on previous visible state variables and action (e.g. no arrow between $s'_{h,1}$ and o'_2, nor between $s'_{v,1}$ and o'_1). Each observation variable is indeed only related to the nature of the corresponding rock.

Formalization. To formally demonstrate how the three previous independence assumptions can be used to factorize B^π, let us recursively define the history $(h_t)_{t\geqslant 0}$ of a π-MOMDP as: $h_0 = \{\beta_0, s_{v,0}\}$ and for each time step $t \geqslant 1$, $h_t = \{o_t, s_{v,t}, a_{t-1}, h_{t-1}\}$. We first prove in the next theorem that the current belief can be decomposed into marginal beliefs dependent on history h_t via the $\min$ aggregation:

Theorem 1. *If $s_{h,1}, \ldots, s_{h,l}$ are initially independent, then at each time step $t > 0$ the belief over hidden states can be written as $\beta_t = \min_{j=1}^{l} \beta_{j,t}$ with $\forall s_{h,j} \in \mathcal{S}_{h,j}$, $\beta_{j,t}(s_{h,j}) = \pi(s_{h,j} \mid h_t)$ the belief over $\mathcal{S}_{h,j}$.*

Proof. First $s_{h,1}, \ldots, s_{h,l}$ are initially independent, then $\exists (\beta_{0,j})_{j=1}^{l}$ such that $\beta_0(s_h) = \min_{j=1}^{l} \beta_{0,j}(s_{h,j})$. The independence between hidden variables conditioned on the history can be shown using the *d-separation* relationship (Pearl 1988) used for example in (Witwicki et al. 2013). In fact, as shown in Figure 4, given $1 \leqslant i < j \leqslant l$, $s'_{h,i}$ and $s'_{h,j}$ are d-separated by the evidence h_{t+1} recursively represented by the light-gray nodes. Thus $\pi(s'_h \mid h_{t+1}) = \min_{j=1}^{l} \pi\left(s'_{h,j} \mid h_{t+1}\right)$ *i.e.* $\beta_t(s'_h) = \min_{j=1}^{l} \beta_{j,t}(s'_{h,j})$. Note however that it would not be true if the same observation variable o would have concerned two different hidden state variables $s_{h,p}$ and $s_{h,q}$: as o is part of the history, there would be a convergent (towards o) relationship between $s_{h,p}$ and $s_{h,q}$ and the hidden state variable would have been dependent (because d-connected) conditioned on history. Moreover if hidden state variable $s'_{h,p}$ could depend on previous hidden state variable $s_{h,q}$, then $s'_{h,p}$ and $s'_{h,q}$ would have been dependent conditioned on history because d-connected through $s_{h,q}$. □

Thanks to the previous theorem, the state space accessible to the agent can now be rewritten as $\mathcal{S}_{v,1} \times \ldots \times \mathcal{S}_{v,m} \times B_1^\pi \times \cdots \times B_l^\pi$ with $B_j^\pi \subsetneq \mathcal{L}^{\mathcal{S}_{h,j}}$. The size of B_j^π is $\#\mathcal{L}^{\#\mathcal{S}_{h,j}} - (\#\mathcal{L}-1)^{\#\mathcal{S}_{h,j}}$. If all state variables are binary, $\#B_j^\pi = 2\#\mathcal{L} - 1$ for all $1 \leqslant i \leqslant l$, so that $\#\mathcal{S}_v \times B^\pi = 2^m(2\#\mathcal{L}-1)^l$: contrary to probabilistic settings, **hidden state variables and visible ones have a similar impact on the solving complexity**, i.e. both singly-exponential in the number of state variables. In the general case, by noting $\kappa = \max\{\max_{1\leqslant i\leqslant m} \#S_{v,i}, \max_{1\leqslant j\leqslant l} \#S_{h,j}\}$, there are $\mathcal{O}(\kappa^m(\#\mathcal{L})^{(\kappa-1)l})$ flattened belief states, which is indeed exponential in the arity of state variables too.

It remains to prove that $s_{v,1}, \ldots, s_{v,m}, \beta_1, \ldots, \beta_l$ are independent post-action variables. This result is based on Lemma 1, which shows how marginal beliefs are actually updated. For this purpose, we recursively define the history concerning hidden variable $s_{h,j}$: $h_{j,0} = \{\beta_{j,0}\}$ and $\forall t \geqslant 0$, $h_{j,t+1} = \{o_{j,t+1}, s_{v,t}, a_t, h_{j,t}\}$. We note $\pi\left(o'_j, s'_{h,j} \,\middle|\, s_v, \beta_j, a\right) = \max_{s_{h,j}} \min\{\pi\left(o'_j \middle| s'_{h,j}, s_v, a\right), \pi\left(s'_{h,j} \middle| s_v, s_{h,j}, a\right), \beta_j(s_{h,j})\}$:

Lemma 1. *If the agent is at time t in visible state s_v, with a belief over j^{th} hidden state $\beta_{j,t}$, executes action a and then gets observation o'_j, the update of the belief state over $\mathcal{S}_{h,j}$ is: $\beta_{j,t+1}(s'_{h,j})$*

$$= \begin{cases} 1 \text{ if } s'_{h,j} \in \underset{s'_{h,j}\in\mathcal{S}_{h,j}}{\operatorname{argmax}} \pi\left(o'_j, s'_{h,j} \,\middle|\, s_v, \beta_{j,t}, a\right) > 0 \\ \pi\left(o'_j, s'_{h,j} \,\middle|\, s_v, \beta_{j,t}, a\right) \text{ otherwise.} \end{cases} \quad (3)$$

Proof. First note that $s_{h,j}$ and $\{o_{m,s}\}_{s\leqslant t, m\neq j} \cup \{s_{v,t}\}$ are d-separated by $h_{j,t}$ then $s_{h,j}$ is independent on $\{o_{m,s}\}_{s\leqslant t, m\neq j} \cup \{s_{v,t}\}$ conditioned on $h_{j,t}$: $\pi(s_{h,j} \mid h_t) = \pi(s_{h,j} \mid h_{j,t})$. Then, possibilistic Bayes' rule as in Equation 1 yields the intended result. □

Finally, Theorem 2 relies on Lemma 1 to ensure independence of all post-action state variables of the belief π-MDP, which allows us to write the possibilistic transition function of the belief-state π-MDP in a factored form:

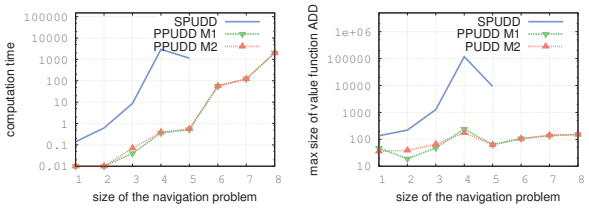
Theorem 2. $\forall \beta, \beta' \in B$, $\forall s_v, s'_v \in \mathcal{S}_v$, $\forall a \in \mathcal{A}$,
$\pi(s'_v, \beta' \mid s_v, \beta, a)$

$$= \min\left\{ \min_{i=1}^{m} \pi\left(s'_{v,i} \,\middle|\, s_v, \beta, a\right), \min_{j=1}^{l} \pi\left(\beta'_j \,\middle|\, s_v, \beta_j, a\right) \right\}$$

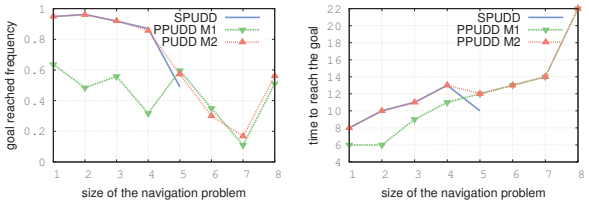
Proof. Observation variables are independent given the past (d-separation again). Moreover, we proved in Lemma 1 that updates of each marginal belief can be performed independently on other marginal beliefs, but depends on the corresponding observation only. Thus, we conclude that the marginal belief state variables are independent given the past. Finally as s'_v and o' are independent given the past,
$\pi(s'_v, \beta' \mid s_v, \beta, a) = \max_{o' \mid \beta' = U(\beta, a, s_v, o')} \pi(s'_v, o' \mid s_v, \beta, a)$

$$= \min\left\{ \pi(s'_v \mid s_v, \beta, a), \max_{o' \mid \beta' = U(\beta, a, s_v, o')} \pi(o' \mid s_v, \beta, a) \right\}$$

$= \min\{\pi(s'_v \mid s_v, \beta, a), \pi(\beta' \mid s_v, \beta, a)\}$ which concludes the proof. □

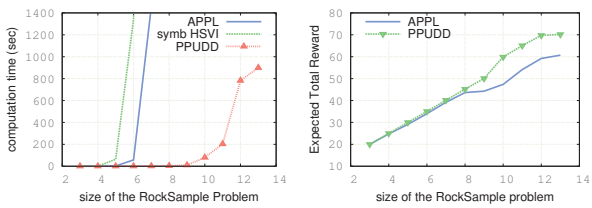


(a) Computation time (b) Size of ADD value function



(c) Goal reached frequency (d) Time to reach the goal

Figure 5: PPUDD vs. SPUDD on the navigation domain



(a) Computation time (b) Expected total reward

Figure 6: PPUDD vs. APPL and symb HSVI (RS)

Experimental results

In this section, we compare our approach against probabilistic solvers in order to answer the following question: what is the efficacy/quality tradeoff achieved by reasoning about an approximate model but with an exact efficient algorithm? Despite radically different methods, possibilistic policies and probabilistic ones are both represented as ADDs that are directly comparable and statistically evaluated under identical settings *i.e.* transition and reward functions defined by the probabilistic model.

We first assessed PPUDD performances on totally observable factored problems since PPUDD is also the first algorithm to solve factored π-MDPs (by inclusion in π-MOMDPs). To this end, we compared PPUDD against SPUDD on the **navigation domain** used in planning competitions (Sanner 2011). In this domain, a robot navigates in a grid where it must reach some goal location most reliably. It can apply actions going north, east, south, west and stay which all cost 1 except on the goal. When moving, it can suddenly disappear with some probability defined as a Bernoulli distribution. This probabilistic model is approximated by two possibilistic ones where: the preference of reaching the goal is 1; in the first model (M1) the highest probability of each Bernoulli distribution is replaced by 1 (for possibility normalization reasons) and the same value for the lowest probability is kept; for the second model (M2), the probability of disappearing is replaced by 1 and the other one is kept. Figure 5a shows that SPUDD runs out of memory from the 6^{th} problem, and PPUDD computation's time outperforms SPUDD's one by many orders of magnitude for the two models. Intuitively, this result comes from the fact that PPUDD's ADDs should be smaller because their leaves' values are in the finite scale $\mathcal{L}$ rather than $\mathbb{R}$, which is indeed demonstrated in Figure 5b. Performances were evaluated with two relevant criteria: frequency of runs where the policy reaches the goal (see Figure 5c), and average length of execution runs that reach the goal (see Figure 5d), that are both functions of the problem's instance. As expected, model (M2) is more cautious than model (M1) and gets a better goal-reached frequency (similar to SPUDD's one for the instances it can solve). The later is more optimistic and gets a better average length of execution runs than model (M2) due to its dangerous behavior. For fairness reasons, we also compared ourselves against APRICODD, which is an approximate algorithm for factored MDPs: parameters impacting the approximation are hard to tune (either huge computation times, or zero qualities) and it is largely outperformed by PPUDD in both time and quality whatever the parameters (curves are not shown since uninformative).

Finally, we compared PPUDD on the **Rocksample problem** (RS) against a recent probabilistic MOMDP planner, APPL (Ong et al. 2010), and a POMDP planner using ADDs, symbolic HSVI (Sim et al. 2008). Both algorithms are approximate and anytime, so we decided to stop them when they reach a precision of 1. Figure 6a, where problem instances increase with grid size and number of rocks, shows that APPL runs out of memory at the 8^{th} problem instance, symbolic HSVI at the 7^{th} one, while PPUDD outperforms them by many orders of magnitude.

Instead of precision, computation time of APPL can be fixed at PPUDD's computation time in order to compare their expected total rewards after they consumed the same CPU time. Surprisingly, Figure 6b shows that rewards gathered are higher with PPUDD than with APPL. The reason is that APPL is in fact an approximate probabilistic planner, which shows that our approach consisting in exactly solving an approximate model can outperform algorithms that approximately solve an exact model.

Conclusion

We presented PPUDD, the first algorithm to the best of our knowledge that solves factored possibilistic (MO)MDPs with symbolic calculations. In our opinion, possibilistic models are a good tradeoff between non-deterministic ones, whose uncertainties are not at all quantified yielding a very approximate model, and probabilistic ones, where uncertainties are fully specified. Moreover, π-MOMDPs reason about finite values in a qualitative scale $\mathcal{L}$ whereas probabilistic MOMDPs deal with values in $\mathbb{R}$, which implies larger ADDs for symbolic algorithms. Also, the former reduce to finite-state belief π-MDPs contrary to the latter that yield *continuous*-state belief MDPs of significantly higher complexity. Our experimental results highlight that using an exact algorithm (PPUDD) for an approximate model (π-MDPs) can bring significantly faster computations than reasoning about exact models, while providing better policies than approxi-

mate algorithms (APPL) for exact models. In the future, we would like to generalize our possibilistic belief factorization theory to probabilistic settings.

References

Araya-López, M.; Thomas, V.; Buffet, O.; and Charpillet, F. 2010. A closer look at MOMDPs. In *Proceedings of the Twenty-Second IEEE International Conference on Tools with Artificial Intelligence (ICTAI-10)*.

Bahar, R. I.; Frohm, E. A.; Gaona, C. M.; Hachtel, G. D.; Macii, E.; Pardo, A.; and Somenzi, F. 1997. Algebric decision diagrams and their applications. *Form. Methods Syst. Des.* 10(2-3):171–206.

Bellman, R. 1957. A Markovian Decision Process. *Indiana Univ. Math. J.* 6:679–684.

Boutilier, C.; Dearden, R.; and Goldszmidt, M. 2000. Stochastic dynamic programming with factored representations. *Artif. Intell.* 121(1-2):49–107.

Boutilier, C. 1997. Correlated action effects in decision theoretic regression. In *UAI*, 30–37.

Boyen, X., and Koller, D. 1999. Exploiting the architecture of dynamic systems. In Hendler, J., and Subramanian, D., eds., *AAAI/IAAI*, 313–320. AAAI Press / The MIT Press.

Cassandra, A.; Littman, M. L.; and Zhang, N. L. 1997. 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*, 54–61. Morgan Kaufmann Publishers.

Dean, T., and Kanazawa, K. 1989. A model for reasoning about persistence and causation. *Comput. Intell.* 5(3):142–150.

Drougard, N.; Teichteil-Konigsbuch, F.; Farges, J.-L.; and Dubois, D. 2013. Qualitative Possibilistic Mixed-Observable MDPs. In *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, 192–201. Corvallis, Oregon: AUAI Press.

Dubois, D., and Prade, H. 1988. *Possibility Theory: An Approach to Computerized Processing of Uncertainty (traduction revue et augmentée de "Théorie des Possibilités")*. New York: Plenum Press.

Dubois, D., and Prade, H. 1990. The logical view of conditioning and its application to possibility and evidence theories. *International Journal of Approximate Reasoning* 4(1):23 – 46.

Dubois, D., and Prade, H. 1995. Possibility theory as a basis for qualitative decision theory. In *IJCAI*, 1924–1930. Morgan Kaufmann.

Dubois, D.; Foulloy, L.; Mauris, G.; and Prade, H. 2004. Probability-possibility transformations, triangular fuzzy sets and probabilistic inequalities. *Reliable Computing* 10:2004.

Dubois, D.; Prade, H.; and Sabbadin, R. 2001. Decision-theoretic foundations of qualitative possibility theory. *European Journal of Operational Research* 128(3):459–478.

Hoey, J.; St-aubin, R.; Hu, A.; and Boutilier, C. 1999. Spudd: Stochastic planning using decision diagrams. In *In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 279–288. Morgan Kaufmann.

Kurniawati, H.; Hsu, D.; and Lee, W. S. 2008. SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Proceedings of Robotics: Science and Systems IV*.

Ong, S. C. W.; Png, S. W.; Hsu, D.; and Lee, W. S. 2010. Planning under uncertainty for robotic tasks with mixed observability. *Int. J. Rob. Res.* 29(8):1053–1068.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pineau, J.; Gordon, G.; and Thrun, S. 2003. Point-based value iteration: An anytime algorithm for pomdps. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1025 – 1032.

Sabbadin, R.; Fargier, H.; and Lang, J. 1998. Towards qualitative approaches to multi-stage decision making. *Int. J. Approx. Reasoning* 19(3-4):441–471.

Sabbadin, R. 1999. 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*, UAI'99, 567–574. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Sabbadin, R. 2000. Empirical comparison of probabilistic and possibilistic markov decision processes algorithms. In Horn, W., ed., *ECAI*, 586–590. IOS Press.

Sabbadin, R. 2001. Possibilistic markov decision processes. *Engineering Applications of Artificial Intelligence* 14(3):287 – 300. Soft Computing for Planning and Scheduling.

Sanner, S. 2011. Probabilistic track of the 2011 international planning competition. http://users.cecs.anu.edu.au/∼ssanner/IPPC_2011.

Shani, G.; Poupart, P.; Brafman, R. I.; and Shimony, S. E. 2008. Efficient add operations for point-based algorithms. In Rintanen, J.; Nebel, B.; Beck, J. C.; and Hansen, E. A., eds., *ICAPS*, 330–337. AAAI.

Sim, H. S.; Kim, K.-E.; Kim, J. H.; Chang, D.-S.; and Koo, M.-W. 2008. Symbolic heuristic search value iteration for factored pomdps. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, 1088–1093. AAAI Press.

Smallwood, R. D., and Sondik, E. J. 1973. *The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon*, volume 21. INFORMS.

Smith, T., and Simmons, R. 2004. Heuristic search value iteration for pomdps. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, UAI '04, 520–527. Arlington, Virginia, United States: AUAI Press.

St-aubin, R.; Hoey, J.; and Boutilier, C. 2000. Apricodd: Approximate policy construction using decision diagrams. In *In Proceedings of Conference on Neural Information Processing Systems*, 1089–1095.

Witwicki, S. J.; Melo, F. S.; Capitan, J.; and Spaan, M. T. J. 2013. A flexible approach to modeling unpredictable events in mdps. In Borrajo, D.; Kambhampati, S.; Oddi, A.; and Fratini, S., eds., *ICAPS*. AAAI.