A Closer Look at the Probabilistic Description Logic Prob-$\mathcal{EL}$

Víctor Gutiérrez-Basulto
Universität Bremen, Germany
victor@informatik.uni-bremen.de

Jean Christoph Jung
Universität Bremen, Germany
jeanjung@informatik.uni-bremen.de

Carsten Lutz
Universität Bremen, Germany
clu@uni-bremen.de

Lutz Schröder
DFKI GmbH, Bremen, Germany
Lutz.Schroeder@dfki.de

Abstract

We study probabilistic variants of the description logic $\mathcal{EL}$. For the case where probabilities apply only to concepts, we provide a careful analysis of the borderline between tractability and EXPTIME-completeness. One outcome is that *any* probability value except zero and one leads to intractability in the presence of general TBoxes, while this is not the case for classical TBoxes. For the case where probabilities can also be applied to roles, we show PSPACE-completeness. This result is (positively) surprising as the best previously known upper bound was 2-EXPTIME and there were reasons to believe in completeness for this class.

Introduction

Classical description logics (DLs) are fragments of first-order logic (FOL) and thus do not provide any built-in means for representing uncertainty. This shortcoming has been addressed in a number of proposals for probabilistic DLs, see for example (Lukasiewicz and Straccia 2008; Jaeger 1994; da Costa and Laskey 2006; Lukasiewicz 2008) and references therein. Recently, a new family of probabilistic DLs was introduced in (Lutz and Schröder 2010), with the distinguishing feature that its members relate to the well-established probabilistic FOL of (Halpern 2003; Bacchus 1990) in the same way as classical DLs relate to traditional FOL. The main purpose of DLs from the new family, from now on called *Prob-DLs*, is to enable concept definitions that require reference to (degrees of) possibility, likelihood, and certainty. To this effect, Prob-DLs provide a probabilistic constructor $P_{\sim p}$ with $\sim \in \{<, \leq, =, \geq, >\}$ and $p \in [0, 1]$ that can be applied to concepts and sometimes also to roles. For example,

$$\mathsf{Patient} \sqcap \exists \mathsf{finding}.(\mathsf{Disease} \sqcap P_{>0.25}\mathsf{Infectious})$$

describes Patients having a disease that is infectious with probability at least .25.

As argued in (Lutz and Schröder 2010), Prob-DLs are well-suited to capture aspects of uncertainty that are present in almost all biomedical ontologies such as SNOMED CT. Such ontologies, which typically reach considerable size but still require efficient reasoning, are often formulated in lightweight DLs of the $\mathcal{EL}$ family for which the central reasoning problem of *subsumption* can be solved in polynomial time (Baader, Brandt, and Lutz 2005; Schulz, Suntisrivaraporn, and Baader 2007). Consequently, studying probabilistic extensions of $\mathcal{EL}$ in the style of the Prob-DL family is particularly relevant in this context. Some initial results in this direction have already been obtained in (Lutz and Schröder 2010).

The purpose of this paper is to establish a more complete picture of subsumption in probabilistic variants of $\mathcal{EL}$. In the first part of the paper, we consider *Prob-$\mathcal{EL}$* in which probabilities can only be applied to concepts, but not to roles. It was known that some concrete combinations of probability constructors such as $P_{>0}$ and $P_{>0.4}$ lead to intractability (in fact, EXPTIME-completeness) of subsumption while a restriction to the probability values zero and one does not. We prove the much more general result that the extension of $\mathcal{EL}$ with *any* single concept constructor $P_{\sim p}$, where $\sim \in \{<, \leq, =, \geq, >\}$ and $p \in (0, 1)$, results in EXPTIME-completeness. More specifically, this result applies to *general TBoxes*, i.e., to sets of concept inclusions $C \sqsubseteq D$ when $\sim \in \{=, \geq, >\}$ and even to the empty TBox when $\sim \in \{<, \leq, \}$. Inspired by the observation that many biomedical ontologies such as SNOMED CT are *classical TBoxes*, i.e., sets of concept definitions $A \equiv D$ with A atomic, we then show that probabilities other than zero and one *can* be used without losing tractability in classical TBoxes for the cases $\sim \in \{>, \geq\}$. More precisely, subsumption in Prob-$\mathcal{EL}$ is tractable when only the constructors $P_{\sim p}$ and $P_{=1}$ are admitted, for any (single!) choice of $\sim \in \{\geq, >\}$ and $p \in (0, 1)$. The resulting logics actually coincide for all possible choices. We also show that when a second probability value from the range $(0, 1)$ sufficiently 'far away' from p is added, the complexity of subsumption snaps back to EXPTIME-completeness.

In the second part, we consider *Prob-$\mathcal{EL}_r$*, where probabilities can be applied to both concepts and roles, concentrating on general TBoxes. While decidability is an open problem for full Prob-$\mathcal{EL}_r$, it was known that subsumption is in 2-EXPTIME and PSPACE-hard in *Prob-$\mathcal{EL}_r^{>0;=1}$*, where probability values are restricted to zero and one. It is interesting to note that Prob-DLs are a special kind of two-dimensional DLs as studied for example in (Gabbay et al. 2003) and that, until now, any two-dimensional extension

of $\mathcal{EL}$ turned out to have the same complexity as the corresponding extension of the expressive DL $\mathcal{ALC}$, see e.g. (Artale et al. 2007). Since subsumption in the $\mathcal{ALC}$-variant of Prob-$\mathcal{EL}_r^{>0;=1}$ is 2-EXPTIME-complete, it was thus tempting to conjecture that the same holds for Prob-$\mathcal{EL}_r^{>0;=1}$. We show that this is not the case by establishing a tight PSPACE upper bound for subsumption in Prob-$\mathcal{EL}_r^{>0;=1}$. This also implies PSPACE-completeness for the two-dimensional DL $\mathsf{S5}_{\mathcal{EL}}$, in sharp contrast with the 2-EXPTIME-completeness of $\mathsf{S5}_{\mathcal{ALC}}$.

Most proofs are deferred to the appendix of the long version, which is available from http://www.informatik.uni-bremen.de/∼clu/papers/index.html.

Preliminaries

Description logic concepts are built from a set of concept names $\mathsf{N_C}$ and a set of role names $\mathsf{N_R}$ (both countably infinite), using the available concept constructors. In the basic description logic $\mathcal{EL}$, these constructors are conjunction and existential restriction, which gives rise to the syntax rule

$$C, D ::= \top \mid A \mid C \sqcap D \mid \exists r.C$$

where $\top$ denotes the 'top-concept' (logical truth), A ranges over $\mathsf{N_C}$ and r over $\mathsf{N_R}$. To obtain a probabilistic version of $\mathcal{EL}$, we can apply probabilities to concepts or roles. Starting with the former, we consider the set of constructors

$$P_{\sim p}C \text{ with } \sim \in \{<, \leq, =, \geq, >\} \text{ and } p \in [0, 1],$$

denoting objects that are an instance of C with probability $\sim p$. The extension of $\mathcal{EL}$ with all these constructors is called *Prob-$\mathcal{EL}$*. For example, the SNOMED CT concept 'animal bite by potentially rabid animal' can be expressed as

$$\mathsf{Bite} \sqcap \exists \mathsf{by}.(\mathsf{Animal} \sqcap P_{>0.5} \exists \mathsf{has.Rabies}).$$

When we admit only a few values for $\sim$ and n, we put them in superscript; for example, Prob-$\mathcal{EL}^{>0.4,<0,1}$ denotes the extension of $\mathcal{EL}$ with $P_{>0.4}C$ and $P_{<0.1}C$.

Probabilities can be applied to roles using the concept constructors $\exists P_{\sim p} r.C$ where $\sim$ and p range over the same values as above, expressing that there is an element satisfying C that is related to the current element by the role name r with probability $\sim p$. For example, the SNOMED CT concept 'disease of possible viral origin' can be modeled as

$$\mathsf{Disease} \sqcap \exists P_{>0} \mathsf{origin.Viral}.$$

We denote the extension of Prob-$\mathcal{EL}$ with all the above (concept and role) constructors with *Prob-$\mathcal{EL}_r$*. We will also consider the restriction of Prob-$\mathcal{EL}_r$ to the constructors $P_{>0}$ and $P_{=1}$ both on concepts and roles, which is called *Prob-$\mathcal{EL}_r^{>0;=1}$*.

The semantics of classical DLs such as $\mathcal{EL}$ is based on interpretations $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, where $\Delta^{\mathcal{I}}$ is a non-empty set called the *domain* and $\cdot^{\mathcal{I}}$ is an *interpretation function* that maps each $A \in \mathsf{N_C}$ to a subset $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ and each $r \in \mathsf{N_R}$ to a subset $r^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$, see (Baader et al. 2003) for more details. The semantics of the probabilistic DLs considered here is given in terms of a *probabilistic interpretation* $\mathcal{I} = (\Delta^{\mathcal{I}}, W, (\mathcal{I}_w)_{w \in W}, \mu)$, where $\Delta^{\mathcal{I}}$ is the (non-empty) domain, W a non-empty set of *possible worlds*, μ a discrete probability distribution on W, and for each $w \in W$, $\mathcal{I}_w$ is a classical DL interpretation with domain $\Delta^{\mathcal{I}}$. We usually write $C^{\mathcal{I},w}$ for $C^{\mathcal{I}_w}$, and likewise for $r^{\mathcal{I},w}$. For concept names A and role names r, we define the probability

- $p_d^{\mathcal{I}}(A)$ that $d \in \Delta^{\mathcal{I}}$ is an A as $\mu(\{w \in W \mid d \in A^{\mathcal{I},w}\})$;
- $p_{d,e}^{\mathcal{I}}(r)$ that $d, e \in \Delta^{\mathcal{I}}$ are related by r as $\mu(\{w \in W \mid (d, e) \in r^{\mathcal{I},w}\})$.

Next, we extend $p_d^{\mathcal{I}}(A)$ to compound concepts C and define the extension $C^{\mathcal{I},w}$ of compound concepts by mutual recursion on C. The definition of $p_d^{\mathcal{I}}(C)$ is exactly as in the base case, with A replaced by C. The extension of compound concepts is defined as follows:

$$\begin{aligned}
\top^{\mathcal{I},w} &= \Delta^{\mathcal{I}} \\
(C \sqcap D)^{\mathcal{I},w} &= C^{\mathcal{I},w} \cap D^{\mathcal{I},w} \\
(\exists r.C)^{\mathcal{I},w} &= \{d \in \Delta^{\mathcal{I}} \mid \exists e \in C^{\mathcal{I},w}.(d, e) \in r^{\mathcal{I},w}\} \\
(P_{\sim p}C)^{\mathcal{I},w} &= \{d \in \Delta^{\mathcal{I}} \mid p_d^{\mathcal{I}}(C) \sim p\} \\
(\exists P_{\sim p} r.C)^{\mathcal{I},w} &= \{d \in \Delta^{\mathcal{I}} \mid \exists e \in C^{\mathcal{I},w} : p_{d,e}^{\mathcal{I}}(r) \sim p\}
\end{aligned}$$

In DLs, an ontology is formalized as a TBox. In this paper, we consider two kinds of TBoxes. A *general TBox* is a finite set of *concept inclusions* $C \sqsubseteq D$, where C, D are concepts. A *classical TBox* is a set of *concept definitions* $A \equiv C$, where A is a concept name and the left-hand sides of concept definitions are unique. Note that cyclic definitions are allowed.

A probabilistic interpretation $\mathcal{I}$ *satisfies* a concept inclusion $C \sqsubseteq D$ if $C^{\mathcal{I},w} \subseteq D^{\mathcal{I},w}$ and a concept definition $A \equiv C$ if $A^{\mathcal{I},w} = C^{\mathcal{I},w}$, for all $w \in W$. $\mathcal{I}$ is a *model* of a TBox $\mathcal{T}$ if it satisfies all inclusions/definitions in $\mathcal{T}$. A concept C is *subsumed by a concept D relative to a TBox $\mathcal{T}$* (written $\mathcal{T} \models C \sqsubseteq D$) if every model $\mathcal{I}$ of $\mathcal{T}$ satisfies the inclusion $C \sqsubseteq D$. Deciding subsumption is the most important reasoning task for DLs as it underlies the computation of the concept hierarchy, a central tool for structuring and accessing ontologies (Baader et al. 2003).

The above definition is the result of transferring the notion of subsumption from standard DLs to probabilistic DLs in a straightforward way. However, there is an alternative variant of subsumption that is natural for probabilistic DLs: a concept C is *positively subsumed by a concept D relative to a TBox $\mathcal{T}$* (written $\mathcal{T} \models^+ C \sqsubseteq D$) if $C^{\mathcal{I},w} \subseteq D^{\mathcal{I},w}$ for every probabilistic model $\mathcal{I} = (\Delta^{\mathcal{I}}, W, (\mathcal{I}_w)_{w \in W}, \mu)$ and every $w \in W$ with $\mu(w) > 0$. Intuitively, classical subsumption is about subsumptions that are *logically implied* whereas positive subsumption is about subsumptions that are *certain*. For example, when $\mathcal{T}_\emptyset$ is the empty TBox, then $\mathcal{T}_\emptyset \not\models P_{=1}A \sqsubseteq A$, but we can only have $d \in (P_{=1}A)^{\mathcal{I},v} \setminus A^{\mathcal{I},v}$ when $\mu(v) = 0$, thus non-subsumption is only witnessed by worlds that we are certain to not be the actual world. Consequently, $\mathcal{T}_\emptyset \models^+ P_{=1}A \sqsubseteq A$. In the extension Prob-$\mathcal{ALC}$ of Prob-$\mathcal{EL}$ with negation studied in (Lutz and Schröder 2010), positive subsumption can easily be reduced to subsumption. This does not seem easily possible in Prob-$\mathcal{EL}$. In fact, we will sometimes use (Turing) reductions in the opposite direction.

Probabilistic Concepts

In (Lutz and Schröder 2010), it was shown that subsumption in Prob-$\mathcal{EL}^{>0;=1}$ with general TBoxes is in PTIME, whereas the same problem is EXPTIME-complete in Prob-$\mathcal{EL}^{>0;>0.4}$ (both in the positive and in the unrestricted case). This raises the question whether *any* probability except 0,1 can be admitted in Prob-$\mathcal{EL}$ without losing tractability. The following theorem provides a strong negative result.

Theorem 1. *For all $p \in (0,1)$, (positive) subsumption in Prob-$\mathcal{EL}^{\sim p}$ relative to*

1. *general TBoxes is* EXPTIME*-hard when $\sim \in \{=, >, \geq\}$*
2. *the empty TBox is* EXPTIME*-hard when $\sim \in \{\leq, <\}$*

Matching upper bounds are an immediate consequence of the fact that each logic Prob-$\mathcal{EL}^{\sim p}$ is a fragment of the DL Prob-$\mathcal{ALC}_c$ for which subsumption was proved EXPTIME-complete in (Lutz and Schröder 2010). To prove the lower bounds, it suffices to show that each logic Prob-$\mathcal{EL}^{\sim p}$ is *non-convex* i.e., that there are a general TBox $\mathcal{T}$ and concepts $C, D_1, \ldots, D_n$, $n \geq 2$, such that $\mathcal{T} \models C \sqsubseteq D_1 \sqcup \cdots \sqcup D_n$, but $\mathcal{T} \not\models C \sqsubseteq D_i$ for all i (the semantics of disjunction is defined in the obvious way). Once that this is established, standard proof techniques from (Baader, Brandt, and Lutz 2005) can be used to reduce satisfiability in $\mathcal{ALC}$ relative to general TBoxes, which is EXPTIME-complete, to subsumption in Prob-$\mathcal{EL}^{\sim p}$. The following constructions work for standard subsumption and positive subsumption alike.

First consider $\sim = \geq$ and assume $p \leq 0.5$. Fix a $k > 0$ such that $k \cdot p > 1$ and set

$$\begin{aligned} \mathcal{T} &= \{A_i \sqcap A_j \sqsubseteq P_{\geq p} B_{ij} \mid 1 \leq i < j \leq k\} \\ C &= P_{\geq p} A_1 \sqcap \cdots \sqcap P_{\geq p} A_k \\ D_{ij} &= P_{\geq p} B_{ij} \end{aligned}$$

Intuitively, the probabilities stipulated by C sum up to > 1, thus some of the A_i have to overlap, but there is a choice as to which ones these are. Formally, we can show non-convexity by proving that $\mathcal{T} \models C \sqsubseteq \bigsqcup_{1 \leq i < j \leq k} D_{ij}$, but $\mathcal{T} \not\models C \sqsubseteq D_{ij}$ for any i, j. The comparisons $\sim \in \{=, >\}$ can be handled similarly. For $\sim = >$ and $p > 0.5$, we use a variation of the above. The main idea is to use $P_{>p} C$ to simulate $P_{>q} C$, for some $q \leq 0.5$, which brings us back to a case already dealt with. More precisely, let $n > 0$ be smallest such that $n > \frac{1}{2(1-p)}$ and set $q = pn - n + 1$. An easy computation shows that $0 \leq q < 0.5$. Moreover, it can be shown that

$$P_{>p} X_1 \sqcap \cdots \sqcap P_{>p} X_n \sqsubseteq P_{>q}(X_1 \sqcap \cdots \sqcap X_n)$$

which allows us to redo the above reduction with probability $q < 0.5$. The comparisons $\sim \in \{=, \geq\}$ can be dealt with similarly.

For the remaining cases $\sim \in \{<, \leq\}$, there is a very simple argument for non-convexity even w.r.t. the empty TBox: we have $\top \sqsubseteq P_{<p} A \sqcup P_{<p} P_{<p} A$, but neither $\top \sqsubseteq P_{<p} A$ nor $\top \sqsubseteq P_{<p} P_{<p} A$, and likewise when $\sim$ is $\leq$.

When $\sim \in \{=, >, \geq\}$, the proof of Theorem 1 relies on general TBoxes in a crucial way. It turns out that when we restrict ourselves to classical TBoxes, tractability can be attained even with probabilities other than 0 and 1.

R1	If $\exists r.B \in C_A$, and $C_{B'} \subseteq C_B$ then replace $A \equiv C_A$ with $A \equiv C_A \cup \{\exists r.B'\}$
R2	If $P_{=1} B \in C_A$ then replace $A \equiv C_A$ with $A \equiv C_A \cup C_B$
R3	If $P_{=1} B \in C_A$ then replace $A \equiv C_A$ with $A \equiv C_A \cup \{P_{\sim p} B\}$
R4	If $P_{\sim p} B \in C_A$, and $D \in \mathsf{cert}(C_B)$ then replace $A \equiv C_A$ with $A \equiv C_A \cup \{D\}$
R5	If $C_B \subseteq \mathsf{cert}(C_A)$ then replace $A \equiv C_A$ with $A \equiv C_A \cup \{P_{=1} B\}$
R6	If $P_{\sim p} B \in C_A$ and $C_{B'} \subseteq \mathsf{cert}(C_A) \cup C_B$ then replace $A \equiv C_A$ with $A \equiv C_A \cup \{P_{\sim p} B'\}$

Figure 1: TBox completion rules for positive subsumption

Theorem 2. *For all $\sim \in \{>, \geq\}$ and $p \in [0,1]$, (positive) subsumption in Prob-$\mathcal{EL}^{\sim p;=1}$ relative to classical TBoxes is in* PTIME.

To prove Theorem 2, we start with positive subsumption. We can assume $p > 0$ since subsumption in Prob-$\mathcal{EL}^{>0;=1}$ is in PTIME even with general TBoxes. To prove a PTIME upper bound, we use a 'consequence-driven' procedure similar to the ones in (Baader, Brandt, and Lutz 2005; Kazakov 2009). A concept name A is *defined* in a classical TBox $\mathcal{T}$ if there is a concept definition $A \equiv C \in \mathcal{T}$, and *primitive* otherwise. We can w.l.o.g. restrict our attention to the subsumption of *defined concept names* relative to TBoxes. We also assume that the input TBox is normalized to a set of concept definitions of the form

$$A \equiv P_1 \sqcap \cdots \sqcap P_n \sqcap C_1 \sqcap \cdots \sqcap C_m$$

$n, m \geq 0$, and where $P_1, \ldots, P_n$ are primitive concept names and $C_1, \ldots, C_m$ are of the form $P_{\sim p} A$, $P_{=1} A$, and $\exists r.A$ with A a defined concept name (note that the top concept is completely normalized away). It is well-known that such a normalization can be achieved in polytime, see (Baader 2003) for details. For a given TBox $\mathcal{T}$ and a defined concept name A in $\mathcal{T}$, we write C_A to denote the *defining concept* for A in $\mathcal{T}$, i.e., $A \equiv C_A \in \mathcal{T}$. Moreover, we deliberately confuse the concept $C_A = D_1 \sqcap \cdots \sqcap D_k$ with the set $\{D_1, \ldots, D_k\}$. We define a set of concepts 'certain for C_A' as

$$\mathsf{cert}(C_A) = \{P_* B \mid P_* B \in C_A\} \cup \bigcup_{P_{=1} B \in C_A} \{C_B\}$$

where, here and in what follows, P_* ranges over $P_{=1}$ and $P_{>p}$. Intuitively $\mathsf{cert}(C_A)$ contains concepts that hold with probability 1 whenever A is satisfied in some world. The algorithm starts with the normalized input TBox and then exhaustively applies the completion rules displayed in Figure 1. As a general proviso, each rule can be applied only if it adds a concept that occurs in $\mathcal{T}$ and actually changes the TBox, e.g., **R1** can only be applied when $\exists r.B'$ occurs

in $\mathcal{T}$ and $\exists r.B' \notin C_A$. Exemplarily, we explain rule **R5** in more detail. If all defining concepts C_B of B are certain for A, then $A \sqsubseteq P_{=1}B$, thus we can add $P_{=1}B$ to C_A. Let $\mathcal{T}^*$ be the result of exhaustive rule application and let C_A^* be the defining concept for A in $\mathcal{T}^*$, for all concept names A. The 'only if' direction requires a careful and surprisingly subtle model construction.

Lemma 3. *For all defined concept names A, B, we have $\mathcal{T} \models^+ A \sqsubseteq B$ iff $C_B^* \subseteq C_A^*$.*

It is easy to see that TBox completion requires only polytime: every rule application extends the TBox, but both the number of concept definitions and of conjuncts in each concept definition is bounded by the size of the original TBox.

To prove Theorem 2 for unrestricted subsumption, we provide a Turing reduction from unrestricted subsumption to positive subsumption. We again assume that the input TBox is in the described normal form and then exhaustively apply the rules shown in Figure 2, calling the result $\mathcal{T}^*$ with defining concept of the form C_A^*.

Lemma 4. *For all defined concept names A, B, we have $\mathcal{T} \models A \sqsubseteq B$ iff $C_B^* \subseteq C_A^*$.*

Clearly, the Turing reduction and thus the overall algorithm runs in polytime.

It is interesting to note that the proof of Theorem 2 is based on exactly the same algorithm, for all $\sim \in \{\geq, >\}$ and $p \in (0, 1]$. It follows that there is in fact only *a single logic* Prob-$\mathcal{EL}^{\sim p}$, for all such $\sim$ and p. Formally, given a Prob-$\mathcal{EL}^{\sim p}$-concept C, $\approx \in \{\geq, >\}$ and $q \in (0, 1]$, let $C_{\approx q}$ denote the result of replacing each subconcept $P_{\sim p}D$ in C with $P_{\approx q}D$ in C and similarly for Prob-$\mathcal{EL}^{\sim p}$-TBoxes $\mathcal{T}$.

Theorem 5. *For any $p, q > 0$, $\sim, \approx \in \{>, \geq\}$, $\mathcal{EL}^{\sim p}$-concepts C, D and -TBox $\mathcal{T}$, we have $\mathcal{T} \models^+ C \sqsubseteq D$ iff $\mathcal{T}_{\approx q} \models^+ C_{\approx q} \sqsubseteq D_{\approx q}$, and likewise for unrestricted subsumption.*

Consequently, the (potentially difficult!) choice of a concrete $\sim \in \{\geq, >\}$ and $p \in (0, 1]$ is moot. In fact, it might be more intuitive to replace the constructor $P_{\sim p}C$ with a constructor $\mathcal{L}\,C$ that describes elements which 'are likely to be a C', and to replace $P_{=1}C$ with the constructor $\mathcal{C}\,C$ to describe elements that 'are certain to be a C', see e.g. (Halpern and Rabin 1987; Herzig 2003) for other approaches to logics of likelihood. Note that the case $p = 0$ is different from the cases considered above: for example, we have $\mathcal{T}_\emptyset \models^+ \exists r.A \sqsubseteq \exists r.P_{>p}A$ iff $p = 0$, and likewise $\mathcal{T}_\emptyset \models P_{>p}\exists r.A \sqsubseteq P_{>p}\exists r.P_{>p}A$ iff $p = 0$. In the spirit of the constructors $\mathcal{C}$ and $\mathcal{L}$, $P_{>0}C$ can be replaced with a constructor $\mathcal{P}C$ that describes elements for which 'it is possible that they are a C'. For example, the SNOMED CT concepts 'definite thrombus' and 'possible thrombus' can then be written as $\mathcal{C}\,\mathsf{Thrombus}$ and $\mathcal{P}\,\mathsf{Thrombus}$ (although we speculate that the SNOMED CT designers mean 'likely' rather than 'possible').

It is a natural question whether the PTIME upper bound for classical TBoxes extends to the case of multiple probability values (except one, which is apparently always uncritical). The following result shows that many combinations of two probability values lead to (non-convexity, thus) intractability, even without any TBox.

S1	If $\exists r.B \in C_A$, and $C_{B'} \subseteq C_B$ then replace $A \equiv C_A$ with $A \equiv C_A \cup \{\exists r.B'\}$
S2	If $\mathcal{T} \models^+ \mathsf{cert}(C_A) \sqsubseteq P_*B$ then replace $A \equiv C_A$ with $A \equiv C_A \cup \{P_*B\}$

Figure 2: TBox completion rules for Turing reduction

Theorem 6. *Let $\sim \in \{>, \geq\}$, and $p, q \in [0, 1)$. Then (positive) subsumption in Prob-$\mathcal{EL}^{\sim p;\sim q}$ relative to the empty TBox is* EXPTIME*-hard if (i) $q = 0$, (ii) $p \leq 1/2$ and $q < p^2$, or more generally (iii) $2p - 1 < q < p^2$.*

In particular, we cannot combine the constructors $\mathcal{P}$ and $\mathcal{L}$ mentioned above without losing tractability. The above formulation of Theorem 6 is actually only a consequence of a more general (but also more complicated to state) result established in the appendix of the long version. We conjecture that (positive) subsumption in Prob-$\mathcal{EL}^{\sim p;\sim q}$ relative to the empty TBox is in PTIME relative to classical TBoxes whenever $p \geq q \geq p^2$ and that, otherwise, it is EXPTIME-hard.

Probabilistic Roles

Adding probabilistic roles to Prob-$\mathcal{EL}$ tends to increase the complexity of subsumption. While for full Prob-$\mathcal{EL}_r$ even decidability is open, it was shown in (Lutz and Schröder 2010) that subsumption is in 2-EXPTIME and PSPACE-hard in Prob-$\mathcal{EL}_r^{>0;=1}$. As discussed in the introduction, there were reasons to believe that this problem is actually 2-EXPTIME-complete. We show that this is not the case by proving a PSPACE upper bound, thus establishing PSPACE-completeness. This result holds both for positive and unrestricted subsumption, we start with the positive case.

We again concentrate on subsumption between concept *names* and assume that the input TBox is in a certain normal form, defined as follows. A *basic* concept is a concept of the form $\top$, A, $P_{>0}A$, $P_{=1}A$, or $\exists\alpha.A$, where A is a concept name and, here and in what follows, α is a *role*, i.e., of the form r, $P_{>0}r$, or $P_{=1}r$ with r a role name. Now, every concept inclusion in the input TBox is required to be of the form

$$X_1 \sqcap \cdots \sqcap X_n \sqsubseteq X$$

with $X_1, \ldots, X_n, X$ basic concepts. It is not hard to show that every TBox can be transformed into this normal form in polynomial time such that (non-)subsumption between the concept names that occur in the original TBox is preserved.

Let $\mathcal{T}$ be the input TBox in normal form, CN the set of concept names that occur in $\mathcal{T}$, BC the set of basic concepts in $\mathcal{T}$, and ROL the set of roles in $\mathcal{T}$. Call a role *probabilistic* if it is of the form $P_{=1}r$ or $P_{>0}r$. Our algorithm maintains the following data structures:

- a mapping Q that associates with each $A \in \mathsf{CN}$ a subset $Q(A) \subseteq \mathsf{BC}$ such that $\mathcal{T} \models A \sqsubseteq X$ for all $X \in Q(A)$;
- a mapping Q_{cert} that associates with each $A \in \mathsf{CN}$ a subset $Q_{\mathsf{cert}}(A) \subseteq \mathsf{BC}$ such that $\mathcal{T} \models A \sqsubseteq P_{=1}X$ for all $X \in Q_{\mathsf{cert}}(A)$;

R1	If $X_1 \sqcap \ldots \sqcap X_n \sqsubseteq X \in \mathcal{T}$ and $X_1, \ldots, X_n \in \Gamma$ then add X to Γ
R2	If $P_{=1}A \in \Gamma$ then add A to Γ
R3	If $\exists P_{=1}r.A \in \Gamma$ then add $\exists r.A$ to Γ
R4	If $A \in \Gamma$ then add $P_{>0}A$ to Γ
R5	If $\exists r.A \in \Gamma$ then add $\exists P_{>0}r.A$ to Γ
R6	If $\exists \alpha.A \in \Gamma$ and $B \in Q(A)$ then add $\exists \alpha.B$ to Γ

Figure 3: Saturation rules for $\mathsf{cl}(\Gamma)$

- a mapping R that associates with each probabilistic role $\alpha \in \mathsf{ROL}$ a binary relation $R(\alpha)$ on CN such that $\mathcal{T} \models A \sqsubseteq P_{>0}(\exists \alpha.B)$ for all $(A, B) \in R(\alpha)$.

Some intuition about the data structures is already provided above; e.g., $X \in Q(A)$ means that $\mathcal{T} \models A \sqsubseteq X$. However, there is also another view on these structures that will be important in what follows: they represent an abstract view of a model of $\mathcal{T}$, where each set $Q(A)$ describes the concept memberships of a domain element d in a world w with $d \in A^{\mathcal{I},w}$ and R describes role memberships, i.e., when $(A, B) \in R(\alpha)$, then $d \in A^{\mathcal{I},w}$ implies that in some world v with positive probability, d has an element described by $Q(B)$ as an α-successor. In this context, $Q_{\mathsf{cert}}(A)$ contains all concepts that must be true with probability 1 for any domain element that satisfies A in *some* world. Note that non-probabilistic roles r and probabilistic roles $P_{=1}r$ are not represented in the $R(\cdot)$ data structure; we will treat them in a more implicit way later on.

The data structures are initialized as follows, for all $A \in \mathsf{CN}$ and relevant roles α:

$$Q(A) = \{\top, A\} \qquad Q_{\mathsf{cert}}(A) = \{\top\} \qquad R(\alpha) = \emptyset.$$

The sets $Q(\cdot)$, $Q_{\mathsf{cert}}(\cdot)$, and $R(\cdot)$ are then repeatedly extended by the application of various rules. Before we can introduce these rules, we need some preliminaries. As the first step, Figure 3 presents a (different!) set of rules that serves the purpose of saturating a set of concepts Γ. We use $\mathsf{cl}(\Gamma)$ to denote the set of concepts that is the result of exhaustively applying the displayed rules to Γ, where any rule can only be applied if the added concept is in BC, but not yet in Γ. The rules access the data structure $Q(\cdot)$ introduced above and shall later be applied to the sets $Q(A)$ and $Q_{\mathsf{cert}}(A)$, but they will also serve other purposes as described below. It is not hard to see that rule application terminates after polynomially many steps.

The rules that are used for completing the data structures $Q(\cdot)$, $Q_{\mathsf{cert}}(\cdot)$, and $R(\cdot)$ are more complex and refer to 'traces' through these data structures, which we introduce next.

Definition 7. Let $B \in \mathsf{CN}$. A *trace to* B is a finite sequence $S, A_1, \alpha_2, A_2, \ldots, \alpha_n, A_n$ where

1. $S = A$ for some $P_{>0}A \in Q(A_1)$ or $S = (r, B)$ for some $(A_1, B) \in R(P_{>0}r)$;
2. each $A_i \in \mathsf{CN}$ and each $\alpha_i \in \mathsf{ROL}$ is a probabilistic role, such that $A_n = B$;
3. $(A_i, A_{i-1}) \in R(\alpha_i)$ for $1 < i \leq n$.

If t is a trace of length n, we use t_k, $k \leq n$, to denote the shorter trace $S, A_1, \alpha_2, \ldots, \alpha_k, A_k$. Intuitively, the purpose of a trace is to deal with worlds that are generated by concepts $P_{>0}A$ and $\exists P_{>0}r.A$; there can be infinitely many such worlds as Prob-$\mathcal{EL}_r^{>0;=1}$ lacks the finite model property, see (Lutz and Schröder 2010). The trace starts at some domain element represented by a set $Q(A_1)$ in the world generated by the first element S of the trace, then repeatedly follows role edges represented by $R(\cdot)$ backwards until it reaches the final domain element represented by $Q(B)$. The importance of traces stems from the fact that information can be propagated along them, as captured by the following notion.

Definition 8. Let $t = S, A_1, \alpha_2, \ldots, \alpha_n, A_n$ be a trace of length n. Then the *type* $\Gamma(t) \subseteq \mathsf{BC}$ of t is

- $\mathsf{cl}(\{A\} \cup Q_{\mathsf{cert}}(A_1))$ if $n = 1$ and $S = A$;
- $\mathsf{cl}(Q_{\mathsf{cert}}(A_1) \cup \{\exists r.B' \in \mathsf{BC} \mid B' \in Q_{\mathsf{cert}}(B)\})$ if $n = 1$ and $S = (r, B)$;
- $\mathsf{cl}(Q_{\mathsf{cert}}(A_n) \cup \{\exists \alpha_n.B' \in \mathsf{BC} \mid B' \in \Gamma(t_{n-1})\})$ if $n > 1$.

Note that the rules **R1** to **R6** are used in every step of this inductive definition. The mentioned propagation of information along traces is now as follows: if there is a trace t to B, then any domain element that satisfies B in *some* world must satisfy the concepts in $\Gamma(t)$ in some other world. So if for example $P_{>0}A \in \Gamma(t)$, we need to add $P_{>0}A$ also to $Q_{\mathsf{cert}}(B)$ and to $Q(B)$.

Figure 4 shows the rules used for completing the data structures $Q(\cdot)$, $Q_{\mathsf{cert}}(\cdot)$, and $R(\cdot)$. Note that **S6** and **S7** implement the propagation of information along traces, as discussed above. Our algorithm for deciding (positive) subsumption starts with the initial data structures defined above and then exhaustively applies the rules shown in Figure 4. To decide whether $\mathcal{T} \models A \sqsubseteq B$, it then simply checks whether $B \in Q(A)$.

Lemma 9. *Let $\mathcal{T}$ be a Prob-$\mathcal{EL}_r^{>0;=1}$-TBox in normal form and A, B be concept names. Then $\mathcal{T} \models^+ A \sqsubseteq B$ iff, after exhaustive rule application, $B \in Q(A)$.*

We now argue that the algorithm can be implemented using only polynomial space. First, it is easy to see that there can be only polynomially many rule applications: every rule application extends the data structures $Q(\cdot)$, $Q_{\mathsf{cert}}(\cdot)$, and $R(\cdot)$, but these structures consist of polynomially many sets, each with at most polynomially many elements. It thus remains to verify that each rule application can be executed using only polyspace, which is obvious for all rules except those involving traces, i.e., **S6** and **S7**. For these rules, we first note that it is not necessary to consider all (infinitely many!) traces. In fact, a straightforward 'pumping argument' can be used to show that there is a trace t to B with some relevant concept $C \in \Gamma(t)$ iff there is a *non-repeating* such trace, i.e., a trace t' of length n such that for all distinct $k, \ell \leq n$, we have $\Gamma(t'_k) \neq \Gamma(t'_\ell)$. Clearly, the length of non-repeating traces is bounded by 2^m, m the size of $\mathcal{T}$. To get to polyspace, we

S1 apply **R1-R6** to $Q(A)$ and $Q_{\mathsf{cert}}(A)$

S2 if $P_*B \in Q(A)$
then add P_*B to $Q_{\mathsf{cert}}(A)$

S3 if $C \in Q_{\mathsf{cert}}(A)$
then add $P_{=1}C$ and C to $Q(A)$

S4 If $\exists\alpha.B \in Q(A)$ with α a probabilistic role
then add (A, B) to $R(\alpha)$.

S5 If $P_{>0}B_1 \in Q(A)$, $(B_1, B_2) \in R(\alpha)$, $B_3 \in Q_{\mathsf{cert}}(B_2)$
then add $\exists\alpha.B_3$ to $Q_{\mathsf{cert}}(A)$

S6 if t is a trace to B and $P_*A \in \Gamma(t)$
then add P_*A to $Q_{\mathsf{cert}}(B)$

S7 if t is a trace to B and $\exists\alpha.A \in \Gamma(t)$
with α a probabilistic role
then add (B, A) to $R(\alpha)$

Figure 4: The rules for completing the data structures.

use a non-deterministic approach, enabled by Savitch's theorem: to check whether there is a trace t to B with $C \in \Gamma(t)$, we guess t step-by-step, at each time keeping only a single A_i, α_i and $\Gamma(t_i)$ in memory. When we reach a situation where $A_i = B$ and $C \in \Gamma(t_i)$, our guessing was successful and we apply the rule. We also maintain a binary counter of the number of steps that have been guessed so far. As soon as this counter exceeds 2^m, the maximum length of non-repeating traces, we stop the guessing and do not apply the rule. Clearly, this yields a polyspace algorithm.

Theorem 10. *Positive subsumption in Prob-$\mathcal{EL}_r^{>0;=1}$ relative to general TBoxes is* PSPACE*-complete.*

As a byproduct, the proof of Lemma 9 yields a unique least model (in the sense of Horn logic), thus proving convexity of Prob-$\mathcal{EL}_r^{>0;=1}$. Note that positive subsumption in Prob-$\mathcal{EL}_r^{>0;=1}$ is actually the same as subsumption in the two-dimensional description logic $\mathsf{S5}_{\mathcal{EL}}$, which is thus also PSPACE-complete. Using a Turing reduction similar to that shown in Figure 2, we can 'lift' the result from positive subsumption to unrestricted subsumption.

Theorem 11. *Subsumption in Prob-$\mathcal{EL}_r^{>0;=1}$ relative to general TBoxes is* PSPACE*-complete.*

Conclusion

We have established a fairly complete picture of the complexity of subsumption in Prob-$\mathcal{EL}$, although some questions remain open. We speculate that Theorem 2 can be proved also when $\sim$ is $=$ with only minor changes (e.g. rule **R3** becomes unsound). It would be interesting to verify the conjecture made below Theorem 6 that (positive) subsumption in Prob-$\mathcal{EL}^{\sim_p;\sim_q}$ relative to classical TBoxes is in PTIME whenever $p \geq q \geq p^2$ and that, otherwise, it is EXPTIME-hard relative to the empty TBox. It is even conceivable that the conjectured PTIME result can be further generalized to any set of probability values $\mathcal{P} \subseteq [0, 1]$ as long as $q \geq p^2$ whenever $p, q \in \mathcal{P}$ and $p \geq q$. Moreover, variants of Theorem 6 that involve, additionally or exclusively, the case where $\sim$ is $=$ would also be of interest.

Acknowledgements

The present work was supported by the DFG project *Probabilistic Description Logics* (LU1417/1-1, SCHR1118/6-1) and DAAD-CONACYT grant 206550.

References

Artale, A.; Kontchakov, R.; Lutz, C.; Wolter, F.; and Zakharyaschev, M. 2007. Temporalising tractable description logics. In *Proc. of TIME07*, 11–22. IEEE Press.

Baader, F.; Calvanese, D.; McGuinness, D. L.; Nardi, D.; and Patel-Schneider, P. F., eds. 2003. *The Description Logic Handbook*. Cambridge University Press.

Baader, F.; Brandt, S.; and Lutz, C. 2005. Pushing the $\mathcal{EL}$ envelope. In *Proc. of IJCAI05*, 364–369. Professional Book Center.

Baader, F. 2003. Terminological cycles in a description logic with existential restrictions. In *Proc. of IJCAI03*, 325–330. Morgan Kaufmann.

Bacchus, F. 1990. *Representing and Reasoning with Probabilistic Knowledge*. Cambridge, MA: MIT Press.

da Costa, P. C. G., and Laskey, K. B. 2006. PR-OWL: A framework for probabilistic ontologies. In *Proc. of FOIS06*, 237–249. IOS Press.

Gabbay, D. M.; Kurucz, Á.; Wolter, F.; and Zakharyaschev, M. 2003. *Many-Dimensional Modal Logics: Theory and Applications*. Elsevier.

Halpern, J. Y., and Rabin, M. O. 1987. A logic to reason about likelihood. *Artificial Intelligence* 32:379–405.

Halpern, J. Y. 2003. *Reasoning About Uncertainty*. MIT Press.

Herzig, A. 2003. Modal probability, belief, and actions. *Fundamentae Informaticae* 57:323–344.

Jaeger, M. 1994. Probabilistic reasoning in terminological logics. In *Proc. of KR94*, 305–316. Morgan Kaufmann.

Kazakov, Y. 2009. Consequence-driven reasoning for Horn SHIQ ontologies. In *Proc. of IJCAI09*, 2040–2045. AAAI Press.

Lukasiewicz, T., and Straccia, U. 2008. Managing uncertainty and vagueness in description logics for the semantic web. *Journal of Web Semantics* 6(4):291–308.

Lukasiewicz, T. 2008. Expressive probabilistic description logics. *Artificial Intelligence* 172:852–883.

Lutz, C., and Schröder, L. 2010. Probabilistic description logics for subjective uncertainty. In *Proc. of KR10*. AAAI Press.

Schulz, S.; Suntisrivaraporn, B.; and Baader, F. 2007. SNOMED CT's problem list: Ontologists' and logicians' therapy suggestions. In *Proc. of Medinfo07 Congress*, IOS Press.