

# *Bridging Commonsense Reasoning and Probabilistic Planning via a Probabilistic Action Language*

Yi Wang\*, Shiqi Zhang#, Joohyung Lee\*

\*Arizona State University, USA # SUNY Binghamton, USA

*submitted 1 January 2003; revised 1 January 2003; accepted 1 January 2003*

## Abstract

To be responsive to dynamically changing real-world environments, an intelligent agent needs to perform complex sequential decision-making tasks that are often guided by commonsense knowledge. The previous work on this line of research led to the framework called *interleaved commonsense reasoning and probabilistic planning* (iCORPP), which used P-log for representing commonsense knowledge and Markov Decision Processes (MDPs) or Partially Observable MDPs (POMDPs) for planning under uncertainty. A main limitation of iCORPP is that its implementation requires non-trivial engineering efforts to bridge the commonsense reasoning and probabilistic planning formalisms. In this paper, we present a unified framework to integrate iCORPP's reasoning and planning components. In particular, we extend probabilistic action language  $p\mathcal{BC}+$  to express utility, belief states, and observation as in POMDP models. Inheriting the advantages of action languages, the new action language provides an elaboration tolerant representation of POMDP that reflects commonsense knowledge. The idea led to the design of the system PBCPLUS2POMDP, which compiles a  $p\mathcal{BC}+$  action description into a POMDP model that can be directly processed by off-the-shelf POMDP solvers to compute an optimal policy of the  $p\mathcal{BC}+$  action description. Our experiments show that it retains the advantages of iCORPP while avoiding the manual efforts in bridging the commonsense reasoner and the probabilistic planner.

(The article is under consideration for acceptance in TPLP.)

## 1 Introduction

Intelligent agents frequently need to perform complex sequential decision making toward achieving goals that require more than one action, in which the agent's utility depends on a sequence of decisions. A common task is to find the policy that maximizes the agent's utility when the environment is partially observable, i.e., the agent knows only partial information about the current state. Partially Observable Markov Decision Processes (POMDPs) (Kaelbling et al. 1998) have been widely used for that purpose. It assumes partial observability of underlying states and can model nondeterministic state transitions and local, unreliable observations using probabilities, and plan toward maximizing long-term rewards under such uncertainties. However, as a very general mathematical framework, POMDPs are not equipped with built-in constructs for representing commonsense knowledge.

Recent works (Zhang and Stone 2015; Zhang et al. 2015) aim at embracing commonsense knowledge into probabilistic planning. In that line of research, a reasoner was used for state estimation with contextual knowledge, and a planner focuses on selecting actions to maximize the long-term reward. More recently, probabilistic logical knowledge has been used for reasoning about both the current state and the dynamics of the world, resulting in the framework called iCORPP (Zhang et al. 2017). iCORPP builds on two formalisms: P-log (Baral et al. 2009) for commonsense reasoning and POMDP (Kaelbling et al. 1998) for probabilistic planning. Reflecting

the commonsense knowledge, iCORPP significantly reduces the complexity of POMDP planning while enabling robot behaviors to adapt to exogenous changes. One example domain in (Zhang et al. 2017) demonstrates that the MDP constructed by iCORPP includes only 60 states whereas the naive way of enumerating all combinations of attribute values produces more than  $2^{69}$  states.

Despite the advantages, iCORPP has the limitation that practitioners must spend non-trivial engineering efforts to bridge the gap between P-log and POMDP in its implementations. One reason is that P-log does not have the built-in notions of utility and partially observable states as in POMDP models. Thus, the work on iCORPP acquired the transitions and their probabilities by running a P-log solver, but then the user has to manually add the information about the rewards and the belief states (Zhang et al. 2017).

In this paper, we present a more principled way to integrate the commonsense reasoning and probabilistic planning components in the iCORPP framework, which serves as the main contribution of this paper. We achieve this by extending probabilistic action language  $p\mathcal{BC}+$  (Lee and Wang 2018; Wang and Lee 2019) to support the representation of and reasoning with utility, belief states, and observation as in POMDP models. Inheriting the advantages of action languages, the new action language provides an elaboration tolerant representation of POMDP that is convenient to encode commonsense knowledge and completely shield users from the syntax or algorithms of POMDPs.

The second contribution is on the design of the system PBCPLUS2POMDP, which can dynamically construct POMDP models given an action description in  $p\mathcal{BC}+$ , and compute action policies using off-the-shelf POMDP solvers. Unlike iCORPP, the semantics of  $p\mathcal{BC}+$  and its reasoning system together support the direct generation of planning models, which can be further used for computing action policies using POMDP solvers. Experimental results show that the extended  $p\mathcal{BC}+$  (and its supporting system) retains the advantages of iCORPP while successfully avoiding the manual efforts in bridging the gap between iCORPP’s commonsense reasoning and probabilistic planning components.

The paper is organized as follows. After reviewing  $p\mathcal{BC}+$  and POMDP in Section 2, we extend  $p\mathcal{BC}+$  and show how it can be used to represent POMDP models in Section 3. In Section 4, we show how we can dynamically generate POMDP models by exploiting the elaboration tolerant representation of  $p\mathcal{BC}+$ . We present the system PBCPLUS2POMDP in Section 5 and experimental results with the system in Section 6. After discussing the related work in Section 7, we conclude in Section 8.

## 2 Preliminaries

Due to the space limit, the review is brief. For more detailed reviews, we refer the reader to (Lee and Wang 2018; Wang and Lee 2019), or the supplementary material corresponding to this paper at the TPLP archives.

### 2.1 Review: $p\mathcal{BC}+$ with Utility

We review  $p\mathcal{BC}+$  as presented in (Wang and Lee 2019), which extends the language in (Lee and Wang 2018) by incorporating the concept of utility.

Like its predecessors  $\mathcal{BC}$  (Lee et al. 2013) and  $\mathcal{BC}+$  (Babb and Lee 2015), language  $p\mathcal{BC}+$  assumes that a propositional signature  $\sigma$  is constructed from “constants” and their “values.” A *constant*  $c$  is a symbol that is associated with a finite set  $Dom(c)$ , called the *domain*. The signature

Causal Laws	Syntax	Translation into $LP^{MLN}$
static law	<b>caused</b> $F$ if $G$ where $F$ and $G$ are fluent formulas	$i : F \leftarrow i : G$ ( $i \in \{0, \dots, m\}$ )
fluent dynamic law	<b>caused</b> $F$ if $G$ after $H$ where $F$ and $G$ are fluent formulas, $H$ is a formula, $F$ does not contain statically determined constants and $H$ does not contain initpf constants	$i+1 : F \leftarrow (i+1 : G) \wedge (i : H)$ ( $i \in \{0, \dots, m-1\}$ )
pf constant declaration	<b>caused</b> $c = \{v_1 : p_1, \dots, v_n : p_n\}$ where $c$ is a pf constant with domain $\{v_1, \dots, v_n\}$ , $0 < p_i < 1$ for each $i \in \{1, \dots, n\}$ and $\sum_{i \in \{1, \dots, n\}} p_i = 1$	For each $j \in \{1, \dots, n\}$ : $ln(p_i) : (i : c) = v_j$ ( $i \in \{0, \dots, m-1\}$ )
utility law	<b>reward</b> $v$ if $F$ after $G$ where $v$ is a real number, $F$ is a fluent formula, $G$ contains fluent constant and action constant only	$\alpha : \text{utility}(v, i+1, id)$ $\leftarrow (i+1 : F) \wedge (i : G)$ where $id$ is a unique identifier
initpf constant declaration	<b>caused</b> $c = \{v_1 : p_1, \dots, v_n : p_n\}$ where $c$ is a initpf constant with domain $\{v_1, \dots, v_n\}$ , $0 < p_i < 1$ for each $i \in \{1, \dots, n\}$ and $\sum_{i \in \{1, \dots, n\}} p_i = 1$	For each $j \in \{1, \dots, n\}$ : $ln(p_i) : (0 : c) = v_j$
initial static law	<b>initially</b> $F$ if $G$ where $F$ is a fluent constant and $G$ is a formula that contains neither action constants nor pf constants	$\perp \leftarrow \neg(0 : F) \wedge 0 : G$

Fig. 1. Causal laws in  $p\mathcal{BC}+$  and their translations into  $LP^{MLN}$

$\sigma$  is constructed from a finite set of constants, consisting of atoms  $c = v$  for every constant  $c$  and every element  $v$  in  $Dom(c)$ . If the domain of  $c$  is  $\{\text{FALSE}, \text{TRUE}\}$ , then we say that  $c$  is *Boolean*, and abbreviate  $c = \text{TRUE}$  as  $c$  and  $c = \text{FALSE}$  as  $\sim c$ .

There are four types of constants in  $p\mathcal{BC}+$ : *fluent constants*, *action constants*, *pf (probability fact) constants* and *initpf (initial probability fact) constants*. Fluent constants are further divided into *regular* and *statically determined*. The domain of every action constant is restricted to Boolean. An *action description* is a finite set of *causal laws*, which describes how fluents depend on each other statically and how their values change from one time step to another. Fig. 1 lists causal laws in  $p\mathcal{BC}+$  and their translations into  $LP^{MLN}$  (Lee and Wang 2016). A *fluent formula* is a formula such that all constants occurring in it are fluent constants.

We use  $\sigma^{fl}$  ( $\sigma^{act}$ ,  $\sigma^{pf}$ , and  $\sigma^{initpf}$ , respectively) to denote the set of all atoms  $c = v$  where  $c$  is a fluent constant (action constant, pf constant, initpf constant, respectively) of  $\sigma$  and  $v$  is in  $Dom(c)$ . For any maximum time step  $m$ , any subset  $\sigma'$  of  $\sigma$  and any  $i \in \{0, \dots, m\}$ , we use  $i : \sigma'$  to denote the set  $\{i : A \mid A \in \sigma'\}$ . For any formula  $F$  of signature  $\sigma$ , by  $i : F$  we denote the result of inserting  $i :$  in front of every occurrence of every constant in  $F$ .

The semantics of a  $p\mathcal{BC}+$  action description  $D$  is defined by a translation into an  $LP^{MLN}$  program  $Tr(D, m) = D_{init} \cup D_m$ . Below we describe the essential part of the translation that turns a  $p\mathcal{BC}+$  description into an  $LP^{MLN}$  program.

The signature  $\sigma_m$  of  $D_m$  consists of atoms of the form  $i : c = v$  such that

- for each fluent constant  $c$  of  $D$ ,  $i \in \{0, \dots, m\}$  and  $v \in Dom(c)$ ,
- for each action constant or pf constant  $c$  of  $D$ ,  $i \in \{0, \dots, m-1\}$  and  $v \in Dom(c)$ .

and atoms of the form  $\text{utility}(v, i, id)$  introduced by each utility law as described in Fig. 1.

$D_m$  contains  $LP^{MLN}$  rules obtained from static laws, fluent dynamic laws, utility laws, and pf constant declarations as described in the third column of Fig. 1, as well as  $\{0 : c = v\}^{ch}$  for every regular fluent constant  $c$  and every  $v \in Dom(c)$ , and  $\{i : c = \text{TRUE}\}^{ch}$ ,  $\{i : c = \text{FALSE}\}^{ch}$  ( $i \in \{0, \dots, m-1\}$ ) for every action constant  $c$  to state that the fluents at time 0 and the actions

at each time are exogenous.<sup>1</sup>  $D_{init}$  contains  $\text{LP}^{\text{MLN}}$  rules obtained from initial static laws and initpf constant declarations as described in the third column of Fig. 1. Both  $D_m$  and  $D_{init}$  also contain constraints asserting that each constant is mapped to exactly one value in its domain. We identify an interpretation of  $\sigma_m$  (or  $\sigma$ ) that satisfies these constraints with the value assignment function mapping each constant to its value.

For any  $\text{LP}^{\text{MLN}}$  program  $\Pi$  of signature  $\sigma_1$  and any interpretation  $I$  of a subset  $\sigma_2$  of  $\sigma_1$ , we say  $I$  is a *residual (probabilistic) stable model* of  $\Pi$  if there exists an interpretation  $J$  of  $\sigma_1 \setminus \sigma_2$  such that  $I \cup J$  is a (probabilistic) stable model of  $\Pi$ .

For any interpretation  $I$  of  $\sigma$ , by  $i: I$  we denote the interpretation of  $i: \sigma$  such that  $i: I \models (i: c) = v$  iff  $I \models c = v$ . For  $x \in \{act, fl, pf\}$ , we use  $\sigma_m^x$  to denote the subset of  $\sigma_m$ , which is  $\{i: c = v \in \sigma_m \mid c = v \in \sigma^x\}$ .

A *state* of  $D$  is an interpretation  $I^{fl}$  of  $\sigma^{fl}$  such that  $0: I^{fl}$  is a residual (probabilistic) stable model of  $D_0$ . A *transition* of  $D$  is a triple  $\langle s, e, s' \rangle$  where  $s$  and  $s'$  are interpretations of  $\sigma^{fl}$  and  $e$  is an interpretation of  $\sigma^{act}$  such that  $0: s \cup 0: e \cup 1: s'$  is a residual stable model of  $D_1$ . A *pf-transition* of  $D$  is a pair  $(\langle s, e, s' \rangle, pf)$ , where  $pf$  is a value assignment to  $\sigma^{pf}$  such that  $0: s \cup 0: e \cup 1: s' \cup 0: pf$  is a stable model of  $D_1$ .

The following simplifying assumptions are made on action descriptions in  $p\mathcal{BC}+$ .

1. **No concurrent execution of actions:** For all transitions  $\langle s, e, s' \rangle$ , we have  $e \models a = \text{TRUE}$  for at most one action constant  $a$ ;
2. **Nondeterministic transitions are determined by pf constants:** For any state  $s$ , any value assignment  $e$  of  $\sigma^{act}$ , and any value assignment  $pf$  of  $\sigma^{pf}$ , there exists exactly one state  $s'$  such that  $(\langle s, e, s' \rangle, pf)$  is a pf-transition;
3. **Nondeterminism on initial states are determined by initpf constants:** For any value assignment  $pf_{init}$  of  $\sigma^{initpf}$ , there exists exactly one value assignment  $fl$  of  $\sigma^{fl}$  such that  $0: pf_{init} \cup 0: fl$  is a stable model of  $D_{init} \cup D_0$ .

With the above three assumptions, the probability of a history, i.e., a sequence of states and actions, can be computed as the product of the probabilities of all the transitions that the history is composed of, multiplied by the probability of the initial state.

A  $p\mathcal{BC}+$  action description defines a probabilistic transition system as follows: A *probabilistic transition system*  $T(D)$  represented by a probabilistic action description  $D$  is a labeled directed graph such that the vertices are the states of  $D$ , and the edges are obtained from the transitions of  $D$ : for every transition  $\langle s, e, s' \rangle$  of  $D$ , an edge labeled  $e: p, u$  goes from  $s$  to  $s'$ , where  $p = P_{D_1}(1: s' \mid 0: s \wedge 0: e)$  and  $u = E[U_{D_1}(0: s \wedge 0: e \wedge 1: s')]$ .<sup>2</sup> The number  $p$  is called the *transition probability* of  $\langle s, e, s' \rangle$ , denoted by  $p(s, e, s')$ , and the number  $u$  is called the *transition reward* of  $\langle s, e, s' \rangle$ , denoted by  $u(s, e, s')$ . The notion of a probabilistic transition system is essentially the same as that of a Markov Decision Process.

## 2.2 Review: POMDP

A Partially Observable Markov Decision Processes (POMDP) is defined as a tuple

$$\langle S, A, T, R, \Omega, O, \gamma \rangle$$

<sup>1</sup>  $\{A\}^{\text{ch}}$  denotes the choice rule  $A \leftarrow \text{not not } A$ .

<sup>2</sup> The *utility* of an interpretation  $I$  under DT-LP<sup>MLN</sup> program  $\Pi$  (Wang and Lee 2019) is defined as  $U_{\Pi}(I) = \sum_{\text{utility}(u,t) \in I} u$  and the *expected utility* of a proposition  $A$  is defined as  $E[U_{\Pi}(A)] = \sum_{I \models A} U_{\Pi}(I) \times P_{\Pi}(I \mid A)$ .

where (i)  $S$  is a set of states; (ii)  $A$  is a set of actions; (iii)  $T : S \times A \times S \rightarrow [0, 1]$  are transition probabilities; (iv)  $R : S \times A \times S \rightarrow \mathbb{R}$  are rewards; (v)  $\Omega$  is a set of observations; (vi)  $O : S \times A \times \Omega \rightarrow [0, 1]$  are observation probabilities; (vii)  $\gamma \in [0, 1]$  is a discount factor.

A *belief state* is a probability distribution over  $S$ . Given the current belief state  $b$ , after taking action  $a \in A$  and observing  $o \in \Omega$ , the updated belief state  $b'$  can be computed as

$$b'(s') = \eta \cdot O(o | s', a) \sum_{s \in S} T(s' | s, a) b(s)$$

where  $s, s' \in S$  are the current and next states respectively;  $b(s)$  is the belief probability in  $b$  corresponding to  $s$ ;  $b'(s')$  is the belief probability in  $b'$  corresponding to  $s'$ ; and  $\eta$  is a normalizer.

A *policy*  $\pi$  is a function from the set of belief states to the set of actions. The *expected total reward* of a stationary policy  $\pi$  starting from the initial belief state  $b_0$  is

$$V^\pi(b_0) = \sum_{t=0}^{\infty} \gamma^t E \left[ R(s_t, \pi(b_t), s_{t+1}) \mid b_0 \right]$$

where  $b_t$  and  $s_t$  are the belief state and the state at time  $t$ , respectively. The optimal policy  $\pi^*$  is obtained by optimizing the long-term reward:  $\pi^* = \underset{\pi}{\operatorname{argmax}} V^\pi(b_0)$ .

### 3 Representing POMDP by Extended $p\mathcal{BC}+$

To be able to express partially observable states, we extend  $p\mathcal{BC}+$  by introducing a new type of constants, called *observation constants*, and a new kind of causal laws called *observation dynamic laws*. An *observation dynamic law* is of the form

$$\text{observed } F \text{ if } G \text{ after } H \tag{1}$$

where  $F$  is a formula containing no constants other than observation constants,  $G$  is a formula containing no constants other than fluent constants, and  $H$  is a formula containing no constants other than action constants and pf constants. Observation constants can occur only in observation dynamic laws. An observation dynamic law  $r$  of the form (1) is translated into the following  $\text{LP}^{\text{MLN}}$  rule:

$$\alpha : (i + 1 : F) \leftarrow (i + 1 : G) \wedge (i : H).$$

For each observation constant  $obs$ ,  $\text{Dom}(obs)$  contains a special value NA (“Not Applicable”). For each observation constant  $obs$  in  $\sigma^{obs}$  and  $v \in \text{Dom}(obs)$ , we include the following  $\text{LP}^{\text{MLN}}$  rule in  $D_m$  to indicate that the initial value of each observation constant is exogenous:

$$\alpha : \{0 : obs = v\}^{\text{ch}}$$

and include the following  $\text{LP}^{\text{MLN}}$  rule in  $D_m$  to indicate that the default value of  $obs$  is NA:

$$\alpha : \{i : obs = \text{NA}\}^{\text{ch}} \quad (i \in \{1, \dots, m\}).$$

For a more flexible representation, we introduce the **if** clause in the pf constant declarations as

$$\text{caused } c = \{v_1 : p_1, \dots, v_n : p_n\} \text{ if } F \tag{2}$$

where  $c$  is a pf constant with the domain  $\{v_1, \dots, v_n\}$ ,  $0 < p_i < 1$  for each  $i \in \{1, \dots, n\}$ ,

$\sum_{i \in \{1, \dots, n\}} p_i = 1$  and  $F$  contains rigid constants only.<sup>3</sup> A pf constant declaration (2) is translated into LP<sup>MLN</sup> rules

$$ln(p_i) : (i : c) = v_j \leftarrow F \quad (3)$$

for  $j \in \{0, \dots, m\}$ . In addition to Assumptions 1–3 above, we add the following assumption:

4. **Rigid constants take the same value over all stable models:** for any rigid constant  $c$ , there exists  $v \in Dom(c)$  such that  $I \models c = v$  for all stable model  $I$  of  $D_m$ .

Under this assumption, the body  $F$  in (3) evaluates to either TRUE or FALSE for all stable models of  $D_m$ , meaning that either (3) can be removed from  $D_m$ , or  $F$  can be removed from the body of (3). Thus, this is not an essential extension but helps us use different probability distributions by changing the condition  $F$ .

Given a  $p\mathcal{BC}+$  action description  $D$ , we use  $\mathbf{S}$  to denote the set of states, i.e, the set of interpretations  $I^{fl}$  of  $\sigma^{fl}$  such that  $0 : I^{fl}$  is a residual (probabilistic) stable model of  $D_0$ . We use  $\mathbf{A}$  to denote the set of interpretations  $I^{act}$  of  $\sigma^{act}$  such that  $0 : I^{act}$  is a residual (probabilistic) stable model of  $D_1$ . Since we assume that at most one action is executed each time step, each element in  $\mathbf{A}$  makes either only one action or none to be true.

*Definition 1*

A  $p\mathcal{BC}+$  action description  $D$ , together with a discount factor  $\gamma$ , defines a POMDP  $M(D) \langle S, A, P, R, \Omega, O, \gamma \rangle$  where

- the state set  $S$  is the same as  $\mathbf{S}$  and the action set  $A$  is the same as  $\mathbf{A}$ ;
- the transition probability  $P$  is defined as  $P(s, a, s') = P_{D_1}(1 : s' \mid 0 : s, 0 : a)$ ;
- the reward function  $R$  is defined as  $R(s, a, s') = E[U_{D_1}(0 : s, 0 : a, 1 : s')]$ ;
- the observation set  $\Omega$  is the set of interpretations  $o$  on  $\sigma^{obs}$  such that  $0 : o$  is a residual stable model of  $D_0$ ;
- the observation probability  $O$  is defined as  $O(s, a, o) = P_{D_1}(1 : o \mid 1 : s, 0 : a)$ .

#### 4 Elaboration Tolerant Representation of POMDP

We illustrate the features of the extended  $p\mathcal{BC}+$  using the “dialog management” example from (Zhang et al. 2017), where a robot is responsible for delivering an item  $i$  to person  $p$  in room  $r$ . The robot needs to ask questions to figure out what  $i, p, r$  are. The challenge comes from the robot’s imperfect speech recognition capability. As a result, repeating questions is sometimes necessary. We use POMDP to model the unreliability from speech recognition, and the robot uses observations to maintain a belief state in the form of a probability distribution. There are two types of questions that the robot can ask:

- Which-Questions: questions about which item/person/room it is, for example, “which item is it?”
- Confirmation-Questions: questions to confirm whether a(n) item/person/room is the requested one, for example, “is the requested item coffee?”

<sup>3</sup> A *rigid* constant is a statically determined fluent constant for which the value is assumed not to change over time (Giunchiglia et al. 2004).

Each of the question-asking actions has a small cost. The robot can execute a *Deliver* action, which consists of an item  $i'$ , person  $p'$  and room  $r'$  as arguments. A *Deliver* action deterministically leads to the terminal state. A reward is obtained with *Deliver* action, determined by to what extent  $i'$ ,  $p'$  and  $r'$  matches  $i$ ,  $p$  and  $r$ . For instance, when all three entries are correctly identified in the *Deliver* action, the agent receives a large reward; when none is correctly identified, the agent receives a large penalty (in the form of a negative reward). Therefore, the agent has the motivation of computing action policies to minimize the cost of its question-asking actions, while maximizing the expected reward by tasking the “correct” delivery action.

This example can be represented in  $pBC+$  as follows. We assume a small domain where  $Item = \{Coffee, Coke, Cookies, Burger\}$ ,  $Person = \{Alice, Bob, Carol\}$ ,  $Room = \{R_1, R_2, R_3\}$ .

---

Notation:  $i, i'$  range over  $Item$ ,  $p, p'$  ranges over  $Person$ ,  $r, r'$  ranges over  $Room$ ,  $c$  ranges over  $\{Yes, No\}$

Observation constant:

*ItemObs*  
*PersonObs*  
*RoomObs*  
*Confirmed*

Domains:

$Item \cup \{NA\}$   
 $Person \cup \{NA\}$   
 $Room \cup \{NA\}$   
 $\{Yes, No, NA\}$

Regular fluent constants:

*ItemReq*  
*PersonReq*  
*RoomReq*  
*Terminated*

Domains:

*Item*  
*Person*  
*Room*  
 Boolean

Action constants:

*WhichItem*, *WhichPerson*, *WhichRoom*,  
*ConfirmItem*( $i$ ), *ConfirmPerson*( $p$ ), *ConfirmRoom*( $r$ ),  
*Deliver*( $i, p, r$ )

Domains:

Boolean

Pf constants:

*Pf\_WhichItem*( $i$ )  
*Pf\_WhichPerson*( $p$ )  
*Pf\_WhichRoom*( $r$ )  
*Pf\_ConfirmWhenCorrect*, *Pf\_ConfirmWhenIncorrect*

Domains:

*Item*  
*Person*  
*Room*  
 $\{Yes, No\}$

---

The action *Deliver* causes the entering of the terminal state:

**caused** *Terminated* **if**  $\top$  **after** *Deliver*( $i, p, r$ ).

The execution of *Deliver* action with the room, the person and the item all correct yields a reward of  $r$ . The execution of *Deliver* action with a wrong item, a wrong person, or a wrong room yield a penalty of  $p_1, p_2, p_3$  each.

**reward**  $r$  **if**  $ItemReq = i \wedge PersonReq = p \wedge RoomReq = r \wedge Deliver(i, p, r) \wedge \sim Terminated$ ,  
**reward**  $-p_1$  **if**  $ItemReq = i \wedge Deliver(i', p', r') \wedge \sim Terminated$  ( $i \neq i'$ ),  
**reward**  $-p_2$  **if**  $PersonReq = p \wedge Deliver(i', p', r') \wedge \sim Terminated$  ( $p \neq p'$ ),  
**reward**  $-p_3$  **if**  $RoomReq = r \wedge Deliver(i', p', r') \wedge \sim Terminated$  ( $r \neq r'$ ).

Asking “which item” question when the actual item being requested is  $i$  returns an item  $i'$  as observation in accordance with the probability distribution defined by pf constant *Pf\_WhichItem*( $i$ ),

shown below. “Which person” and “Which room” questions are represented in a similar way.

$$\begin{aligned}
& \mathbf{observed} \text{ ItemObs} = i' \text{ if } \text{ItemReq} = i \wedge \sim \text{Terminated} \text{ after } \text{WhichItem} \wedge \text{Pf\_WhichItem}(i) = i', \\
& \mathbf{caused} \text{ Pf\_WhichItem}(\text{Coffee}) = \{\text{Coffee} : 0.7, \text{Coke} : 0.1, \text{Cookies} : 0.1, \text{Burger} : 0.1\}, \\
& \mathbf{caused} \text{ Pf\_WhichItem}(\text{Coke}) = \{\text{Coffee} : 0.1, \text{Coke} : 0.7, \text{Cookies} : 0.1, \text{Burger} : 0.1\}, \\
& \mathbf{caused} \text{ Pf\_WhichItem}(\text{Cookies}) = \{\text{Coffee} : 0.1, \text{Coke} : 0.1, \text{Cookies} : 0.7, \text{Burger} : 0.1\}, \\
& \mathbf{caused} \text{ Pf\_WhichItem}(\text{Burger}) = \{\text{Coffee} : 0.1, \text{Coke} : 0.1, \text{Cookies} : 0.1, \text{Burger} : 0.7\},
\end{aligned} \tag{4}$$

When the robot asks the confirmation question “is the item  $i$ ?”, the human’s answer could be sometimes mistakenly recognized, and the probability distribution of the answer depends on whether the item  $i$  is indeed what the human asked for. We use two pf constants,  $\text{Pf\_ConfirmWhenCorrect}$  and  $\text{Pf\_ConfirmWhenIncorrect}$  to specify each of the probability distributions depending on whether the robot’s guess is correct or not. When the robot asks to confirm if the item requested is  $i$ , which is indeed what the human requested:

$$\begin{aligned}
& \mathbf{observed} \text{ Confirmation} = v \text{ if } \text{ItemReq} = i \wedge \sim \text{Terminated} \\
& \quad \mathbf{after} \text{ ConfirmItem}(i) \wedge \text{Pf\_ConfirmWhenCorrect} = v. \quad (v \in \{\text{Yes}, \text{No}\}) \\
& \mathbf{caused} \text{ Pf\_ConfirmWhenCorrect} = \{\text{Yes} : 0.8, \text{No} : 0.2\}.
\end{aligned}$$

When the robot asks to confirm if the requested item is  $i'$  whereas the actual item the human requested is  $i$ :

$$\begin{aligned}
& \mathbf{observed} \text{ Confirmation} = v \text{ if } \text{ItemReq} = i \wedge \sim \text{Terminated} \\
& \quad \mathbf{after} \text{ ConfirmItem}(i') \wedge \text{Pf\_ConfirmWhenIncorrect} = v \quad (i \neq i'), \\
& \mathbf{caused} \text{ Pf\_ConfirmWhenIncorrect} = \{\text{Yes} : 0.2, \text{No} : 0.8\}.
\end{aligned}$$

(The probability distributions of these pf constants do not have to be complementary.)

The formulations of person- and room-related questions are described similarly, and omitted from the paper.

Asking which-questions has a cost of  $c_1$ ; asking confirmation-questions has a cost of  $c_2$ .

$$\begin{aligned}
& \mathbf{reward} -c_1 \text{ if } \top \text{ after } \text{WhichItem}, & \mathbf{reward} -c_2 \text{ if } \top \text{ after } \text{ConfirmItem}(i), \\
& \mathbf{reward} -c_1 \text{ if } \top \text{ after } \text{WhichPerson}, & \mathbf{reward} -c_2 \text{ if } \top \text{ after } \text{ConfirmPerson}(p), \\
& \mathbf{reward} -c_1 \text{ if } \top \text{ after } \text{WhichRoom}, & \mathbf{reward} -c_2 \text{ if } \top \text{ after } \text{ConfirmRoom}(r).
\end{aligned}$$

Finally, all regular fluents in this domain are inertial:

$$\mathbf{inertial} \text{ rf} \quad (\text{rf} \in \{\text{ItemReq}, \text{PersonReq}, \text{RoomReq}, \text{Terminated}\}).$$

In the following subsections, we illustrate the elaboration tolerance of the above  $p\mathcal{BC}+$  action description. It should be noted that using a vanilla POMDP method, manipulating states, actions, or observation functions requires significant engineering efforts, and a developer frequently has to tune prohibitively a large number of parameters. iCORPP and this research aim to avoid that through probabilistic reasoning about actions. In this work, we move forward from iCORPP to shield a developer from the syntax or algorithms of POMDPs.

#### 4.1 Elaboration 1: Unavailable items

When an item becomes unavailable for delivery, we can simply remove that item from the domains of relevant constants. For example, when *Coke* becomes unavailable, we simply replace



the pf constant declarations in (4) with

$$\begin{aligned} \text{caused } Pf\_WhichItem(Coffee) &= \{Coffee : 0.78, Cookies : 0.11, Burger : 0.11\}, \\ \text{caused } Pf\_WhichItem(Cookies) &= \{Coffee : 0.11, Cookies : 0.78, Burger : 0.11\}, \\ \text{caused } Pf\_WhichItem(Burger) &= \{Coffee : 0.11, Cookies : 0.11, Burger : 0.78\}. \end{aligned}$$

#### 4.2 Elaboration 2: Reflecting personal preference in reward function

We use a rigid fluent  $Interchangeable(p, i_1, i_2)$  with the integer domain to represent to what degree the two items  $i_1, i_2$  are interchangeable for person  $p$ . For example, Alice does not mind when the robot delivers coke while she actually ordered coffee but she does mind when the robot delivers burger instead of coffee. We add the following elaboration to represent object interchangeability.

$$\begin{aligned} \text{caused } Interchangeable(Alice, Coffee, Coke) &= 5, \\ \text{caused } Interchangeable(Alice, Coffee, Cookies) &= 1, \\ \text{caused } Interchangeable(Alice, Coffee, Burger) &= -3. \end{aligned}$$

We add the following causal law to reflect the interchangeability of the items.

$$\text{reward } x \text{ if } ItemReq = i \wedge Interchangeable(p, i, i') = x \wedge PersonReq(p) \text{ after } Deliver(i', p', r').$$

Such knowledge can be used to enable the robot to be more conservative in delivering items, such as *burger*, due to their low interchangeability with other items.

#### 4.3 Elaboration 3: Changing Perception Model

The speech recognition system may have different accuracies depending on the environment. For example, when there is background noise, its accuracy could drop. In this case, we can update the probability distribution for the relevant pf constant, controlled by auxiliary constants indicating the situation. We introduce a rigid constant called *Noise*, and then replace (4) with

$$\begin{aligned} \text{caused } Pf\_WhichItem(Coffee) &= \{Coffee : 0.7, Coke : 0.1, Cookies : 0.1, Burger : 0.1\} \text{ unless } ab \\ \text{caused } Pf\_WhichItem(Coke) &= \{Coffee : 0.1, Coke : 0.7, Cookies : 0.1, Burger : 0.1\} \text{ unless } ab \\ \text{caused } Pf\_WhichItem(Cookies) &= \{Coffee : 0.1, Coke : 0.1, Cookies : 0.7, Burger : 0.1\} \text{ unless } ab \\ \text{caused } Pf\_WhichItem(Burger) &= \{Coffee : 0.1, Coke : 0.1, Cookies : 0.1, Burger : 0.7\} \text{ unless } ab \end{aligned} \tag{5}$$

to make them defeasible. We then define the probability distribution to override the original ones when there is loud background noise.

$$\begin{aligned} \text{caused } Pf\_WhichItem(Coffee) &= \{Coffee : \frac{6}{10}, Coke : \frac{4}{30}, Cookies : \frac{4}{30}, Burger : \frac{4}{30}\} \text{ if } Noise, \\ \text{caused } Pf\_WhichItem(Coke) &= \{Coffee : \frac{4}{30}, Coke : \frac{6}{10}, Cookies : \frac{4}{30}, Burger : \frac{4}{30}\} \text{ if } Noise, \\ \text{caused } Pf\_WhichItem(Cookies) &= \{Coffee : \frac{4}{30}, Coke : \frac{4}{30}, Cookies : \frac{6}{10}, Burger : \frac{4}{30}\} \text{ if } Noise, \\ \text{caused } Pf\_WhichItem(Burger) &= \{Coffee : \frac{4}{30}, Coke : \frac{4}{30}, Cookies : \frac{4}{30}, Burger : \frac{6}{10}\} \text{ if } Noise. \end{aligned}$$

We add

**caused** *ab if Noise*

to indicate that by default there is no background noise. When the robot agent detects that there is background noise, we add

**caused** *Noise*

to the action description to update the generated POMDP to incorporate the new speech recognition probabilities. It should be noted that the speech recognition component is generally unreliable, though background noise further reduces its reliability.

## 5 System PBCPLUS2POMDP

We implemented the prototype system PBCPLUS2POMDP, which takes a  $p\mathcal{BC}+$  action description  $D$  as input and outputs the POMDP  $M(D)$  in the input language of the POMDP solver APPL.<sup>4</sup> The system uses LPMLN2ASP (Lee et al. 2017) with exact inference on  $D_1$  and  $D_0$  to generate the components of POMDP: all states, all actions, all transitions and their probabilities, all observations and their probabilities and transition rewards as defined in Definition 1. The system is publicly available at <https://github.com/ywang485/pbcplus2pomdp>, along with several examples.

Even though we limit the computation to  $D_0$  and  $D_1$ , i.e., at most one step action execution is considered, the number of stable models may become too large to enumerate all. Since the transition probabilities, rewards, observation probabilities are per each action, the system implements a compositional way to generate the POMDP model by partitioning the actions in different groups and generating the POMDP model per each group by omitting the causal laws involving other actions and their pf constants. This “compositional” mode often saves the POMDP generation time drastically.<sup>5</sup>

## 6 Evaluation

All experiments reported in this section were performed on a machine powered by 4 Intel(R) Core(TM) i5-2400 CPU with OS Ubuntu 14.04.5 LTS and 8G memory.

### 6.1 Evaluation of Planning Efficiency

We report the running statistics of POMDP generation with our PBCPLUS2POMDP system and POMDP planning with APPL on the dialog example (as described in Section 4) in Table 1. We test domains with different numbers of items, people, and rooms. PBCPLUS2POMDP(NAIVE) generates POMDP in a non-compositional way while PBCPLUS2POMDP(COMPO) generates POMDP in a compositional way (as described in Section 5) by partitioning actions into  $\{ConfirmItem(i) \mid i \in Item\}$ ,  $\{ConfirmPerson(p) \mid p \in Person\}$ ,  $\{ConfirmRoom(r) \mid r \in Room\}$ ,  $\{WhichItem\}$ ,  $\{WhichPerson\}$ ,  $\{WhichRoom\}$ ,  $\{Deliver(i, p, r) \mid i \in Item, p \in Person, r \in Room\}$ .

$\gamma$  is a discount factor. “POMDP solving time (APPL)” refers to the running time of APPL

<sup>4</sup> <http://bigbird.comp.nus.edu.sg/pmwiki/farm/appl/>

<sup>5</sup> The more detailed description of the algorithm is given in Appendix B of the supplementary material corresponding to this paper at the TPLP archives.

Domain Size	POMDP Generation Time		POMDP Solving Time (APPL)		
	PBCPLUS2POMDP (naive)	PBCPLUS2POMDP (compo)	$\gamma = 0.9$	$\gamma = 0.8$	$\gamma = 0.7$
<i>2i2p2r</i> #states = 16 #actions = 18 #observations = 9	49m10.495s	0m13.611s	0m6.123s	0m0.680s	0m0.249s
<i>2i3p2r</i> #states = 24 #actions = 23 #observations = 10	> 1hr	0m22.723s	4m43.572s	0m21.939s	0m2.294s
<i>3i3p2r</i> #states = 36 #actions = 30 #observations = 11	> 1hr	0m41.944s	> 1hr	8m14.415s	0m37.944s
<i>4i3p2r</i> #states = 48 #actions = 37 #observations = 12	> 1hr	2m56.652s	> 1hr	> 1hr	10m50.248s

Table 1. Running Statistics of POMDP Model Generation and Solving in Dialog Example

until the convergence to a target precision of 0.1. The PBCPLUS2POMDP(COMPO) mode is much more efficient than the PBCPLUS2POMDP(NAIVE) mode for the dialog domain.

## 6.2 Evaluation of Solution Quality

*pBC+* provides a high-level description of POMDP models such that various elaborations on the underlying action domain can be easily achieved by changing a small part of the *pBC+* action description, whereas such elaboration would require a complete reconstruction of transition/reward/observation matrices at POMDP level. In Sections 4.1, 4.2 and 4.3, we have illustrated this point with the three example elaborations. In this subsection, we evaluate the impact of the three elaborations on dynamic planning, in the sense that the low-level POMDP (planning module) can be updated automatically once the high-level *pBC+* action description (reasoning module) detects changes in the environment to generate better plans. For each of the three elaborations, we compare the plan generated from a static POMDP that does not reflect environmental changes, and the one generated from the adaptive POMDP that is updated by *pBC+* reasoning to reflect environmental changes.

Fig. 2 compares the policies generated from the static POMDPs (baseline) and from the POMDP dynamically generated using *pBC+*, where the two items of burger and cookies might be unavailable (Elaboration 1). We have run 1000 simulation trials. The diagram on the left compares them in terms of average total reward from the simulation runs, and the right is in terms of average QA cost (accumulated penalty by asking questions). In this experiment, the discount factor is 0.95 (which offers the dialog agent a relatively long horizon),  $c_1$  is 4.0,  $c_2$  is 2.0,  $r$  is 20.0,  $p_2$  is 20.0, and  $p_3$  is 30.0. Action policies are generated using APPL in at most 120 seconds. We observe that the adaptive POMDP (ours) achieves a higher average total reward when the

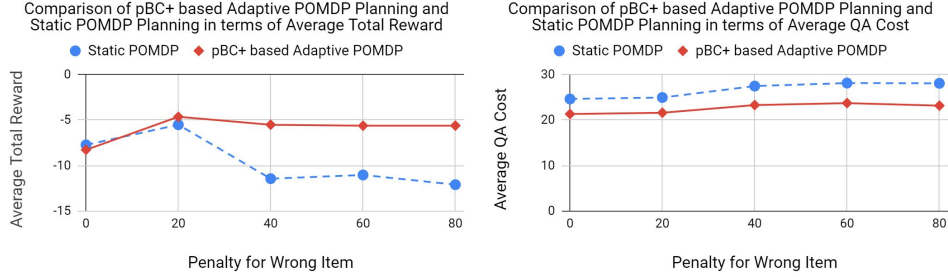


Fig. 2. Impact of Elaboration 1 on Policy Generated

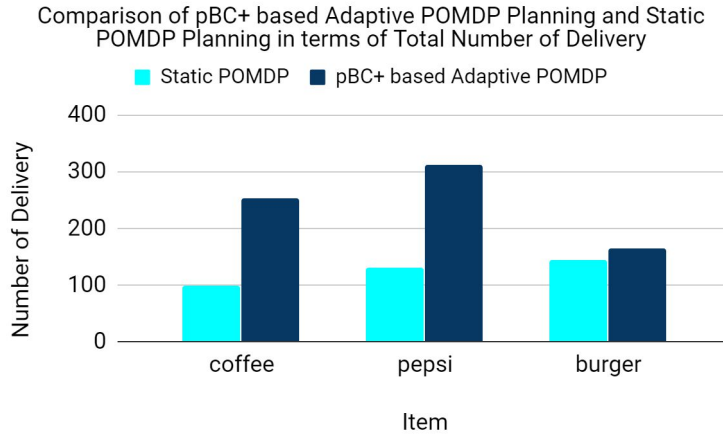


Fig. 3. Impact of Elaboration 2 on Policy Generated

penalty for the wrong item is positive, and the adaptive POMDPs are able to complete deliveries with less QA costs. It is worth noting that by reflecting unavailable items,  $pBC+$  reduces the size of the generated POMDP models, resulting in shorter POMDP-solving times. As can be seen from Table 1, for a domain that contains 2 items, 3 people and 2 rooms, POMDP generation plus POMDP solving takes way less time than POMDP solving on a domain with 4 items, 3 people and 2 rooms.

Fig. 3 compares the policies generated from the static POMDP and from  $pBC+$  based adaptive POMDP when item interchangeability is introduced (Elaboration 2). We replaced cookies with pepsi in the domain, added causal laws to indicate that when coke is being requested, delivering pepsi instead yields a reward of 15, delivering coffee instead yields a reward of 5 and delivering burger instead yields an additional penalty of 20 (in the presence of penalty  $p_1$ ). We have run 10000 simulations, and for all of the simulations, the actual item being requested is fixed to be coke.<sup>6</sup> For the static POMDP, 9628 deliveries were correct, and for the adaptive POMDP, 9270 deliveries were correct. Note that although the static POMDP achieves more correct deliveries, the dynamically generated POMDPs (our approach) achieved higher average total reward by

<sup>6</sup> The item is fixed to be coke only during simulation, not during policy generation.

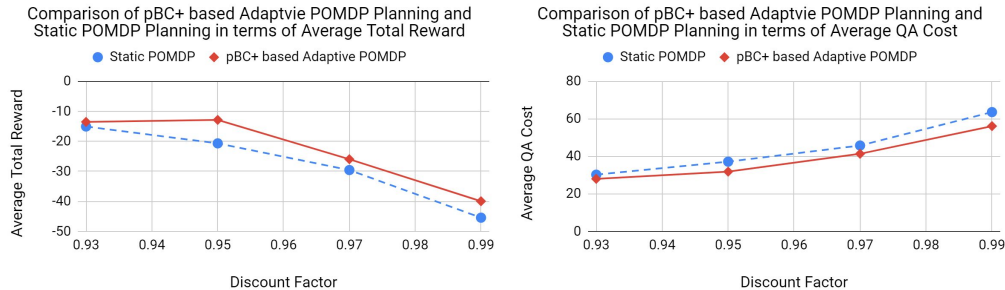


Fig. 4. Impact of Elaboration 3 on Policy Generated

asking fewer questions. The policy generated from the static POMDP gives similar numbers of deliveries for each item that is not coke, while the policy generated from the adaptive POMDP delivered pepsi the most and burger the least, which is aligned with our setting of interchangeability. The discount factor for this experiment is set to be 0.99.  $c_1$  is 6,  $c_2$  is 4,  $r$  is 5,  $p_1$  is 5,  $p_2$  is 20 and  $p_3$  is 30. Policies from both POMDPs are generated by APPL with 120 seconds.

Fig. 4 compares the policies generated from the static POMDP and from  $p\mathcal{BC}+$  based adaptive POMDP when there is a background noise (Elaboration 3). To reflect environmental noise, we lowered the observation probability of correct answers by 0.1 (and the remaining answers are uniformly distributed). We have run 1000 simulations. The diagram on the left compares them in term of average total reward from the simulation runs, and the diagram on the right compares them in terms of average QA cost (accumulated cost from questions asked) from the simulation runs. In this experiment,  $c_1$  is 4,  $c_2$  is 2,  $r$  is 20,  $p_2$  is 20 and  $p_3$  is 30. Policies from both POMDPs are generated by APPL with 120 seconds. It can be seen from the diagrams that while the average total reward of both POMDPs decreases as the discount factor increases, the adaptive POMDP achieves higher average total reward by asking fewer questions.

## 7 Related Work

Intelligent agents need the capabilities of both reasoning about declarative knowledge, and probabilistic planning toward achieving long-term goals. A variety of algorithms have been developed to integrate commonsense reasoning and probabilistic planning (Hanheide et al. 2017; Zhang et al. 2015; Zhang and Stone 2015; Sridharan et al. 2019; Chitnis et al. 2018; Zhang et al. 2017; Amiri et al. 2018; Veiga et al. 2019), and some of them, such as (Sridharan et al. 2019) and (Amiri et al. 2018), also include non-deterministic dynamic laws for observations. Although the algorithms use very different computational paradigms for representing and reasoning with human knowledge (e.g., logics, probabilities, graphs, etc), they all share the goal of leveraging declarative knowledge to improve the performance in probabilistic planning. In these works, the hypothesis is that human knowledge potentially can be useful in guiding robot behaviors in the real world, while the challenge is that human knowledge is sparse, incomplete, and sometimes unreliable. In this research, we share the same goal of utilizing contextual knowledge from people to help intelligent agents in sequential decision-making tasks while accounting for the uncertainty in perception and action outcomes.

Among the algorithms that integrate commonsense reasoning and probabilistic planning paradigms, iCORPP enabled an agent to reason with contextual knowledge to dynamically construct com-

plete probabilistic planning models (Zhang et al. 2017) for adaptive robot control, where P-log was used for logical-probabilistic reasoning (Baral et al. 2009). Depending on the observability of world states, iCORPP uses either Markov Decision Processes (MDPs) (Puterman 2014) or Partially Observable MDPs (POMDPs) (Kaelbling et al. 1998) for probabilistic planning. As a result, iCORPP has been applied to robot navigation, dialog system, and manipulation tasks (Zhang et al. 2017; Amiri et al. 2018). In this work, we develop a unified representation and a corresponding implementation for iCORPP, where the entire reasoning and planning system can be encoded using a single program, and practitioners are completely shielded from the technical details of formulating and solving (PO)MDPs. In comparison, iCORPP requires significant engineering efforts (e.g., using Python or C++) for “gluing” the computational paradigms used by the commonsense reasoning and probabilistic planning components.

Recently, researchers have developed algorithms to incorporate knowledge representation and reasoning into reinforcement learning (RL) (Sutton and Barto 2018), where the goal is to provide the learning agents with guidance in action selections through reasoning with declarative knowledge. Notable examples include (Leonetti et al. 2016; Yang et al. 2018; Jiang et al. 2018; Lu et al. 2018; Lyu et al. 2019; Kim et al. 2019). In this research, we assume the availability of world models, including both states and dynamics, in a declarative form. In case of world models being unavailable, incomplete, or dynamically changing, there is the potential of combining the above “knowledge-driven RL” algorithms, particularly the ones using model-based RL such as (Lu et al. 2018), with our new representation to enable agents to simultaneously learn and reason about world models to compute action policies.

In an earlier work (Tran and Baral 2004), the authors show how Pearl’s probabilistic causal model can be encoded in a probabilistic action language PAL (Baral et al. 2002).

## 8 Conclusion and Future Work

In this paper, we present a principled way of integrating probabilistic logical reasoning and probabilistic planning. This is done by extending probabilistic action language  $p\mathcal{BC}+$  (Lee and Wang 2018; Wang and Lee 2019) to be able to express utility, belief states, and observation as in POMDP models. Inheriting the advantages of action languages, the new action language provides an elaboration tolerant representation of POMDP that is convenient to encode commonsense knowledge.

One of the well known problems limiting applications of POMDPs is sensitivity of the optimal behavior to the small changes in the reward function and the probability distribution. Because of this sensitivity care must be taken in choosing the reward function as well as the probability distribution. The choice of these, and especially of the latter is a non-trivial problem, which is outside of the scope of the paper. POMDP algorithms perform poorly in scalability in many applications. Although the language and system developed in this paper can potentially alleviate this issue, we believe this is a challenging problem that deserves more effort, and we leave it to future work.

The current prototype implementation is not highly scalable when the number of transitions becomes large. For a more scalable generation of the POMDP input using the  $LP^{MLN}$  system, we could use the sampling method in  $LP^{MLN}$  inference, which we leave for future work.

**Acknowledgements:** We are grateful to the anonymous referees for their useful comments. The

first and the third author’s work was partially supported by the National Science Foundation under Grant IIS-1815337.

## References

- AMIRI, S., WEI, S., ZHANG, S., SINAPOV, J., THOMASON, J., AND STONE, P. 2018. Multi-modal predicate identification using dynamically learned robot controllers. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*.
- BABB, J. AND LEE, J. 2015. Action language  $\mathcal{BC}+$ . *Journal of Logic and Computation*, exv062.
- BARAL, C., GELFOND, M., AND RUSHTON, J. N. 2009. Probabilistic reasoning with answer sets. *Theory and Practice of Logic Programming* 9, 1, 57–144.
- BARAL, C., TRAN, N., AND TUAN, L.-C. 2002. Reasoning about actions in a probabilistic setting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 507–512.
- CHITNIS, R., KAEHLING, L. P., AND LOZANO-PÉREZ, T. 2018. Integrating human-provided information into belief state representation using dynamic factorization. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3551–3558. IEEE.
- GIUNCHIGLIA, E., LEE, J., LIFSCHITZ, V., MCCAIN, N., AND TURNER, H. 2004. Nonmonotonic causal theories. *Artificial Intelligence* 153(1–2), 49–104.
- HANHEIDE, M., GÖBELBECKER, M., HORN, G. S., PRONOBIS, A., SJÖÖ, K., AYDEMIR, A., JENSFELT, P., GRETTON, C., DEARDEN, R., JANICEK, M., ET AL. 2017. Robot task planning and explanation in open and uncertain worlds. *Artificial Intelligence* 247, 119–150.
- JIANG, Y., YANG, F., ZHANG, S., AND STONE, P. 2018. Integrating task-motion planning with reinforcement learning for robust decision making in mobile robots. *CoRR abs/1811.08955*.
- KAEHLING, L. P., LITTMAN, M. L., AND CASSANDRA, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1-2, 99–134.
- KIM, B., KAEHLING, L. P., AND LOZANO-PÉREZ, T. 2019. Adversarial actor-critic method for task and motion planning problems using planning experience. In *AAAI Conference on Artificial Intelligence (AAAI)*.
- LEE, J., LIFSCHITZ, V., AND YANG, F. 2013. Action language  $\mathcal{BC}$ : Preliminary report. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*.
- LEE, J., TALSANIA, S., AND WANG, Y. 2017. Computing LPMLN using ASP and MLN solvers. *Theory and Practice of Logic Programming*.
- LEE, J. AND WANG, Y. 2016. Weighted rules under the stable model semantics. In *Proceedings of International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pp. 145–154.
- LEE, J. AND WANG, Y. 2018. A probabilistic extension of action language  $\mathcal{BC}+$ . *Theory and Practice of Logic Programming* 18(3–4), 607–622.
- LEONETTI, M., IOCCHI, L., AND STONE, P. 2016. A synthesis of automated planning and reinforcement learning for efficient, robust decision-making. *Artificial Intelligence* 241, 103–130.
- LU, K., ZHANG, S., STONE, P., AND CHEN, X. 2018. Robot representing and reasoning with knowledge from reinforcement learning. *CoRR abs/1809.11074*.
- LYU, D., YANG, F., LIU, B., AND GUSTAFSON, S. 2019. Sdrl: Interpretable and data-efficient deep reinforcement learning leveraging symbolic planning. In *AAAI*.
- PUTERMAN, M. L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.
- SRIDHARAN, M., GELFOND, M., ZHANG, S., AND WYATT, J. 2019. REBA: A refinement-based architecture for knowledge representation and reasoning in robotics. *Journal of Artificial Intelligence Research* 65, 87–180.
- SUTTON, R. S. AND BARTO, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.
- TRAN, N. AND BARAL, C. 2004. Encoding probabilistic causal model in probabilistic action language. In *Proceedings of the National Conference on Artificial Intelligence*.

- VEIGA, T. S., SILVA, M., VENTURA, R., AND LIMA, P. U. 2019. A hierarchical approach to active semantic mapping using probabilistic logic and information reward pomdps. In *Proceedings of the International Conference on Automated Planning and Scheduling*.
- WANG, Y. AND LEE, J. 2019. Elaboration tolerant representation of markov decision process via decision theoretic extension of action language pbc+. In *LPNMR*. To appear.
- YANG, F., LYU, D., LIU, B., AND GUSTAFSON, S. 2018. Peorl: integrating symbolic planning and hierarchical reinforcement learning for robust decision-making. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 4860–4866.
- ZHANG, S., KHANDELWAL, P., AND STONE, P. 2017. Dynamically constructed (PO)MDPs for adaptive robot planning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.
- ZHANG, S., SRIDHARAN, M., AND WYATT, J. L. 2015. Mixed logical inference and probabilistic planning for robots in unreliable worlds. *IEEE Transactions on Robotics* 31, 3, 699–713.
- ZHANG, S. AND STONE, P. 2015. CORPP: Commonsense reasoning and probabilistic planning, as applied to dialog with a mobile robot. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.



Online appendix for the paper

*Bridging Commonsense Reasoning and Probabilistic Planning via  
a Probabilistic Action Language*

published in Theory and Practice of Logic Programming

Yi Wang\*, Shiqi Zhang#, Joohyung Lee\*

\*Arizona State University, USA # SUNY Binghamton, USA

**Appendix A Extended Review of Preliminaries**

**A.1 Review: Language  $LP^{MLN}$**

An  $LP^{MLN}$  program is a finite set of weighted rules  $w : R$  where  $R$  is a rule and  $w$  is a real number (in which case, the weighted rule is called *soft*) or  $\alpha$  for denoting the infinite weight (in which case, the weighted rule is called *hard*). Throughout the paper, we assume that the language is propositional. Schematic variables can be introduced via grounding as usual in answer set programming.

For any  $LP^{MLN}$  program  $\Pi$  and any interpretation  $I$ ,  $\bar{\Pi}$  denotes the usual (unweighted) ASP program obtained from  $\Pi$  by dropping the weights, and  $\Pi_I$  denotes the set of  $w : R$  in  $\Pi$  such that  $I \models R$ .

In general, an  $LP^{MLN}$  program may even have stable models that violate some hard rules, which encode definite knowledge. However, throughout the paper, we restrict attention to  $LP^{MLN}$  programs whose stable models do not violate hard rules. More precisely, given a ground  $LP^{MLN}$  program  $\Pi$ ,  $SM[\Pi]$  denotes the set

$$\{I \mid I \text{ is a (deterministic) stable model of } \Pi_I \text{ that satisfies all hard rules in } \Pi\}.$$

The weight of an interpretation  $I$ , denoted  $W_{\Pi}(I)$ , is defined as

$$W_{\Pi}(I) = \begin{cases} \exp\left(\sum_{w:R \in \Pi_I} w\right) & \text{if } I \in SM[\Pi]; \\ 0 & \text{otherwise,} \end{cases}$$

and the probability of  $I$ , denoted  $P_{\Pi}(I)$ , is defined as

$$P_{\Pi}(I) = \frac{W_{\Pi}(I)}{\sum_{J \in SM[\Pi]} W_{\Pi}(J)}.$$

**A.2 Review: DT- $LP^{MLN}$**

We extend the syntax and the semantics of  $LP^{MLN}$  to DT- $LP^{MLN}$  by introducing atoms of the form

$$\text{utility}(u, \mathbf{t}) \tag{A1}$$

where  $u$  is a real number, and  $\mathbf{t}$  is an arbitrary list of terms. These atoms can only occur in the head of hard rules of the form

$$\alpha : \text{utility}(u, \mathbf{t}) \leftarrow \text{Body} \tag{A2}$$

where *Body* is a list of literals. We call these rules *utility rules*.

The weight and the probability of an interpretation are defined the same as in  $\text{LP}^{\text{MLN}}$ . The *utility* of an interpretation  $I$  under  $\Pi$  is defined as

$$U_{\Pi}(I) = \sum_{\text{utility}(u,t) \in I} u.$$

The *expected utility* of a proposition  $A$  is defined as

$$E[U_{\Pi}(A)] = \sum_{I \models A} U_{\Pi}(I) \times P_{\Pi}(I \mid A). \quad (\text{A3})$$

### A.3 Review: Multi-Valued Probabilistic Programs

Multi-valued probabilistic programs (Lee and Wang 2016) are a simple fragment of  $\text{LP}^{\text{MLN}}$  that allows us to represent probability more naturally.

We assume that the propositional signature  $\sigma$  is constructed from “constants” and their “values.” A *constant*  $c$  is a symbol that is associated with a finite set  $\text{Dom}(c)$ , called the *domain*. The signature  $\sigma$  is constructed from a finite set of constants, consisting of atoms  $c = v$ <sup>7</sup> for every constant  $c$  and every element  $v$  in  $\text{Dom}(c)$ . If the domain of  $c$  is  $\{\text{FALSE}, \text{TRUE}\}$  then we say that  $c$  is *Boolean*, and abbreviate  $c = \text{TRUE}$  as  $c$  and  $c = \text{FALSE}$  as  $\sim c$ .

We assume that constants are divided into *probabilistic* constants and *non-probabilistic* constants. A multi-valued probabilistic program  $\Pi$  is a tuple  $\langle PF, \Pi \rangle$ , where

- $PF$  contains *probabilistic constant declarations* of the following form:

$$p_1 :: c = v_1 \mid \cdots \mid p_n :: c = v_n \quad (\text{A4})$$

one for each probabilistic constant  $c$ , where  $\{v_1, \dots, v_n\} = \text{Dom}(c)$ ,  $v_i \neq v_j$ ,  $0 \leq p_1, \dots, p_n \leq 1$  and  $\sum_{i=1}^n p_i = 1$ . We use  $M_{\Pi}(c = v_i)$  to denote  $p_i$ . In other words,  $PF$  describes the probability distribution over each “random variable”  $c$ .

- $\Pi$  is a set of rules such that the head contains no probabilistic constants.

The semantics of such a program  $\Pi$  is defined as a shorthand for  $\text{LP}^{\text{MLN}}$  program  $T(\Pi)$  of the same signature as follows.

- For each probabilistic constant declaration (A4),  $T(\Pi)$  contains, for each  $i = 1, \dots, n$ , (i)  $\text{ln}(p_i) : c = v_i$  if  $0 < p_i < 1$ ; (ii)  $\alpha : c = v_i$  if  $p_i = 1$ ; (iii)  $\alpha : \perp \leftarrow c = v_i$  if  $p_i = 0$ .
- For each rule  $\text{Head} \leftarrow \text{Body}$  in  $\Pi$ ,  $T(\Pi)$  contains  $\alpha : \text{Head} \leftarrow \text{Body}$ .
- For each constant  $c$ ,  $T(\Pi)$  contains the uniqueness of value constraints

$$\alpha : \perp \leftarrow c = v_1 \wedge c = v_2 \quad (\text{A5})$$

for all  $v_1, v_2 \in \text{Dom}(c)$  such that  $v_1 \neq v_2$ , and the existence of value constraint

$$\alpha : \perp \leftarrow \neg \bigvee_{v \in \text{Dom}(c)} c = v. \quad (\text{A6})$$

In the presence of the constraints (A5) and (A6), assuming  $T(\Pi)$  has at least one (probabilistic) stable model that satisfies all the hard rules, a (probabilistic) stable model  $I$  satisfies  $c = v$  for exactly one value  $v$ , so we may identify  $I$  with the value assignment that assigns  $v$  to  $c$ .

<sup>7</sup> Note that here “=” is just a part of the symbol for propositional atoms, and is not equality in first-order logic.

#### A.4 Review: Action Language $p\mathcal{BC}+$ with Utility

##### Syntax of $p\mathcal{BC}+$

We assume a propositional signature  $\sigma$  as defined in Section A.3. We further assume that the signature of an action description is divided into four groups: *fluent constants*, *action constants*, *pf (probability fact) constants* and *initpf (initial probability fact) constants*. Fluent constants are further divided into *regular* and *statically determined*. The domain of every action constant is Boolean. A *fluent formula* is a formula such that all constants occurring in it are fluent constants.

The following definition of  $p\mathcal{BC}+$  is based on the definition of  $\mathcal{BC}+$  language from (Babb and Lee 2015).

A *static law* is an expression of the form

$$\text{caused } F \text{ if } G \tag{A7}$$

where  $F$  and  $G$  are fluent formulas.

A *fluent dynamic law* is an expression of the form

$$\text{caused } F \text{ if } G \text{ after } H \tag{A8}$$

where  $F$  and  $G$  are fluent formulas and  $H$  is a formula, provided that  $F$  does not contain statically determined constants and  $H$  does not contain initpf constants.

A *pf constant declaration* is an expression of the form

$$\text{caused } c = \{v_1 : p_1, \dots, v_n : p_n\} \tag{A9}$$

where  $c$  is a pf constant with domain  $\{v_1, \dots, v_n\}$ ,  $0 < p_i < 1$  for each  $i \in \{1, \dots, n\}$ <sup>8</sup>, and  $p_1 + \dots + p_n = 1$ . In other words, (A9) describes the probability distribution of  $c$ .

An *initpf constant declaration* is an expression of the form (A9) where  $c$  is an initpf constant.

An *initial static law* is an expression of the form

$$\text{initially } F \text{ if } G \tag{A10}$$

where  $F$  is a fluent constant and  $G$  is a formula that contains neither action constants nor pf constants.

A *causal law* is a static law, a fluent dynamic law, a pf constant declaration, an initpf constant declaration, or an initial static law. An *action description* is a finite set of causal laws.

We use  $\sigma^{fl}$  to denote the set of fluent constants,  $\sigma^{act}$  to denote the set of action constants,  $\sigma^{pf}$  to denote the set of pf constants, and  $\sigma^{initpf}$  to denote the set of initpf constants. For any signature  $\sigma'$  and any  $i \in \{0, \dots, m\}$ , we use  $i : \sigma'$  to denote the set  $\{i : a \mid a \in \sigma'\}$ .

By  $i : F$  we denote the result of inserting  $i :$  in front of every occurrence of every constant in formula  $F$ . This notation is straightforwardly extended when  $F$  is a set of formulas.

##### Semantics of $p\mathcal{BC}+$

Given a non-negative integer  $m$  denoting the maximum length of histories, the semantics of an action description  $D$  in  $p\mathcal{BC}+$  is defined by a reduction to multi-valued probabilistic program  $Tr(D, m)$ , which is the union of two subprograms  $D_m$  and  $D_{init}$  as defined below.

<sup>8</sup> We require  $0 < p_i < 1$  for each  $i \in \{1, \dots, n\}$  for the sake of simplicity. On the other hand, if  $p_i = 0$  or  $p_i = 1$  for some  $i$ , that means either  $v_i$  can be removed from the domain of  $c$  or there is not really a need to introduce  $c$  as a pf constant. So this assumption does not really sacrifice expressivity.

For an action description  $D$  of a signature  $\sigma$ , we define a sequence of multi-valued probabilistic program  $D_0, D_1, \dots$ , so that the stable models of  $D_m$  can be identified with the paths in the transition system described by  $D$ .

The signature  $\sigma_m$  of  $D_m$  consists of atoms of the form  $i : c = v$  such that

- for each fluent constant  $c$  of  $D$ ,  $i \in \{0, \dots, m\}$  and  $v \in \text{Dom}(c)$ ,
- for each action constant or pf constant  $c$  of  $D$ ,  $i \in \{0, \dots, m-1\}$  and  $v \in \text{Dom}(c)$ .

For  $x \in \{act, fl, pf\}$ , we use  $\sigma_m^x$  to denote the subset of  $\sigma_m$

$$\{i : c = v \mid i : c = v \in \sigma_m \text{ and } c \in \sigma^x\}.$$

For  $i \in \{0, \dots, m\}$ , we use  $i : \sigma^x$  to denote the subset of  $\sigma_m^x$

$$\{i : c = v \mid i : c = v \in \sigma_m^x\}.$$

We define  $D_m$  to be the multi-valued probabilistic program  $\langle PF, \Pi \rangle$ , where  $\Pi$  is the conjunction of

$$i : F \leftarrow i : G \tag{A11}$$

for every static law (A7) in  $D$  and every  $i \in \{0, \dots, m\}$ ,

$$i+1 : F \leftarrow (i+1 : G) \wedge (i : H) \tag{A12}$$

for every fluent dynamic law (A8) in  $D$  and every  $i \in \{0, \dots, m-1\}$ ,

$$\{0 : c = v\}^{\text{ch}} \tag{A13}$$

for every regular fluent constant  $c$  and every  $v \in \text{Dom}(c)$ ,

$$\{i : c = \text{TRUE}\}^{\text{ch}}, \quad \{i : c = \text{FALSE}\}^{\text{ch}} \tag{A14}$$

for every action constant  $c$ , and  $PF$  consists of

$$p_1 :: i : pf = v_1 \mid \dots \mid p_n :: i : pf = v_n \tag{A15}$$

( $i = 0, \dots, m-1$ ) for each pf constant declaration (A9) in  $D$  that describes the probability distribution of  $pf$ .

In addition, we define the program  $D_{init}$ , whose signature is  $0 : \sigma^{\text{initpf}} \cup 0 : \sigma^{\text{fl}}$ .  $D_{init}$  is the multi-valued probabilistic program

$$D_{init} = \langle PF^{\text{init}}, \Pi^{\text{init}} \rangle$$

where  $\Pi^{\text{init}}$  consists of the rule

$$\perp \leftarrow \neg(0 : F) \wedge 0 : G$$

for each initial static law (A10), and  $PF^{\text{init}}$  consists of

$$p_1 :: 0 : pf = v_1 \mid \dots \mid p_n :: 0 : pf = v_n$$

for each initpf constant declaration (A9).

We define  $Tr(D, m)$  to be the union of the two multi-valued probabilistic program  $\langle PF \cup PF^{\text{init}}, \Pi \cup \Pi^{\text{init}} \rangle$ .

For any  $\text{LP}^{\text{MLN}}$  program  $\Pi$  of signature  $\sigma$  and a value assignment  $I$  to a subset  $\sigma'$  of  $\sigma$ , we say  $I$  is a *residual (probabilistic) stable model* of  $\Pi$  if there exists a value assignment  $J$  to  $\sigma \setminus \sigma'$  such that  $I \cup J$  is a (probabilistic) stable model of  $\Pi$ .

For any value assignment  $I$  to constants in  $\sigma$ , by  $i : I$  we denote the value assignment to constants in  $i : \sigma$  so that  $i : I \models (i : c) = v$  iff  $I \models c = v$ .

We define a *state* as an interpretation  $I^{fl}$  of  $\sigma^{fl}$  such that  $0 : I^{fl}$  is a residual (probabilistic) stable model of  $D_0$ . A *transition* of  $D$  is a triple  $\langle s, e, s' \rangle$  where  $s$  and  $s'$  are interpretations of  $\sigma^{fl}$  and  $e$  is an interpretation of  $\sigma^{act}$  such that  $0 : s \cup 0 : e \cup 1 : s'$  is a residual stable model of  $D_1$ . A *pf-transition* of  $D$  is a pair  $(\langle s, e, s' \rangle, pf)$ , where  $pf$  is a value assignment to  $\sigma^{pf}$  such that  $0 : s \cup 0 : e \cup 1 : s' \cup 0 : pf$  is a stable model of  $D_1$ .

A *probabilistic transition system*  $T(D)$  represented by a probabilistic action description  $D$  is a labeled directed graph such that the vertices are the states of  $D$ , and the edges are obtained from the transitions of  $D$ : for every transition  $\langle s, e, s' \rangle$  of  $D$ , an edge labeled  $e : p$  goes from  $s$  to  $s'$ , where  $p = Pr_{D_1}(1 : s' \mid 0 : s, 0 : e)$ . The number  $p$  is called the *transition probability* of  $\langle s, e, s' \rangle$ .

The soundness of the definition of a probabilistic transition system relies on the following proposition.

*Proposition 1*

For any transition  $\langle s, e, s' \rangle$ ,  $s$  and  $s'$  are states.

We make the following simplifying assumptions on action descriptions:

1. **No concurrent execution of actions:** For all transitions  $\langle s, e, s' \rangle$ , we have  $e \models a = \text{TRUE}$  for at most one action constant  $a$ ;
2. **Nondeterministic transitions are determined by pf constants:** For any state  $s$ , any value assignment  $e$  of  $\sigma^{act}$ , and any value assignment  $pf$  of  $\sigma^{pf}$ , there exists exactly one state  $s'$  such that  $(\langle s, e, s' \rangle, pf)$  is a pf-transition;
3. **Nondeterminism on initial states are determined by initpf constants:** For any value assignment  $pf_{init}$  of  $\sigma^{initpf}$ , there exists exactly one value assignment  $fl$  of  $\sigma^{fl}$  such that  $0 : pf_{init} \cup 0 : fl$  is a stable model of  $D_{init} \cup D_0$ .

For any state  $s$ , any value assignment  $e$  of  $\sigma^{act}$  such that at most one action is true, and any value assignment  $pf$  of  $\sigma^{pf}$ , we use  $\phi(s, e, pf)$  to denote the state  $s'$  such that  $(\langle s, e, s' \rangle, pf)$  is a pf-transition (According to Assumption 2, such  $s'$  must be unique). For any interpretation  $I$ ,  $i \in \{0, \dots, m\}$  and any subset  $\sigma'$  of  $\sigma$ , we use  $I|_{i:\sigma'}$  to denote the value assignment of  $I$  to atoms in  $i : \sigma'$ . Given any value assignment  $TC$  of  $0 : \sigma^{initpf} \cup \sigma_m^{pf}$  and a value assignment  $A$  of  $\sigma_m^{act}$ , we construct an interpretation  $I_{TC \cup A}$  of  $Tr(D, m)$  that satisfies  $TC \cup A$  as follows:

- For all atoms  $p$  in  $\sigma_m^{pf} \cup 0 : \sigma^{initpf}$ , we have  $I_{TC \cup A}(p) = TC(p)$ ;
- For all atoms  $p$  in  $\sigma_m^{act}$ , we have  $I_{TC \cup A}(p) = A(p)$ ;
- $(I_{TC \cup A})|_{0:\sigma^{fl}}$  is the assignment such that  $(I_{TC \cup A})|_{0:\sigma^{fl} \cup 0:\sigma^{initpf}}$  is a stable model of  $D_{init} \cup D_0$ .
- For each  $i \in \{1, \dots, m\}$ ,

$$(I_{TC \cup A})|_{i:\sigma^{fl}} = \phi((I_{TC \cup A})|_{(i-1):\sigma^{fl}}, (I_{TC \cup A})|_{(i-1):\sigma^{act}}, (I_{TC \cup A})|_{(i-1):\sigma^{pf}}).$$

By Assumptions 2 and 3, the above construction produces a unique interpretation.

It can be seen that in the multi-valued probabilistic program  $Tr(D, m)$  translated from  $D$ , the probabilistic constants are  $0 : \sigma^{initpf} \cup \sigma_m^{pf}$ . We thus call the value assignment of an interpretation  $I$  on  $0 : \sigma^{initpf} \cup \sigma_m^{pf}$  the *total choice* of  $I$ . The following theorem asserts that the probability of a stable model under  $Tr(D, m)$  can be computed by simply dividing the probability of the total choice associated with the stable model by the number of choice of actions.

*Theorem 1*

For any value assignment  $TC$  of  $0 : \sigma^{initpf} \cup \sigma_m^{pf}$  and any value assignment  $A$  of  $\sigma_m^{act}$ , there exists exactly one stable model  $I_{TC \cup A}$  of  $Tr(D, m)$  that satisfies  $TC \cup A$ , and the probability of  $I_{TC \cup A}$  is

$$Pr_{Tr(D, m)}(I_{TC \cup A}) = \frac{\prod_{c=v \in TC} M(c=v)}{(|\sigma^{act}| + 1)^m}.$$

The following theorem tells us that the conditional probability of transiting from a state  $s$  to another state  $s'$  with action  $e$  remains the same for all timesteps, i.e., the conditional probability of  $i+1 : s'$  given  $i : s$  and  $i : e$  correctly represents the transition probability from  $s$  to  $s'$  via  $e$  in the transition system.

*Theorem 2*

For any state  $s$  and  $s'$ , and action  $e$ , we have

$$Pr_{Tr(D, m)}(i+1 : s' \mid i : s, i : e) = Pr_{Tr(D, m)}(j+1 : s' \mid j : s, j : e)$$

for any  $i, j \in \{0, \dots, m-1\}$  such that  $Pr_{Tr(D, m)}(i : s) > 0$  and  $Pr_{Tr(D, m)}(j : s) > 0$ .

For every subset  $X_m$  of  $\sigma_m \setminus \sigma_m^{pf}$ , let  $X^i (i < m)$  be the triple consisting of

- the set consisting of atoms  $A$  such that  $i : A$  belongs to  $X_m$  and  $A \in \sigma^{fl}$ ;
- the set consisting of atoms  $A$  such that  $i : A$  belongs to  $X_m$  and  $A \in \sigma^{act}$ ;
- the set consisting of atoms  $A$  such that  $i+1 : A$  belongs to  $X_m$  and  $A \in \sigma^{fl}$ .

Let  $p(X^i)$  be the transition probability of  $X^i$ ,  $s_0$  is the interpretation of  $\sigma_0^{fl}$  defined by  $X^0$ , and  $e_i$  be the interpretations of  $i : \sigma^{act}$  defined by  $X^i$ .

Since the transition probability remains the same, the probability of a path given a sequence of actions can be computed from the probabilities of transitions.

*Corollary 1*

For every  $m \geq 1$ ,  $X_m$  is a residual (probabilistic) stable model of  $Tr(D, m)$  iff  $X^0, \dots, X^{m-1}$  are transitions of  $D$  and  $0 : s_0$  is a residual stable model of  $D_{init}$ . Furthermore,

$$Pr_{Tr(D, m)}(X_m \mid 0 : e_0, \dots, m-1 : e_{m-1}) = p(X^0) \times \dots \times p(X^m) \times Pr_{Tr(D, m)}(0 : s_0).$$

*pBC+ with Utility*

Wang and Lee (2019) has extended *pBC+* with the notion of utility as follows.

We extend *pBC+* by introducing the following expression called *utility law* that assigns a reward to transitions:

$$\text{reward } v \text{ if } F \text{ after } G \tag{A16}$$

where  $v$  is a real number representing the reward,  $F$  is a formula that contains fluent constants only, and  $G$  is a formula that contains fluent constants and action constants only (no pf, no initpf constants). We extend the signature of  $Tr(D, m)$  with a set of atoms of the form (A1). We turn a utility law of the form (A16) into the  $LP^{MLN}$  rule

$$\alpha : \text{utility}(v, i+1, id) \leftarrow (i+1 : F) \wedge (i : G) \tag{A17}$$

where  $id$  is a unique number assigned to the  $LP^{MLN}$  rule and  $i \in \{0, \dots, m-1\}$ .

Given a nonnegative integer  $m$  denoting the maximum timestamp, a *pBC+* action description

$D$  with utility over multi-valued propositional signature  $\sigma$  is defined as a high-level representation of the DT-LP<sup>MLN</sup> program  $(Tr(D, m), \sigma_m^{act})$ .

We extend the definition of a probabilistic transition system as follows: A *probabilistic transition system*  $T(D)$  represented by a probabilistic action description  $D$  is a labeled directed graph such that the vertices are the states of  $D$ , and the edges are obtained from the transitions of  $D$ : for every transition  $\langle s, e, s' \rangle$  of  $D$ , an edge labeled  $e : p, u$  goes from  $s$  to  $s'$ , where  $p = Pr_{D_1}(1 : s' \mid 0 : s \wedge 0 : e)$  and  $u = E[U_{D_1}(0 : s \wedge 0 : e \wedge 1 : s')]$ . The number  $p$  is called the *transition probability* of  $\langle s, e, s' \rangle$ , denoted by  $p(s, e, s')$ , and the number  $u$  is called the *transition reward* of  $\langle s, e, s' \rangle$ , denoted by  $u(s, e, s')$ .

## Appendix B PBCPLUS2POMDP in Compositional Way

In particular, the inputs of PBCPLUS2POMDP(COMPO) include the following:

- LP<sup>MLN</sup> program  $\Pi(m)$ , parameterized with maximum timestep  $m$ , that contains LP<sup>MLN</sup> translation of fluent dynamic laws, observation dynamic laws and utility laws with no occurrence of action constant, and static laws, as well as pf constant declarations of pf constants that occur in those causal laws (see Figure 1);
- For each group of actions  $a_i \in \{a_1, \dots, a_n\}$ , an LP<sup>MLN</sup> program  $\Pi_i(m) \cup C_i(m)$ , parameterized with maximum timestep  $m$ ;  $\Pi_i(m)$  contains translation of fluent dynamic laws, observation dynamic laws and utility laws where only actions in  $a_i$  can occur in the body, as well as pf constant declarations of pf constants that occurs in those causal laws;  $C_i(m)$  contains choice rules (possibly with cardinality bounds) to generate exactly one action in the group  $a_i$ ; It is up to the user how to group the actions;
- Discount factor.

The system outputs the POMDP definition  $M(D)$ , so that  $D_m = \Pi(m) \cup \Pi_1(m) \cup \dots \cup \Pi_n(m) \cup C(m)$ , where  $C(m)$  is the choice rule with cardinality constraint to generate at most one action in  $a_1, \dots, a_n$  for each timestep  $i \in \{0, \dots, m-1\}$ . The transition probabilities, observation probabilities and reward function of  $M(D)$  are obtained by conjoining those from each of  $\Pi \cup \Pi_i \cup C_i$  ( $i \in \{1, \dots, n\}$ ).

Formally, let  $\mathbf{S}, \Omega, P_{M(D)}, O_{M(D)}, R_{M(D)}$  be the set of states, the set of observations, transition probabilities, observation probabilities and reward function of  $M(D)$ , resp. system PBCPLUS2POMDP calls LPMLN2ASP first to solve  $\Pi(0)$  to obtain  $\mathbf{S}$ , and then  $\Pi(1) \cup \Pi_i(1) \cup C_i(1)$  to obtain  $P_{M(D)}, O_{M(D)}, R_{M(D)}$  as follows:

$$P_{M(D)}(s, a, s') = P_{\Pi(1) \cup \Pi_i(1) \cup C_i(1)}(1 : s' \mid 0 : s, 0 : a)$$

$$O_{M(D)}(s, a, o) = P_{\Pi(1) \cup \Pi_i(1) \cup C_i(1)}(1 : o \mid 1 : s, 0 : a)$$

$$R_{M(D)}(s, a, s') = E[U_{\Pi(1) \cup \Pi_i(1) \cup C_i(1)}(0 : s, 0 : a, 1 : s')]$$

for each  $a \in a_i, s, s' \in \mathbf{S}$  and  $o \in \Omega$ .

### Example 1

For the dialog example, we group the actions as follows:  $\{\text{ConfirmItem}(i) \mid i \in \text{Item}\}, \{\text{ConfirmPerson}(p) \mid p \in \text{Person}\}, \{\text{ConfirmRoom}(r) \mid r \in \text{Room}\}, \{\text{WhichItem}\}, \{\text{WhichPerson}\}, \{\text{WhichRoom}\}, \{\text{Deliver}(i, p, r) \mid i \in \text{Item}, p \in \text{Person}, r \in \text{Room}\}$ .

$\Pi$  is

```

astep(0..m-1).
step(0..m).
boolean(t; f).
item(coffee; coke; cookies; burger).
person(alice; bob; carol).
room(r1; r2; r3).

% UEC
:- obs_Item(X1, I), obs_Item(X2, I), X1 != X2.
:- not obs_Item(coffee, I), not obs_Item(coke, I),
   not obs_Item(cookies, I), not obs_Item(burger, I),
   not obs_Item(na, I), step(I).
:- obs_Person(X1, I), obs_Person(X2, I), X1 != X2.
:- not obs_Person(alice, I), not obs_Person(bob, I),
   not obs_Person(carol, I), not obs_Person(na, I),
   step(I).
:- obs_Room(X1, I), obs_Room(X2, I), X1 != X2.
:- not obs_Room(r1, I), not obs_Room(r2, I),
   not obs_Room(r3, I), not obs_Room(na, I),
   step(I).
:- obs_Confirmed(X1, I), obs_Confirmed(X2, I), X1 != X2.
:- not obs_Confirmed(yes, I), not obs_Confirmed(no, I),
   not obs_Confirmed(na, I), step(I).

:- fl_ItemReq(X1, I), fl_ItemReq(X2, I), X1 != X2.
:- not fl_ItemReq(coffee, I), not fl_ItemReq(coke, I),
   not fl_ItemReq(cookies, I), not fl_ItemReq(burger, I),
   not fl_ItemReq(na, I), step(I).
:- fl_PersonReq(X1, I), fl_PersonReq(X2, I), X1 != X2.
:- not fl_PersonReq(alice, I), not fl_PersonReq(bob, I),
   not fl_PersonReq(carol, I), not fl_PersonReq(na, I),
   step(I).
:- fl_RoomReq(X1, I), fl_RoomReq(X2, I), X1 != X2.
:- not fl_RoomReq(r1, I), not fl_RoomReq(r2, I),
   not fl_RoomReq(r3, I), not fl_RoomReq(na, I),
   step(I).
:- fl_Terminated(X1, I), fl_Terminated(X2, I), X1 != X2.
:- not fl_Terminated(t, I), not fl_Terminated(f, I), step(I).

%% No two observations can occur at the same time step
:- obs_Item(It, I), obs_Person(P, I), It != na, P != na.
:- obs_Item(It, I), obs_Room(R, I), It != na, R != na.
:- obs_Item(It, I), obs_Confirmed(C, I), It != na, C != na.
:- obs_Person(P, I), obs_Room(R, I), P != na, R != na.
:- obs_Person(P, I), obs_Confirmed(C, I), P != na, C != na.
:- obs_Room(R, I), obs_Confirmed(C, I), R != na, C != na.

% Inertial Fluents
{fl_ItemReq(It, I+1)} :- fl_ItemReq(It, I), astep(I).
{fl_PersonReq(P, I+1)} :- fl_PersonReq(P, I), astep(I).
{fl_RoomReq(R, I+1)} :- fl_RoomReq(R, I), astep(I).
{fl_Terminated(B, I+1)} :- fl_Terminated(B, I), astep(I).

% Initial value of regular fluents and observation constants are exogenous

```



```

{fl_Terminated(B, 0)} :- boolean(B).
{fl_ItemReq(It, 0)} :- item(It).
{fl_PersonReq(P, 0)} :- person(P).
{fl_RoomReq(R, 0)} :- room(R).
{obs_Item(It, 0)} :- item(It).
{obs_Person(P, 0)} :- person(P).
{obs_Room(R, 0)} :- room(R).
{obs_Confirmed(yes, 0); obs_Confirmed(no, 0)}.

```

```

% By default, observation constant has na value
{obs_Item(na, I)} :- step(I).
{obs_Person(na, I)} :- step(I).
{obs_Room(na, I)} :- step(I).
{obs_Confirmed(na, I)} :- step(I).

```

$\Pi_1$  contains definition of action *Ask2ConfirmItem*:

```

% Action: ConfirmItem
:- c(It, X1, I), act_ConfirmItem(It, X2, I), X1 != X2.
:- not act_ConfirmItem(It, t, I), not act_ConfirmItem(It, f, I), item(It), astep(I).

:- pf_ConfirmWhenCorrect(X1, I), pf_ConfirmWhenCorrect(X2, I), X1 != X2.
:- not pf_ConfirmWhenCorrect(yes, I), not pf_ConfirmWhenCorrect(no, I), astep(I).
:- pf_ConfirmWhenIncorrect(X1, I), pf_ConfirmWhenIncorrect(X2, I), X1 != X2.
:- not pf_ConfirmWhenIncorrect(yes, I), not pf_ConfirmWhenIncorrect(no, I), astep(I).

@log(0.8) pf_ConfirmWhenCorrect(yes, I) :- astep(I).
@log(0.2) pf_ConfirmWhenCorrect(no, I) :- astep(I).

@log(0.2) pf_ConfirmWhenIncorrect(yes, I) :- astep(I).
@log(0.8) pf_ConfirmWhenIncorrect(no, I) :- astep(I).

```

```

obs_Confirmed(C, I+1) :- fl_ItemReq(It, I+1), fl_Terminated(f, I+1),
    act_ConfirmItem(It, t, I), pf_ConfirmWhenCorrect(C, I).
obs_Confirmed(C, I+1) :- fl_ItemReq(It, I+1), fl_Terminated(f, I+1),
    act_ConfirmItem(It1, t, I), It1 != It, pf_ConfirmWhenIncorrect(C, I).

```

```

{act_ConfirmItem(It, B, I)} :- item(It), boolean(B), astep(I).
:- not 1{act_ConfirmItem(It, t, I) : item(It)}1, astep(I).

```

Similarly,  $\Pi_2$ ,  $\Pi_3$  contains definition of actions *ConfirmPerson* and *ConfirmRoom*.

$\Pi_4$  contains definition of actions *WhichItem(t)*:

```

% Action WhichItem
:- act_WhichItem(X1, I), act_WhichItem(X2, I), X1 != X2.
:- not act_WhichItem(t, I), not act_WhichItem(f, I), astep(I).

:- pf_WhichItem(It, X1, I), pf_WhichItem(It, X2, I), X1 != X2.
:- not pf_WhichItem(It, coffee, I), not pf_WhichItem(It, coke, I),
    not pf_WhichItem(It, cookies, I), not pf_WhichItem(It, burger, I),
    item(It), astep(I).

@log(0.7) pf_WhichItem(coffee, coffee, I) :- astep(I).
@log(0.1) pf_WhichItem(coffee, coke, I) :- astep(I).
@log(0.1) pf_WhichItem(coffee, cookies, I) :- astep(I).

```

```

@log(0.1) pf_WhichItem(coffee, burger, I) :- astep(I).
@log(0.1) pf_WhichItem(coke, coffee, I) :- astep(I).
@log(0.7) pf_WhichItem(coke, coke, I) :- astep(I).
@log(0.1) pf_WhichItem(coke, cookies, I) :- astep(I).
@log(0.1) pf_WhichItem(coke, burger, I) :- astep(I).
@log(0.1) pf_WhichItem(cookies, coffee, I) :- astep(I).
@log(0.1) pf_WhichItem(cookies, coke, I) :- astep(I).
@log(0.7) pf_WhichItem(cookies, cookies, I) :- astep(I).
@log(0.1) pf_WhichItem(cookies, burger, I) :- astep(I).
@log(0.1) pf_WhichItem(burger, coffee, I) :- astep(I).
@log(0.1) pf_WhichItem(burger, coke, I) :- astep(I).
@log(0.1) pf_WhichItem(burger, cookies, I) :- astep(I).
@log(0.7) pf_WhichItem(burger, burger, I) :- astep(I).

obs_Item(It1, I+1) :- fl_ItemReq(It, I+1), fl_Terminated(f, I+1),
                    act_WhichItem(t, I), pf_WhichItem(It, It1, I).

%{act_WhichItem(B, I)} :- boolean(B), astep(I).
act_WhichItem(t, I) :- astep(I).

```

Similarly,  $\Pi_5$  and  $\Pi_6$  contain definitions of actions *WhichPerson(t)* and *WhichRoom(t)*.

$\Pi_7$  contains definitions of action *Deliver(i, p, r)*:

```

% Action: Deliver
:- act_Deliver(It, P, R, X1, I), act_Deliver(It, P, R, X2, I), X1 != X2.
:- not act_Deliver(It, P, R, t, I), not act_Deliver(It, P, R, f, I), item(It), person(P), room(R), as
).

utility(1, I+1, It) :- fl_ItemReq(It, I+1), act_Deliver(It, P, R, t, I), fl_Terminated(f, I).
utility(1, I+1, P) :- fl_PersonReq(P, I+1), act_Deliver(It, P, R, t, I), fl_Terminated(f, I).
utility(1, I+1, R) :- fl_RoomReq(R, I+1), act_Deliver(It, P, R, t, I), fl_Terminated(f, I).

fl_Terminated(t, I+1) :- act_Deliver(It, P, R, t, I).

{act_Deliver(It, P, R, B, I)} :- item(It), person(P), room(R), boolean(B), astep(I).
:- not 1{act_Deliver(It, P, R, t, I) : item(It), person(P), room(R)}1, astep(I).

```

$\Pi_8$  contains definitions of no-action:

```
act_noact(I) :- astep(I).
```

With this way of grouping actions, system PBCPLUS2POMDP(COMPO) can generate POMDP for this example with  $\sim 5$  minutes.

## Appendix C Tiger Example

### Example 2

**(Two Tigers Example).** Consider a variant of the well-known tiger example extended with two tigers. Each of the three doors has either a tiger or a prize behind. The agent can open either of the three doors. The agent can also listen to get a better idea of where the tiger is. Listening yields the correct information about where each of the two tigers is with probability 0.85. This example can be represented in the extended  $p\mathcal{BC}+$  as follows:

Notation:  $l, l_1, l_2, l_3$  range over Left, Middle, Right,  $y$  ranges over Tiger1, Tiger2

Observation constants:	Domains:
$TigerPositionObserved(y)$	{Left, Middle, Right, NA}
Regular fluent constants:	Domains:
$TigerPosition(y)$	{Left, Middle, Right}
Action constants:	Domains:
$Listen$	Boolean
$OpenDoor(l)$	Boolean
Pf constants:	Domains:
$Pf\_Listen$	Boolean
$Pf\_FailedListen(y)$	{Left, Middle, Right}

---

A reward of 10 is obtained for opening the door with no tiger behind.

**reward** 10 **if**  $TigerPosition(Tiger1) = l_1 \wedge TigerPosition(Tiger2) = l_2$  **after**  $OpenDoor(l_3)$   
 $(l_1 \neq l_3, l_2 \neq l_3)$ .

A penalty of 100 is imposed on opening a door with a tiger behind.

**reward** -100 **if**  $TigerPosition(y) = l$  **after**  $OpenDoor(l)$ .

Executing the action  $Listen$  has a small penalty of 1.

**reward** -1 **if**  $\top$  **after**  $Listen$ .

Two tigers cannot be in the same position.

**caused**  $\perp$  **if**  $TigerPosition(Tiger1) = l \wedge TigerPosition(Tiger2) = l$ .

Successful listening reveals the positions of the two tigers.

**observed**  $TigerPositionObserved(y) = l$  **if**  $TigerPosition(y) = l$  **after**  $Listen \wedge Pf\_Listen$ .

Failed listening yields a random position for each tiger.

**caused**  $Pf\_FailedListen(y) = \{Left : \frac{1}{3}, Middle : \frac{1}{3}, Right : \frac{1}{3}\}$ ,

**observed**  $TigerPositionObserved(y) = l$  **if**  $\top$  **after**  $Listen \wedge \sim Pf\_Listen \wedge Pf\_FailedListen(y) = l$ .

The positions of tigers observe the commonsense law of inertia.

**inertial**  $TigerPosition(y)$ .

The action  $Listen$  has a success rate of 0.85.

**caused**  $Pf\_Listen = \{TRUE : 0.85, FALSE : 0.15\}$ .