Künstl Intell
DOI 10.1007/s13218-016-0455-7

TECHNICAL CONTRIBUTION

Reasoning about Imprecise Beliefs in Multi-Agent Systems with PDT Logic

Karsten Martiny[1] · **Ralf Möller**[1]

Received: 29 January 2016 / Accepted: 28 September 2016

Abstract We present Probabilistic Doxastic Temporal (PDT) Logic, a formalism to represent and reason about probabilistic beliefs and their finite temporal evolution in multi-agent systems. This formalism enables the quantification of agents' beliefs through probability intervals and incorporates an explicit notion of time. In this work, we give an overview of recent contributions on PDT Logic. After describing the syntax and semantics of this formalism, we show that two alternative representation forms are available to model problems in PDT Logic. Furthermore, we outline how abductive reasoning can be performed in PDT Logic and how this formalism can be extended to infinite time frames.

Keywords Knowledge representation · Belief updates · Imprecise probabilities

1 Introduction

Logical analysis of knowledge and belief has been an active topic of research in diverse fields such as philosophy [10], economics [1], and computer science [7].

In most realistic scenarios, an agent has only incomplete and inaccurate information about the actual state of the world, and thus considers several different worlds as actually being possible. As it receives new information, it has to update its beliefs about possible worlds. These updates can for example result in regarding some worlds as impossible or judging some worlds to be more likely than before.

When multiple agents are involved in such a setting, an agent may not only have varying beliefs regarding the facts of the actual world, but also regarding the beliefs of other agents. In many scenarios, the actions of one agent will not only depend on its belief in ontic facts (i.e., facts of the actual world), but also on its beliefs in some other agent's beliefs.

To formalize reasoning about such beliefs in multi-agent settings, we have developed Probabilistic Doxastic Temporal (PDT) Logic, as introduced in [14]. This formalism merges established concepts from epistemic logic with recent contributions from temporal logic, namely Annotated Probabilistic Temporal (APT) Logic [20], and enables a representation of uncertain knowledge with imprecise probabilities (i.e., using probability intervals instead of single values). This representation yields two main advantages. On the one hand, the use of probability intervals significantly eases the task of formally representing existing knowledge of a human domain expert. In most cases, a domain expert can give reasonable probability estimates of her knowledge, but will inevitably fail at giving precise numerical values on these probabilities. Consider for instance a weather forecast: most people find it easy to give coarse probabilistic quantifications such as "the chance of rain is high", while virtually nobody could quantify this through an exact numerical value. Employing exact numerical values in a formal representation would then inevitably introduce errors in the probability model. Thus, the use of probability intervals provides means to express probabilistic knowledge as precisely as possible without enforcing unrealistic precision. On the other hand,

✉ Karsten Martiny
karsten.martiny@uni-luebeck.de

Ralf Möller
moeller@uni-luebeck.de

[1] Institute of Information Systems, Universität zu Lübeck, Lübeck, Germany

Published online: 13 October 2016

there are many scenarios where probabilities are simply unavailable, while bounds on these values may be known. To illustrate this, consider a modified version of the Three Prisoners Puzzle [8, 9]:

Example 1 (Modified Three Prisoners Puzzle, adapted from [9]) Of three prisoners a, b, and c, two are to be executed, but a does not know which. The probability that a will be executed is 2/3, while the respective probabilities for b and c being executed are unknown. a says to the jailer, "since either b or c is certainly going to be executed, you will give me no information about my own chances if you give me the name of one man, either b or c, who is going to be executed."

Now, it is easy to see that a's chance of surviving is 1/3, while no such a probability value can be given for the survival of b or c. However, it follows immediately from a's survival chance that neither b's or c's chance of survival can exceed 2/3. Using imprecise probabilities, we can correctly specify probability intervals [0, 2/3] for b and c, while any attempt of modeling precise values would fail. We will return to this example below and give a formal representation after summarizing the syntax and semantics of PDT Logic in the next section.

2 Related Work

Early research on epistemic logic culminated in the influential work [7], which provides a unified presentation of various preceding contributions on epistemic logic with "sharp" knowledge, i.e., it does not consider uncertainty. Several works have extended epistemic logic to represent dynamically changing knowledge, i.e., evolutions of knowledge are represented through the step-by-step results of certain actions, while no explicit model of time is given. The first formal analysis of this approach has been presented in [19]. Subsequent works—e.g., [2, 3]—have generalized this approach to incorporate a variety of complex epistemic actions. A thorough treatment of Dynamic Epistemic Logic can be found in [5].

An alternative approach of modeling the evolution of knowledge is to combine epistemic logic with some temporal system. One example for this are the interpreted systems from [7], where time is represented through sequences of global states. A similar approach is Epistemic Temporal Logic (ETL) [18], where situations are represented through sets of histories. The temporal model employed in PDT Logic is closely related to ETL, with the additional introduction of frequency functions (as described below), which—compared to existing temporal epistemic logics—enable reasoning about a wider range of temporal relationships.

Various works have augmented epistemic logics with probabilities to represent uncertainty. An early influential work in this area is [6], which combines epistemic logic with lower-bound probability operators. Kooi [11] limits the approach of [6] to measurable sets and combines it with dynamic epistemic logic to obtain a dynamic version of probabilistic epistemic logic. This work is extended in [4] to analyze the results of various epistemic actions. In contrast to these works, PDT Logic uses an explicit representation of imprecise probabilities (i.e., it uses a probability operator with both lower and upper bounds). In most related approaches, there is a difference between belief with probability 1 and knowledge, while these concepts are unified in PDT Logic. To illustrate this, consider repeatedly flipping a coin: the probability that it will eventually show head is 1 for an infinite number of repetitions, while nobody can know that it will ever show head. As PDT Logic is restricted to finite domains, these concepts are equivalent in our work, and thus, for beliefs with probability 1, our formalism corresponds to classic epistemic logic.

3 PDT Logic Programs: Syntax and Semantics

We now summarize the syntax and semantics of PDT Logic, as introduced in [14].

3.1 Syntax

We assume the existence of a function-free first order logic language $\mathcal{L}$ with finite sets of constant symbols $\mathcal{L}_{cons}$ and predicate symbols $\mathcal{L}_{pred}$, and an infinite set of variable symbols $\mathcal{L}_{var}$. Every predicate symbol $p \in \mathcal{L}_{pred}$ has an *arity*. A *term* is any member of the set $\mathcal{L}_{cons} \cup \mathcal{L}_{var}$. A term is called a *ground term* if it is a member of $\mathcal{L}_{cons}$. If $t_1, .., t_n$ are (ground) terms, and p is a predicate symbol in $\mathcal{L}_{pred}$ with arity n, then $p(t_1, \ldots, t_n)$ is a (ground) atom. If a is a (ground) atom, then a and $\neg a$ are (ground) *literals*. The former is called a *positive literal*, the latter is called a *negative literal*. The set of all ground literals is denoted by $\mathcal{L}_{lit}$. As usual, $\mathcal{B}$ denotes the Herbrand Base of $\mathcal{L}$.

Time is modeled as a set τ of discrete time points $\tau = \{1, \ldots, t_{max}\}$. The set of agents is denoted by $\mathcal{A}$. To describe what agents observe, we define observation atoms as follows:

Definition 1 (*Observation atoms*) For any non-empty group of agents $\mathcal{G} \subseteq \mathcal{A}$ and ground literal $l \in \mathcal{L}_{lit}$, $Obs_{\mathcal{G}}(l)$ is an *observation atom*. The set of all observation atoms is denoted by $\mathcal{L}_{obs}$.

Both atoms and observation atoms are formulae. If F and G are formulae, then $F \wedge G$, $F \vee G$, and $\neg F$ are formulae.

Intuitively, the meaning of a statement of the form $Obs_{\mathcal{G}}(l)$ is that all agents in the group $\mathcal{G}$ observe that the fact l holds. Since l may be a negative literal, we can explicitly specify observations of certain facts being false (such as "it is not raining"). Note that the formal concept of observations is not limited to express passive acts of observing facts, but can instead be used to model a wide range of actions. In the line of [2], observations can be viewed as the effects of private group announcements of a fact l to a group $\mathcal{G}$ (i.e., l becomes common knowledge *within* $\mathcal{G}$, while all agents *outside* of $\mathcal{G}$ remain entirely oblivious of the observation): it represents an epistemic action, i.e., it alters the belief states of all agents (as formally defined below) in $\mathcal{G}$, but does not influence the ontic facts of the respective world.

To express temporal relationships, we define temporal rules following the approach of APT rules from [20]. The definition of temporal rules already relies on the concept of frequency functions, even though these are defined in the next section. We still introduce temporal rules now to enable a clearly separated presentation of syntax and semantics of PDT Logic. For now, it suffices to note that frequency functions provide information about temporal connections between events.

Definition 2 (*Temporal rules*) Let F, G be two formulae, Δt a time interval, and fr a frequency function (as defined below in Sect. 3.2.5). Then $r^{\mathrm{fr}}_{\Delta t}(F, G)$ is called a temporal rule.

The meaning of such an expression is to be understood as "F is followed by G in Δt time units w.r.t. fr".

Now, we can define the belief operator $B^{l,u}_{i,t'}$ to express agents' beliefs. Intuitively, $B^{l,u}_{i,t'}(\varphi)$ means that at time t', agent i believes that some fact φ is true with a probability $p \in [\ell, u]$. Particularly, the intuitive meaning of belief in a temporal rule is that agent i believes that G will hold according to $r^{\mathrm{fr}}_{\Delta t}(F, G)$, given that F holds at some time point. We call the probability interval $[\ell, u]$ the *quantification* of agent i's belief. We use F_t to denote that formula F holds at time t and, accordingly, $Obs_{\mathcal{G}}(l)_t$ to denote that an observation $Obs_{\mathcal{G}}(l)$ occurs at time t. We call these expressions time-stamped formulae and time-stamped observation atoms, respectively.

Definition 3 (*Belief formulae*) Let i be an agent, t' a time point, and $[\ell, u] \subseteq [0, 1]$. Then, *belief formulae* are inductively defined as follows:

1. If F is a formula and t is a time point, then $B^{l,u}_{i,t'}(F_t)$ is a belief formula.
2. If $r^{\mathrm{fr}}_{\Delta t}(F, G)$ is a temporal rule, then $B^{l,u}_{i,t'}(r^{\mathrm{fr}}_{\Delta t}(F, G))$ is a belief formula.
3. If $\mathscr{F}$ and $\mathscr{G}$ are belief formulae, then so are $B^{l,u}_{i,t'}(\mathscr{F})$, $\mathscr{F} \wedge \mathscr{G}$, $\mathscr{F} \vee \mathscr{G}$, and $\neg\mathscr{F}$.

We use script fonts (e.g., $\mathscr{F}$) to distinguish belief formulae from standard formulae.

3.2 Semantics

In this section, we summarize the formal semantics for PDT Logic that captures the intuitions explained above. To ease understanding of the presentation, we use the example from [14], which we will return to repeatedly when introducing the various concepts of the semantics.

Example 2 (Trains [14]) Let Alice and Bob be two agents living in two different cities C_A and C_B, respectively. Suppose that Alice wants to take a train to visit Bob. Unfortunately, there is no direct connection between cities C_A and C_B, so Alice has to change trains at a third city C_C. We assume that train T_1 connects C_A and C_C, and train T_2 connects C_C and C_B. Both trains usually require 2 time units for their trip, but they might be running late and arrive one time unit later than scheduled. Alice requires one time unit to change trains at city C_C. If T_1 runs on time, she has a direct connection to T_2, otherwise she has to wait for two time units until the next train T_2 leaves at city C_C. If a train is running late, she can call Bob to let him know. These calls can be modeled as shared observations between Alice and Bob. For instance, if Alice wants to tell Bob that train T_1 is running late (i.e., T_1 does not arrive at C_C at the expected time, this can be modeled as $Obs_{\{AB\}}(\neg at(T_1, C_C))$ at the expected arrival time.

3.2.1 Possible Worlds

Ontic facts and according observations (e.g., as described in the above example) form *worlds* (or *states* in the terminology of [7]). A world ω consists of a set of ground atoms and a set of observation atoms, i.e., $\omega \in 2^{\mathcal{B} \cup \mathcal{L}_{obs}}$. We use $a \in \omega$ and $Obs_{\mathcal{G}}(l) \in \omega$ to denote that an atom a (resp. observation atom $Obs_{\mathcal{G}}(l)$) holds in world ω. Since agents can only observe facts that actually hold in the respective world, we can define admissibility conditions of worlds w.r.t. the set of observations:

Definition 4 (*Admissible worlds*) A world ω is admissible, iff for every observation atom $Obs_{\mathcal{G}}(l) \in \omega$

1. The observed fact holds, i.e., $x \in w$ if l is a positive literal x, and $x \notin w$ if l is a negative literal $\neg x$, and
2. For every subgroup $\mathcal{G}' \subset \mathcal{G}$, $Obs_{\mathcal{G}'}(l) \in \omega$.

We use $adm(\omega)$ to denote that a world ω is admissible.

The set of all possible worlds is denoted by Ω and the set of admissible worlds by $\hat{\Omega}$.

Example 3 (Trains continued) For Example 2, we have ground terms A, B, C_A, C_B, C_C, T_1, and T_2, representing Alice, Bob, three cities, and two trains. Furthermore, we have atoms $on(x, y)$ indicating that person y is on train x, and $at(x, z)$ indicating that train x is at city z. Finally, we have observation atoms of the kind $Obs_{\mathcal{G}}(at(x,z))$, indicating that the agents in $\mathcal{G}$observe that train x is at station z. A possible world can for example be $\omega_1 = \{at(T_1, C_A), on(T_1, A), Obs_A(at(T_1, A))\}$, indicating that train T_1 is at city C_A and A has boarded that train.

We define satisfaction of a ground formula F by a world ω, in the usual way:

Definition 5 (*Satisfaction of ground formulae*) Let F, F', F'' be ground formulae and ω a world. Then, F is satisfied by ω (denoted $\omega \models F$) if and only if:

- Case: $F = a$ for some ground atom a, then $a \in \omega$.
- Case: $F = \neg F'$ for some ground formula F', then $\omega \not\models F'$.
- Case: $F = F' \wedge F''$ for formulae F' and F'', then $\omega \models F'$ and $\omega \models F''$.
- Case: $F = F' \vee F''$ for formulae F' and F'', then $\omega \models F'$ or $\omega \models F''$.

3.2.2 Threads

We use the definition of *threads* from [20]:

Definition 6 (*Thread*) A *thread Th*is a mapping from the set of time points τ to the set of admissible worlds: $Th : \tau \rightarrow \hat{\Omega}$.

Thus, a thread is a sequence of worlds and $Th(i)$ identifies the actual world at time i according to thread Th. The set of all possible threads (i.e., all possible sequences constructible from τ and $\hat{\Omega}$) is denoted by $\mathcal{T}$. We refrain from directly working with $\mathcal{T}$, and instead assume that any meaningful problem specification gives information about possible temporal evolutions of the system. We use $\hat{\mathcal{T}}$to represent this set of relevant possible threads. For notational convenience, we assume that there is an additional prior world $Th(0)$ for every thread.

We assume that the system is synchronous, i.e., the agents have a global clock. Thus, even if an agent does not observe anything in world $Th(t)$, it is still aware of time passing and can therefore distinguish between worlds $Th(t)$ and $Th(t-1)$.

Example 4 (Trains continued) The description from Example 2 yields the set of possible threads $\hat{\mathcal{T}}$ depicted in Fig. 1.

3.2.3 Kripke Structures

With the definition of threads, we can use a slightly modified version of Kripke structures [12]. As usual, we define a Kripke structure as a tuple $\langle \hat{\Omega}, \mathcal{K}_1, \ldots, \mathcal{K}_n \rangle$, with the set of admissible worlds $\hat{\Omega}$ and binary relations $\mathcal{K}_i$on

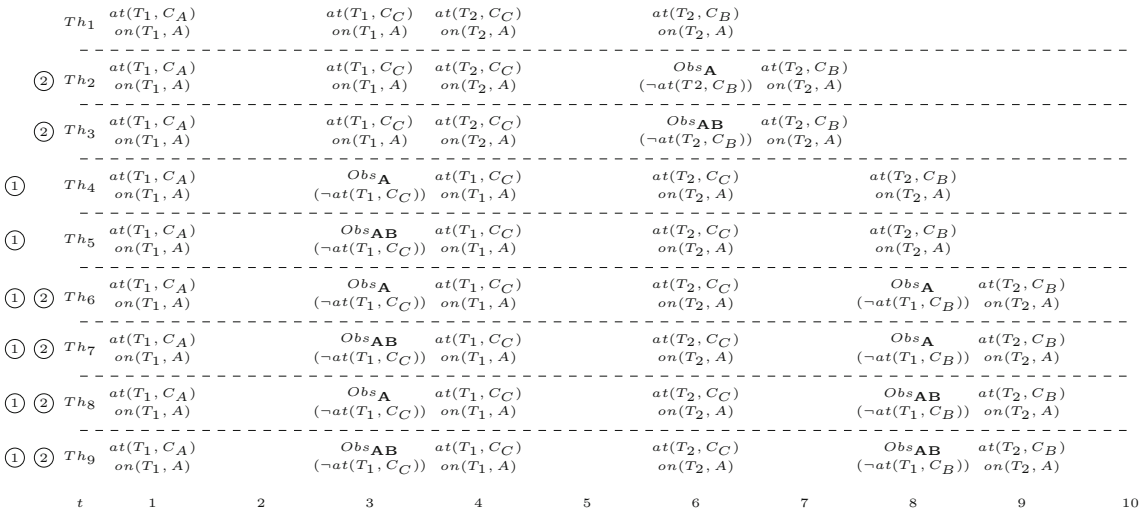

Fig. 1 Visualization of the possible threads Th_k from Example 2: $at(T_i, C_j)$ denotes that train T_i is currently at city C_j, $on(T_i, A)$ that Alice is currently on train T_i, and $Obs_{AG}(\neg at(T_i, C_j))$ denotes a call from Alice to inform Bob that train T_i is currently not at city C_j. For the sake of simplicity, facts irrelevant to the analysis (such as $on(T_i, A)$ for time points *2* and *5*) are omitted from the presentation. Note that if a train is running late (the respective threads are marked with according *circles*), there are always two possible threads: one where only *A* observes this and one where both share the observation. For an easier distinction, we have marked the according group of an observation with *boldface* indices

$\hat{\Omega}$ for every agent $i \in \mathcal{A}$. Thus, the Kripke structure for agent i at world ω is defined as

$$\mathcal{K}_i(\omega) = \{\omega' : (\omega, \omega') \in \mathcal{K}_i\} \tag{1}$$

Intuitively, $(\omega, \omega') \in \mathcal{K}_i$ specifies that in world ω , agent i considers ω' as a possible world.

We initialize the Kripke structure such that the set of possible worlds are the worlds that occur at time $t = 1$ in some thread Th':

$$\forall Th \in \hat{\mathcal{T}} : \mathcal{K}_i(Th(0)) = \bigcup_{Th' \in \hat{\mathcal{T}}} \{Th'(1)\}, \quad i = 1, \ldots, n \tag{2}$$

With the evolution of time, each agent can eliminate the worlds that do not comply with its respective observations. Through the elimination of worlds, an agent will also reduce the set of threads it considers possible (if — due to some observation — a world ω is considered impossible at a time point t, then all threads Th with $Th(t) = \omega$ are considered impossible). We assume that agents have perfect recall and therefore will not consider some thread possible again if it was considered impossible at one point. Thus, $\mathcal{K}_i$is updated w.r.t. the agent's respective observations, such that it considers all threads possible that both comply with its current observations and were considered possible at the previous time point:

$$\begin{aligned} &\mathcal{K}_i(Th(t)) := \{Th'(t) : (Th'(t-1) \in \mathcal{K}_i \\ &(Th(t-1)) \wedge \{Obs_{\mathcal{G}}(l) \in Th(t) : i \in \mathcal{G}\} \\ &= \{Obs_{\mathcal{G}}(l) \in Th'(t) : i \in \mathcal{G}\})\} \end{aligned} \tag{3}$$

Example 5 (Trains continued) From Fig. 1, we obtain that at time 1, the only possible world is $\{\{at(T_1, C_A), on(T_1, A)\}\}$, which is contained in all possible threads. Thus, $\mathcal{K}_i(Th_j(1))$ contains exactly this world for all agents i and threads j. Consequently, both agents consider all threads as possible at time 1.

Now, assume that time evolves for two steps and the actual thread is Th_4 (i.e., train T_1 is running late, but A does not inform B about this). Both agents will update their possibility relations accordingly, yielding

$$\mathcal{K}_A(Th_4(3)) = \{\{Obs_{\{A\}}(\neg at(T_1, C_C))\}\}$$

and

$$\mathcal{K}_B(Th_4(3)) = \{\{at(T_1, C_C), on(T_1, A)\}, \quad \{Obs_{\{A\}}(\neg at(T_1, C_C))\}\},$$

i.e., A knows that T_1 is not on time, while B is unaware of this.

3.2.4 Subjective Posterior Temporal Probabilistic Interpretations

Each agent has probabilistic beliefs about the expected evolution of the world over time. This is expressed through subjective temporal probabilistic interpretations:

Definition 7 (*Subjective posterior probabilistic temporal interpretation*) Given a set of possible threads $\hat{\mathcal{T}}$, some thread $\mathring{T}h \in \hat{\mathcal{T}}$, a time point $t' > 0$ and an agent i, $\mathcal{I}_{i,t'}^{\mathring{T}h} : \hat{\mathcal{T}} \rightarrow [0, 1]$ specifies the *subjective posterior probabilistic temporal interpretation* from agent i's point of view at time t' in thread $\mathring{T}h$, i.e., a probability distribution over all possible threads: $\sum_{Th \in \hat{\mathcal{T}}} \mathcal{I}_{i,t'}^{\mathring{T}h}(Th) = 1$. We call $\mathring{T}h$the *point of view (pov) thread* of interpretation $\mathcal{I}_{i,t'}^{\mathring{T}h}$.

The prior probabilities of each agent for all threads are then given by $\mathcal{I}_{i,0}^{\hat{T}}h(Th)$. Since all threads are indistinguishable a priori, there is only a *single* prior distribution for each agent. Furthermore, in order to be able to reason about nested beliefs (as discussed below), we assume that the prior probability assessments of all agents are commonly known (i.e., all agents know how all other agents assess the prior probabilities of each thread). This in turn requires that all agents have exactly the same prior probability assessment over all possible threads: if two agents have different, but commonly known prior probability assessments, we essentially have an instance of Aumann's well-known problem of "agreeing to disagree" [1]. Intuitively, if differing priors are commonly known, it is common knowledge that (at least) one of the agents is at fault and should revise its probability assessments. As a result, we have only one prior probability distribution which is the same from all viewpoints, denoted by $\mathcal{I}$.

Example 6 *(Trains continued)* A meaningful interpretation is

$$\mathcal{I} = (0.7 \quad 0.02 \quad 0.09 \quad 0.02 \quad 0.09 \quad 0.01 \quad 0.02 \quad 0.02 \quad 0.03),$$

which assigns the highest probability to Th_1 (no train running late), lower probabilities to the threads where one train is running late and A informs B (Th_3 and Th_5), even lower probabilities to the events that either both trains are running late and A informs B (Th_7, Th_8, and Th_9) or that one train is running late and A does not inform B (Th_2 and Th_4), and lowest probability to the thread where both trains are running late and A does not inform B (Th_6).

Even though we only have a single prior probability distribution over the set of possible threads, it is still necessary to distinguish the viewpoints of different agents in different threads, as the following definition of interpretation updates shows.

Definition 8 (*Interpretation update*) Let i be an agent, t' a time point, and $\mathring{T}h$a pov thread. Then, if the system is actually in thread $\mathring{T}h$at time t', agent i's probabilistic interpretation over the set of possible threads is given by the update rule:

$$\mathcal{I}^{\mathring{Th}}_{i,t'}(Th) = \begin{cases} \frac{1}{\alpha^{\mathring{Th}}_{i,t'}} \cdot \mathcal{I}^{\mathring{Th}}_{i,t'-1}(Th) & \text{if Th (t}') \in \mathcal{K}_{\text{i}} \text{ (Th (t}')) \\ 0 & \text{if Th (t}') \notin \mathcal{K}_{\text{i}} \text{ (Th (t}')) \end{cases} \tag{4}$$

with $\frac{1}{\alpha^{\mathring{Th}}_{i,t'}}$ being a normalization factor to ensure that all interpretations sum to one.

Essentially, the update rule assigns all impossible threads a probability of zero and scales the probabilities of the remaining threads such that they are proportional to the probabilities of the previous time point. With a given prior probability distribution $\mathcal{I}$ over the set of possible threads, the subjective posterior probabilities $\mathcal{I}^{\mathring{Th}}_{i,t'}$ in a specific pov thread $\mathring{Th}$ for all agents i and all time points t' are induced by the respective observations contained in $\mathring{Th}$. We use $\mathcal{I}^{\mathring{Th}}$ to denote the set of all subjective posterior interpretations $\mathcal{I}^{\mathring{Th}}_{i,t'}$ induced in pov thread $\mathring{Th}$.

Example 7 (Trains continued) Applying the update rule from (4) to the situation described in Example 5, with $\mathcal{I}$as given in Example 6, yields the updated interpretation for A:

$$\mathcal{I}^{\mathring{Th}_4}_{A,3} = (0 \quad 0 \quad 0 \quad 0.4 \quad 0 \quad 0.2 \quad 0 \quad 0.4 \quad 0) \tag{5}$$

i.e., A considers exactly those threads possible, where the train is running late and she does not inform B (threads Th_4, Th_6, and Th_8). Due to the lack of any new information, B can only eliminate the situations where A does inform him about being late, and thus B's interpretation is updated to:

$$\mathcal{I}^{\mathring{Th}_4}_{B,3} \approx (.814 \quad .023 \quad .105 \quad .023 \quad 0 \quad .0120 \quad .023 \quad 0). \tag{6}$$

3.2.5 Frequency Functions

To represent temporal relationships within threads, we adapt the concept of *frequency functions* as introduced in [20]. Frequency functions provide a flexible way of representing temporal relations between the occurrence of specific events. To illustrate the motivation behind using frequency functions, consider the exemplary thread Th depicted in Fig. 2. In this thread, one of the events F or G occurs at every time point from $t = 1$ to $t = 8$. As discussed in [20], there are multiple ways of characterizing temporal relationships between the events F and G: For instance, one might specify how often the event F is followed by the event G in, say, exactly 2 time points. According to Fig. 2, this happens in one out of four occurrences of F in Th. It might prove meaningful to exclude the final occurrence of F in $Th$when determining this frequency, because naturally an occurrence of F at t_{max}cannot be followed by a subsequent occurrence of

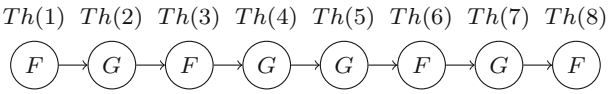

Fig. 2 Example thread Th with $\tau = \{1, \ldots, 8\}$, adopted from [20]. This figure shows each world that satisfies formula F or formula G

G. Excluding the final occurrence of F would yield one out of three for the desired frequency. Alternatively, one could also specify how often F is followed by G within the next two time points. For the exemplary thread from Fig. 2, this would produce frequencies of 1 and 0.75 respectively, again depending on whether the final occurrence of F is included.

This example illustrates already four different possible definitions of temporal relations between events. To maintain flexibility in expressing temporal relations, we do not commit to specific definitions in PDT Logic, but instead we adapt an axiomatic definition of frequency functions:

Definition 9 (*Frequency functions, adapted from* [20]) Let Th be a thread, F and G be ground formulae, and $\Delta t \geq 0$ be an integer. A *frequency function* frmaps quadruples of the form $(Th, F, G, \Delta t)$ to [0, 1] such that the following axioms hold:

(FF1) If G is a tautology, then $\mathsf{fr}(Th, F, G, \Delta t) = 1$.
(FF2) If F is a tautology and G is a contradiction, then $\mathsf{fr}(Th, F, G, \Delta t) = 0$.
(FF3) If F is a contradiction, $\mathsf{fr}(Th, F, G, \Delta t) = 1$.
(FF4) If G is not a tautology, and either F or $\neg G$ is not a tautology, and F is not a contradiction, then there exist threads Th_1, $Th_2 \in \mathcal{T}$ such that $\mathsf{fr}(Th_1, F, G, \Delta t) = 0$ and $\mathsf{fr}(Th_2, F, G, \Delta t) = 1$.

Axioms (FF1) to (FF3) ensure that frequency functions behave as temporal implications with premise F and conclusion G. Axiom (FF4) enforces non-trivial frequency functions by requiring that in all cases not covered by the first three axioms, there must be at least one thread that perfectly contradicts and one that perfectly satisfies the conditional, respectively.

To illustrate the concept of frequency functions, we now present adapted formal definitions for point and existential frequency functions from [20] that represent the informal descriptions of frequencies from above:

The point frequency function *pfr* expresses how frequently some event F is followed by another event G in *exactly* Δt time units:

$$pfr(Th, F, G, \Delta t) = \frac{|\{t : Th(t) \models F \wedge Th(t + \Delta t) \models G\}|}{|\{t : (t \leq t_{max} - \Delta t) \wedge Th(t) \models F\}|} \tag{7}$$

If the denominator is zero, we define *pfr* to be 1.

The existential frequency function *efr* expresses how frequently some event F is followed by another event G *within* the next Δt time units:

$$efr(Th, F, G, \Delta t) = \frac{efn(Th, F, G, \Delta t, 0, t_{max})}{efd(Th, F, G, \Delta t)} \tag{8}$$

$$\text{with } efn(\mathrm{Th}, \mathrm{F}, \mathrm{G}, \Delta t, t_1, t_2) = |\{t : (t_1 < t \leq t_2) \wedge \mathrm{Th}(t) \models \mathrm{F} \wedge \exists t' \in [t, \min(t_2, t + \Delta t)]\ (Th(t') \models G)\}|$$

$$\text{and } efd(\mathrm{Th}, \mathrm{F}, \mathrm{G}, \Delta t) = |\{t : (t \leq t_{\max} - \Delta t) \wedge \mathrm{Th}(t) \models \mathrm{F}\}| + efn(Th, F, G, \Delta t, t_{max} - \Delta t, t_{max}),$$

3.2.6 Semantics of the Belief Operator

Now, with the definitions of subjective posterior probabilistic temporal interpretations and the introduction of frequency functions, we can build upon the definitions from [20] for the satisfiability of interpretations to provide a formal semantics for the belief operators defined in Sect. 3.1:

Definition 10 (*Belief Semantics*) Let i be an agent and $\mathring{\mathcal{I}}_{i,t'}^{\mathring{T}h}$ be agent i's interpretation at time t' in pov thread $\mathring{T}h$. Then, it follows from this interpretation that agent i believes at time t' with a probability in the range $[\ell, u]$ that

1. (Belief in ground formulae)

A formula F holds at time t (denoted by $\mathring{\mathcal{I}}_{i,t'}^{\mathring{T}h} \models B_{i,t'}^{l,u}(F_t)$) iff:

$$\ell \leq \sum_{Th \in \hat{\mathcal{T}}, Th(t) \models F} \mathring{\mathcal{I}}_{i,t'}^{\mathring{T}h}(Th) \leq u. \tag{9}$$

2. (Belief in rules)

A temporal rule $r_{\Delta t}^{\mathrm{fr}}(F, G)$ holds (denoted by $\mathring{\mathcal{I}}_{i,t'}^{\mathring{T}h} \models B_{i,t'}^{l,u}(r_{\Delta t}^{\mathrm{fr}}(F, G))$) iff:

$$\ell \leq \sum_{Th \in \mathcal{T}^{\wedge}} \mathring{\mathcal{I}}_{i,t'}^{\mathring{T}h}(Th) \cdot \mathsf{fr}(Th, F, G, \Delta t) \leq u. \tag{10}$$

3. (Nested beliefs)

A belief $B_{j,t}^{l_j,u_j}(\varphi)$ of some other agent j holds at time t' (denoted by

$\mathring{\mathcal{I}}_{i,t'}^{\mathring{T}h} \models B_{i,t'}^{l,u}(B_{j,t}^{l_j,u_j}(\varphi))$) iff:

$$\ell \leq \sum_{\substack{Th \in \hat{\mathcal{T}} \\ \mathcal{I}_{j,t}^{Th} \models B_{j,t}^{l_j,u_j}(\varphi)}} \mathring{\mathcal{I}}_{i,t'}^{\mathring{T}h}(Th) \leq u. \tag{11}$$

Note that agent i does not know the actual beliefs of agent j. However, due to the assumption of common and equal priors, agent i is able to reason about agent j's hypothetical interpretation updates given that the system is in a specific thread. Thus, agent i is able to compute (11) without knowing j's exact beliefs.

Example 8 (Trains continued) We can use a point frequency function to express beliefs about the punctuality of trains. Assume that both A and B judge the probability of a train running late (i.e., arriving after 3 instead of 2 time units, expressed through the temporal rule r_3^{pfr}) as being at most 0.4. This yields the following belief formulae

$$\begin{matrix} B_{i,0}^{0,0.4}\left(r_3^{pfr}(at(T_1, C_A), at(T_1, C_C))\right) \\ B_{i,0}^{0,0.4}\left(r_3^{pfr}(at(T_2, C_C), at(T_2, C_B))\right) \end{matrix}, \quad i \in \{A, B\}.$$

One can easily verify that these formulae are satisfied by the interpretation given in Example 6.

To illustrate the evolution of beliefs, we finish the example with an analysis of expected arrival times.

Example 9 (Trains continued) From the interpretation given in Example 6, we can infer that Bob (and of course Alice, too) can safely assume at time 1 that Alice will arrive at time 8 at the latest (i.e., the actual thread is one of $Th_1, \ldots, Th_5$) with a probability in the range [0.9, 1] because from Definition 10 we obtain that the following belief holds for $t = 1$:

$$\mathscr{B}_{Bob,t} \equiv B_{B,t}^{0.9,1}\left(r_8^{efr}(on(T_1, A), (at(T_2, C_B) \wedge on(T_2, A))\right). \tag{12}$$

Now, consider the previously described situation, where T_1 is running late and A does not inform B about it. This leads to the updated interpretations given in (5) and (6). These updates lead to a significant divergence in the belief of the expected arrival time: Alice's belief exhibits a drastically reduced certainty and changes to

$$B_{A,3}^{0.4,1}\left(r_8^{efr}(on(T_1, A), (at(T_2, C_B) \wedge on(T_2, A))\right), \tag{13}$$

while Bob's previous belief (12) remains valid.

Even though Alice's beliefs have changed significantly, she is aware that Bob maintains beliefs conflicting with her own, as is shown by the following valid expression of nested beliefs:

$$B_{A,3}^{0.6,1}(\mathscr{B}_{Bob,3}) \tag{14}$$

Finally, consider the situation where everything is as described before with the only difference that Alice now

shares her observation of the delayed train with Bob. It immediately follows that Bob updates his beliefs in the same way as Alice, which in turn yields an update in Alice's beliefs about Bob's beliefs so that (14) is no longer valid.

This example shows how Alice can reason about the influence of her own actions on Bob's belief state and therefore she can decide on actions that improve Bob's utility (as he does not have to wait in vain).

4 Problem Representations in PDT Logic

Two alternative approaches to represent problems in PDT Logic are available. The first approach requires an exhaustive specification of all possible threads (cf. Fig. 1 from the previous section's train example). If such a problem specification is given, satisfiability can be checked in a single pass through a straight-forward application of the given semantics. Thus, for a fixed set of threads, the resulting decision procedure's time complexity is in PTIME.

While there are some problem domains where exhaustive thread specifications are available (e.g., when analyzing attack graphs in cyber security scenarios, cf. [17]), in most cases such a representation is neither available nor easy to derive. Therefore, in [16] we have developed an alternative problem specification that does not rely on a fully materialized set of all possible threads. Informally speaking, this representation takes a set $\mathfrak{B}$ of PDT Logic belief formulae to specify how the target domain may evolve over time. From this set, we can employ heuristic search strategies to find a model that proofs satisfiability without having to materialize every possible thread. While the worst-case complexity of this problem is in EXPSPACE, many problems can be decided with considerably small computational effort by heuristically guiding the search for possible models.

Now that formal syntax and semantics for PDT Logic are established, we can return to the Three Prisoners Puzzle from Example 1 to give an intuition about problem specifications through a set of belief formulae.

Let us assume that at time point 1 the jailer responds to a's question and at time point 2 the executions will be carried out. Let $\mathtt{A}, \mathtt{B}$, and $\mathtt{C}$ be the respective events that prisoners a, b, and c will survive. Then, a's initial beliefs about these events can be expressed as

$$\mathfrak{B}_1 = \left\{ B^{1/3,1/3}_{a,0}(\mathtt{A}_2),\ B^{0,2/3}_{a,0}(\mathtt{B}_2),\ B^{0,2/3}_{a,0}(\mathtt{C}_2) \right\}. \tag{15}$$

The story continues and the jailer indeed gives the name of one man, either b or c, who is going to be executed. Assuming that $Obs_a(r(\neg\mathtt{B}))$ models the jailer's response that b is going to be executed (and accordingly for c), we obtain the additional formula

$$\mathfrak{B}_2 = \left\{ B^{1,1}_{a,0}(Obs_a(r(\neg\mathtt{B})))_1 \vee Obs_a(r(\neg\mathtt{C})))_1 \right\}. \tag{16}$$

To model the facts that the jailer answers truthfully, and that only one of the three prisoners can survive, we use the following formulae, respectively:

$$\mathfrak{B}_3 = \left\{ B^{1,1}_{a,0}(r^{pfr}_1(Obs_a(r(\neg X), \neg X))),\ B^{1,1}_{a,0}(r^{pfr}_0(\mathtt{A}, \neg(\mathtt{B} \vee \mathtt{C})) \wedge r^{pfr}_0(\mathtt{B}, \neg(\mathtt{A} \vee \mathtt{C})) \wedge r^{pfr}_0(\mathtt{C}, \neg(\mathtt{A} \vee \mathtt{B}))) \right\}. \tag{17}$$

Combining the sets $\mathfrak{B}_1, \ldots \mathfrak{B}_3$ then gives us a complete problem specification $\mathfrak{B}$.

From this set $\mathfrak{B}$, we can derive a set of representative threads and corresponding linear constraints for the respective probabilistic interpretations. As the resulting representation is beyond the space limits of this work, we simplify the example such that events $\mathtt{A}$, $\mathtt{B}$, and $\mathtt{C}$, are equally likely, i.e., we replace (15) with

$$\mathfrak{B}'_1 = \left\{ B^{1/3,1/3}_{a,0}(\mathtt{A}_2),\ B^{1/3,1/3}_{a,0}(\mathtt{B}_2),\ B^{1/3,1/3}_{a,0}(\mathtt{C}_2) \right\}. \tag{18}$$

Of course, with this adjustment we are not working with imprecise probabilities any longer. Still, this example serves to illustrate the following procedure and enables a significantly reduced presentation.

After hearing the jailer's response, the number of potential survivors decreases from 3 to 2, which in turn might lead a to believe (incorrectly!) that his chance of surviving increases to 1/2, regardless of the jailer's actual answer. While this update might appear reasonable at first glance, the result is somewhat puzzling: even though a has not received any significant new information (as he already knew that b or c is going to be executed), his beliefs are altered. As pointed out for example in [9], this result is due to updating in the *naive* problem space, while a correct representation in a *sophisticated* space will leave a's beliefs about his own survival unchanged. By creating a set of threads induced by $\mathfrak{B}$, we illustrate that probabilistic updates in PDT Logic do not suffer from problems induced by naive conditioning.

From $\mathfrak{B}$, we obtain the following set of threads:

Th_1	$Obs_a(r(\neg\mathtt{B}))$	$\neg\mathtt{A}, \neg\mathtt{B}, \mathtt{C}$	$\mathcal{I}_{a,0}(Th_1) = 1/3$
Th_2	$Obs_a(r(\neg\mathtt{B}))$	$\mathtt{A}, \neg\mathtt{B}, \neg\mathtt{C}$	$\mathcal{I}_{a,0}(Th_2) = 1/6$
Th_3	$Obs_a(r(\neg\mathtt{C}))$	$\neg\mathtt{A}, \mathtt{B}, \neg\mathtt{C}$	$\mathcal{I}_{a,0}(Th_3) = 1/3$
Th_4	$Obs_a(r(\neg\mathtt{C}))$	$\mathtt{A}, \neg\mathtt{B}, \neg\mathtt{C}$	$\mathcal{I}_{a,0}(Th_4) = 1/6$
t	1	2	

In this model, the prior interpretations $\mathcal{I}_{a,0}$ for threads

$Th_1, \ldots Th_4$ correctly represent the beliefs specified in (18). Then, regardless of the jailer's actual answer at time 1, a will disregard exactly two (either Th_1 and Th_2 or Th_3 and Th_4) of the given threads. Then, two threads remain, such that $\mathtt{A}$ holds in one of these threads (Th_2 or Th_4), and $\neg\mathtt{A}$ in the other. Independently from the actual response at time 1, updating the probabilistic interpretations according to (4) then results in the unchanged belief $B_{a,1}^{1/3,/13}(\mathtt{A})$. Without going through all technical details discussed in [16], this small example provides some intuition on how to derive a set of threads from a given set of belief belief formulae $\mathfrak{B}$. Furthermore, this example shows that probabilistic updates in PDT Logic yield desired results without running into pitfalls such as in the Three Prisoners Puzzle or the Monty Hall Problem.

5 Further Contributions

Next to the two alternative decision procedures, in [15] we have also formalized abductive reasoning for PDT Logic. As described before, we can provide a set of belief formulae $\mathfrak{B}$ to model a specific problem. Then, we can check if and how it is possible to induce a certain goal belief G for some agent. The core idea of abduction in PDT Logic is that (i) a set of belief formulae $\mathfrak{B}$ spans a set of of possible threads and (ii) the belief state of an agent can only be influenced through observations of the respective agent. Thus, we have shown in [15] how to construct a hypothesis space for the abduction problem automatically from a set of belief formulae $\mathfrak{B}$. Based on the decision procedure for $\mathfrak{B}$, we have provided a sound and complete algorithm to find a minimal solution to the abduction problem. Deciding whether an instance of the abduction has a solution is Σ_2^P-complete [15].

The methods discussed so far are only able to represent problems with finite time frames. In [13] we have developed an extension that enables the use of PDT Logic to reason about infinite streams of possible worlds under certain conditions. For this, it is required that the modeled domain can be represented through a stream that is aperiodic (i.e., states can occur at irregular times) and positive recurrent (i.e., every state has a finite mean recurrence time). Then, we can represent this stream as a sequence of finite-length segments and a transition model to represent the transition probabilities from one segment to the next. Within the resulting infinite stream of possible worlds, we can define arbitrary finite time windows and then use the segment and transition model to obtain a set of all possible threads within this time window. Based on this set of threads, we can then carry out satisfiability checks as described before.

References

1. Aumann RJ (1976) Agreeing to disagree. Ann Stat 4(6):1236–1239
2. Baltag A, Moss LS (2004) Logics for epistemic programs. Synthese 139(2):165–224
3. Baltag A, Moss LS, Solecki S (1998) The logic of public announcements, common knowledge, and private suspicions. In: Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge. TARK '98Morgan Kaufmann Publishers Inc., San Francisco, CA pp 43–56
4. van Benthem J (2003) Conditional probability meets update logic. J Logic Lang Inform 12(4):409–421. doi:10.1023/A:1025002917675
5. van Ditmarsch H, van der Hoek W, Kooi B (2007) Dynamic Epistemic Logic, 1st edn. Springer, New York
6. Fagin R, Halpern JY (1994) Reasoning about knowledge and probability. J ACM 41:340–367
7. Fagin R, Halpern JY, Moses Y, Vardi MY (1995) Reasoning about knowledge. MIT Press, Cambridge (1995)
8. Gardner M (1961) The Second Scientific American Book of mathematical puzzles and diversions. Simon and Schuster, New York
9. Grünwald PD, Halpern JY (2003) Updating probabilities. J Artif Int Res 19(1):243–278
10. Hintikka J (1962) Knowledge and belief. An introduction to the logic of the two notions. Cornell University Press, Ithaca, NY
11. Kooi BP (2003) Probabilistic dynamic epistemic logic. J Logic Lang Inform 12(4):381–408. doi:10.1023/A:1025050800836
12. Kripke SA (1963) Semantical considerations on modal logic. Acta Philos Fenn 16(1963):83–94
13. Martiny K, Moeller R (2014) PDT Logic for Stream Reasoning in Multi-agent Systems. In: SCSS 2014. 6th International Symposium on Symbolic Computation in Software Science. EPiC Series, vol 30. EasyChair, pp 35–46
14. Martiny K, Möller R (2015) A probabilistic doxastic temporal logic for reasoning about beliefs in multi-agent systems. In: ICAART 2015—Proceedings of the 7th International Conference on Agents and Artificial Intelligence. SciTePress, Portugal
15. Martiny K, Möller R (2015) Abduction in PDT Logic. In: AI 2015: advances in artificial intelligence. Springer International Publishing, New York
16. Martiny K, Möller R (2015) PDT logic: a probabilistic doxastic temporal logic for reasoning about beliefs in multi-agent systems. J Artif Int Res 57(1):39–112
17. Martiny K, Motzek A, Möller R (2015) Formalizing agents beliefs for cyber-security defense strategy planning. In: CISIS 2015—Proceedings of the 8th International Conference on Computational Intelligence in Security for Information Systems
18. Parikh R, Ramanujam R (2003) A knowledge based semantics of messages. J Logic Lang Inform 12(4):453–467. doi:10.1023/A:1025007018583
19. Plaza J (1989) Logics of public communications. In: Proceedings of the 4th international symposium on methodologies for intelligent systems: Poster session program. Oak Ridge National Laboratory, USA, pp 201–216
20. Shakarian P, Parker A, Simari G, Subrahmanian VS (2011) Annotated probabilistic temporal logic. ACM Trans Comput Logic 12(2):14:1–14:44